Shaping the Selection of Fields of Study in Afghanistan through Educational Data Mining Approaches

vorgelegt von
M.Sc.
Abdul Rahman Sherzad
geb. in Herat, Afghanistan

von der Fakultät IV – Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften
– Dr.-Ing. –

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr. Ziawasch Abedjan, Technische Universität Berlin
Gutachter:	Prof. Dr. Uwe Nestmann, Technische Universität Berlin
Gutachter:	Prof. Dr. Niels Pinkwart, Humboldt-Universität zu Berlin
Gutachter:	Prof. Dr. Sebastian Bab, Fachhochschule Dortmund
Gutachter:	Dr. Nazir Peroz, Technische Universität Berlin

Tag der wissenschaftlichen Aussprache: 13. Juli 2018

Berlin 2018

ABSTRACT

Every year around 250000 high school graduates participate in *'Kankor'*, *the Afghan national university entrance exam*, while the seating capacity of the country's 36 public universities is one-fourth of that number. Currently, public and private sectors lack advisory systems to guide the increasing number of participants to choose their fields of study. This is further exacerbated by the fact that the Afghan high school education system and the Kankor are non-specialized and participants have only *two chances* to take the exam in their lifetime. Thus, participants make uninformed choices. This research is an effort to introduce a data-oriented system to support Afghan policymakers to reshape the education system and to empower advisors to recommend disciplines to test takers based on their skills and interests. This is the first such attempt in Afghanistan. The findings of this research are mainly based on 1.5 million records of Kankor candidates (2003-2015), 6000 records of high school students' marks, more than 3000 questionnaires and interviews, author's observations as well as empirical studies.

The research findings show that high school marks alone cannot be a reliable data to replace Kankor for university admission. These research findings suggest strategic solutions for Kankor itself: an extensive model to introduce data culture in the core of the Afghan education system. The solutions offered create an evolving opportunity for the educational institutions to standardize testing approaches by considering the actual capacities and interests of test takers. This research proposes and develops the following three methods to better understand the interests, preferences and competences of applicants: (1) *'assessment test'* to better understand the interest and tendency of Kankor applicants (2) *'predictive models using data mining algorithms on Kankor data'* and (3) *'recommender systems approach using high school marks'* to identify disciplines which are closest to Kankor candidates' knowledge set, skills and preferences. The *'e-Kankor'* is proposed to simulate the *'actual Kankor exam'* allowing candidates to learn the importance of timeliness, methods and techniques through exam cram and to estimate the candidates' scores prior to taking the *actual Kankor*. The scores are then used in the *'prediction models'* as input data.

The models proposed and developed in this research, if executed systematically, will on average support 250000 Kankor participants every year in choosing their fields of study judiciously. Moreover, they will support nearly 1000000 students who are currently in high school. That would shape and transform the selection of careers for generations to come. The developed models provide a roadmap for systematic student advising to become an integral part of the education structures. This research establishes the foundations to streamline standardized testing in a country which is in the very initial stages of experiencing the power of data. Such a transformation, powered by insight from data, may be the beginning of a paradigm shift. The models and findings of this research have the potential to be contextualized and applied in other countries that share a similar education system to Afghanistan. This research is also important in the field of educational data mining independent of the case study on Afghanistan. Finally, this work opens up numerous opportunities for further investigations in the future.

ZUSAMMENFASSUNG

Jedes Jahr nehmen rund 250000 Schulabsolventen an der afghanischen nationalen Aufnahmeprüfung Kankor teil, wobei die Kapazität der 36 staatlichen Universitäten des Landes ein Viertel dieser Zahl ausmacht. Derzeit fehlt es dem öffentlichen und dem privaten Sektor an Beratungssystemen, um die steigende Zahl von Teilnehmern bei der Auswahl ihrer Studienfächer zu unterstützen. Dies wird noch dadurch verstärkt, dass das afghanische Schulsystem und Kankor nicht spezialisiert sind und die Teilnehmer nur zwei Chancen haben, das Examen in ihrem Leben zu absolvieren. Somit treffen die Teilnehmer uninformierte Entscheidungen. Diese Forschung ist ein Versuch, ein datenorientiertes System einzuführen, um afghanische Politiker dabei zu unterstützen, das Bildungssystem umzugestalten und Berater zu befähigen, den Testteilnehmern Fachdisziplinen basierend auf ihren Fähigkeiten und Interessen zu empfehlen. Dies ist der erste derartige Versuch in Afghanistan. Die Ergebnisse dieser Forschung basieren hauptsächlich auf 1,5 Millionen Aufzeichnungen von Kankor-Kandidaten (2003-2015), 6000 Aufzeichnungen von Abschlussnoten, mehr als 3000 Fragebögen und Interviews, Autorenbeobachtungen sowie empirischen Studien.

Die Forschungsergebnisse zeigen, dass High-School-Noten allein keine verlässlichen Daten sein können, um Kankor für die Aufnahme in die Universität zu ersetzen. Die Forschungsergebnisse dieser Arbeit schlagen strategische Lösungen für Kankor selbst vor: ein umfassendes Modell zur Einführung einer Datenkultur in den Kern des afghanischen Bildungssystems. Die angebotenen Lösungen schaffen eine sich entwickelnde Möglichkeit für die Bildungseinrichtungen, Testansätze zu standardisieren, indem die tatsächlichen Fähigkeiten und Interessen der Testteilnehmer berücksichtigt werden. Diese Forschung empfiehlt und entwickelt die folgenden drei Methoden, um die Interessen, Präferenzen und Kompetenzen der Bewerber besser zu verstehen: (1) ein "Assessment-Test", um das Interesse und die Tendenz der Kankor-Bewerber besser zu verstehen (2) "Vorhersagemodelle auf Basis von Data-Mining-Algorithmen auf den Kankordaten" sowie (3)"Empfehlungssystem-Ansätze auf Basis von Schulnoten", um Disziplinen zu ermitteln, die den Kenntnissen, Fähigkeiten und Vorlieben der Kankor-Kandidaten am nächsten sind. Der "E-Kankor" wird vorgeschlagen, um die "tatsächliche Kankor-Prüfung" zu simulieren, die es Kandidaten ermöglicht, die Wichtigkeit von Aktualität, Methoden und Techniken der Prüfung zu lernen und die Punktzahlen der Kandidaten zu schätzen, bevor der eigentliche Kankortest absolviert wird. Die Bewertungen werden dann in den "Vorhersagemodellen" als Input Daten verwendet.

Die in dieser Arbeit vorgeschlagenen und entwickelten Modelle werden, wenn sie systematisch durchgeführt werden, im Durchschnitt 250000 Kankor-Teilnehmer pro Jahr bei der Auswahl ihrer Studienfächer unterstützen. Darüber hinaus werden sie fast 1000000 Schüler unterstützen, die derzeit in der Schule sind. Das würde die Auswahl an Karrieren für künftige Generationen prägen und verändern. Die entwickelten Modelle bieten eine Roadmap für die systematische Beratung von Studierenden, um ein integraler Bestandteil der Bildungsstrukturen zu werden. Diese Untersuchung schafft die Grundlagen für die Rationalisierung von standardisierten Tests in einem Land, das sich gerade in der Anfangsphase hinsichtlich eines Erlebens der Macht der Daten befindet. Eine solche Transformation, die auf Daten basiert, könnte der Beginn eines Paradigmenwechsels sein. Die Modelle und Ergebnisse dieser Forschung können in anderen Ländern, die ein ähnliches Bildungssystem wie Afghanistan haben, kontextualisiert und angewendet werden. Diese Forschung ist auch unabhängig von der Fallstudie Afghanistan für den Bereich des Educational Data Mining relevant. Schließlich eröffnet diese Arbeit zahlreiche Möglichkeiten für weitere Untersuchungen in der Zukunft.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my advisor professors, Prof. Dr.-Ing. Uwe Nestmann, Prof. Dr. Sebastian Bab, and Dr. Nazir Peroz for their unceasing support towards my PhD; not just for their insightful guidance and encouragement but also for their thoughtful and cross-cutting intricacy which broadened both the technical and philosophical aspects of this research from various angles. Their guidance helped me throughout my research and writing of this dissertation.

My gratitude and appreciations also go to Prof. Dr. Niels Pinkwart, who provided me an opportunity to present my topic to his research team, and willingly accepted to be my referee as well as to Prof. Dr. Ziawasch Abedjan, the chairperson of the doctoral examination board. Thank you very much, Prof. Pinkwart and Prof. Abedjan.

Regarding the paperwork, administration processes and other support I would like to express my sincere thanks to the ZiiK team of TU-Berlin (Agnieszka Zielinska, Andrea Hillenbrand, Anna Marković, Chi-Thanh Christopher Nguyen, Daniel Tippmann, Marius Mailänder, Stefan Heil, Tilman Schieber, Vanessa Hüber and other team members) for their gracious support and guidance. It has been instrumentally helpful.

Kudos to DAAD, TU-Berlin and the Government of Germany for their generous support and commitment towards capacity building and capacity strengthening programs for the Afghan people. As a beneficiary of DAAD, I am confident that investment in the future of a modern Afghanistan will truly material through such commitments.

I would like to extend my appreciations to the government of Afghanistan and the Afghan Ministry of Higher Education. I would like to request that they consider using the results of this work for making it easier for Afghan students to choose suitable university majors, for introducing an efficient data architecture for saving Kankor and other related data, and also for using data as a crucial asset for improving education across Afghanistan.

I am also very grateful to Bashir Ahmad, who helped me in proofreading all my chapters from the beginning; and Moslem Shah and Nadir Noorzai for their quick review of my PhD dissertation as well as for their encouragement and motivation. Furthermore, I sincerely thank my former students as well as my colleagues in Afghanistan who helped me in the data collection process involved in this research.

Last but not the least, I would like to thank my family: my parents, brothers and sisters for supporting me spiritually throughout writing this dissertation and my life in general. Your prayers for me was what sustained me thus so far.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Chapter 1. INTRODUCTION

The net increase in the enrollment rate in education, expansion of access to internet as well as the gradual adoption of technology in education sector have led to the aggregation of large amounts of student data at educational institutions (schools, colleges, and universities), which makes it vital to use data mining methods to improve and better manage the needed development supported by informed decisions. In this growth, institutions have become both the owner and supplier of increasing amounts of biodata and academic records. To get essential benefits from the data, powerful techniques are required to extract the useful knowledge which is valuable and significant for the decision of policymakers. Data mining methods are widely used in customer segmentation, fraud detection, intrusion detection and other areas which have shown remarkable results. In the domain of education, there is a new emerging field, called Educational Data Mining (EDM), in which methods are developed to discover knowledge from data originating from educational environments. EDM is useful in many different areas including identifying students at-risk of attrition, recommending the right course of study for students, supporting students in the selection of their majors, and effectively evaluating institutional performance.

Moreover, recommender systems have become increasingly popular in recent years in various industries, for instance, supporting users to make decisions about what movies to watch, what products to buy, what jobs to look at, or what books to read. While these systems have shown their effectiveness in e-commerce, movies, music and social networks, the field of education is an emerging and very promising application area. Technology has consequently transformed how education is provided as a service, students and educators are at the forefront of benefiting the use of educational tools, software, and platforms, not only to pursue research but also to collaborate on a wide variety of academic activities. Recommender systems enable students to better calibrate information on type of courses they are most likely to benefit from, using their background and prior experiences.

The main goal of this thesis (dissertation) is to provide the basis for educational data mining and recommendation techniques in the context of a developing country, particularly Afghanistan, to help policymakers in shaping the education system. The proposed and developed methods laid out in this thesis also intend to allow advisors to provide guidance to *Kankor candidates* choosing suitable fields of study (majors or disciplines) which are consistent with *their high school marks, Kankor practice tests performance and score through e-Kankor, their choice of field of study and their interest.* Previous *Kankor results' data* and *high school students' marks* will be used to develop and test the prediction models, the data extracted from candidate's profile in the *e-Kankor* and other supportive tools will be used as input to the prediction models to suggest the suitable choices for the tested candidate.

In almost all education systems, students must make decisions about their academic career by choosing a field of study. Some countries, such as the US, have a flexible structure in higher education, such that students do not have to start specializing until the 2nd or 3rd year, whereas other countries such as Iran, Turkey, India, China, and UK require students to declare their specialty before higher education. According to *World Education Services* (WES Staff 2017; Asadzadeh 2015; Sedgwick 2000; WES Staff 2013), in Iran at the end of grade 9, students are divided into one of the following three streams; academic, technical or vocational. In Turkey (Kevin and WES Staff 2017), students are assessed after grade 8 and they can then study at general high schools, or vocational & technical high schools. In China (Sedgwick and Chen 2002), students in regular high schools typically must choose between the science stream and the art stream prior to the start of grade 11. In India (Clark 2006), the second phase of secondary education is designed to allow for diversification and specialization.

Choosing a field of study should be planned and based on knowledge the student has, and if and when the wrong choice is made, the implications can be severe, alternatively, reversing the decision may be costly for students (Richard L. and Jessica G. 2014, 1:55). Hence, informed decisions to choose field of study can have a very significant influence on students' academic careers. The more accurate those decisions are, the better the development of a student's potential and the quality of education would be.

Generally, education systems vary from country to country and region to region. In many education systems, the important decision of choosing a field of study should be made in early stages when students are likely underprepared. The students do not have enough knowledge of the requirements and challenges ahead of them to make an effective choice. Initially, decisions are made by students themselves based on diverse factors such as, general interest and preferences, personal goals and future job market, and sometime just a random decision. Students tend to look to others (advisors, parents, peers, and high school counselors) for feedback.

Therefore, due to the potential positive or negative impact the choice of field of study can have on the student experience, it is crucial for educational institutions to meet and overcome this challenge and improve career decision-making processes systematically. Hence, researchers from a variety of disciplines including computer science, education, psychology, and statistics and other domains together found the community of EDM. A significant number of interdisciplinary researches have been carried out to better understand students, and the settings in which they learn (C. Romero and Ventura 2007, 2010; Cristobal Romero et al. 2010). The EDM community was established to offer data mining solutions in the education contexts and methods to investigate how data mining approaches could eventually address the root causes of common education challenges. Moreover, collaboration among academic advisors and university researchers has the potential to increase the effectiveness of academic advising (Gordon, Habley, and Grites 2008). Thus, researchers have carried out studies and introduced institutionalized counseling mechanisms, through so-called advisors whose role is to help students with their academic careers methodically.

In recent years (Richard L. and Jessica G. 2014), academic advising has become much more prominent on most school and college campuses whereby advising responsibilities has also undergone a tremendous change toward the development of students. Generally (Lowe and Toney 2000), academic advisors are expected to guide students in making practical academic decisions, discover a range of options available to them, and discuss the consequences of their choices.

In the case of Afghanistan, the context varies profoundly. Within the current structures, students are not offered specialized studies at high schools, per *UNESCO Office in Kabul* (Padilla, n.d.), adult literacy rates are among the lowest in the world, thus parents are not a suitable source of support to guide students in choosing a career. Structurally, high schools also lack the existence of basic career counseling services, and there are no academic advisory organizations outside educational institutions to guide the students on their critical career decisions, in other words, student counseling and advisory are literally nonexistent.

In Afghanistan, university applicants can opt for five fields of study through the *National University Entrance Exam* (*Kankor*, from the French word Concours), but there are several flaws. Factors leading to inappropriate selection of field of study and affecting the quality of education identified in this dissertation are classified in the following:

1. High school education is general, thus leads to misinformed selection of field of study by applicants.
2. Kankor candidates are permitted to take the *Kankor* only twice in their lifetime. If they fail this examination two times, they will not be able to enroll in public universities (aka state universities) and semi-higher education institutions (Technical and Vocational institutions, and Teacher Training Colleges) thereafter.
3. There are more than 100 fields of study, which are not categorized.
4. There is unhealthy competition among families and they want their children to select fields of study which are deemed more prestigious, such as medicine and engineering, this leads to psychological pressures on Kankor participants.
5. High school teachers are generally not able to help Kankor candidates with making career decisions
6. There is a sharp and continuous increase in the number of Kankor participants.

The Ministry of Education (MoE), and Ministry of Higher Education (MoHE) are the two main bodies directly responsible for overseeing general education and higher education. To gain admission into higher education, all eligible high school graduates (Kankor candidates) must successfully pass *Kankor*. In this examination, students are required to choose five fields of study. Most of the students have no knowledge of the requirements and prerequisites of the offered fields of study, and select randomly. The results are poor performance or high rates of dropout in higher education.

While enrollment in higher education has been increasing, poor graduation and high attrition rates are widespread in higher education, with a significant economic and social impact for students and the society. To confirm this argumentation, the author of this dissertation collected the numbers of Kankor participants admitted into different faculties (fields of study) at Herat University in 2009, 2010, and 2011, and the successful graduation rates for

the same participants in 2013, 2014, and 2015 respectively, to compare the ratio of admission vs. successful graduation. As shown in **Figure 1-1**, almost one-third of the participants admitted into Herat University could not complete their higher education studies – dropout, leaving university before completion, taking longer to complete, change of discipline could be the reasons. Academic advising (Richard L. and Jessica G. 2014, 1:5) can be one key to overcoming this problem and help students reach their educational and career goals in the most efficient, cost-effective manner possible.

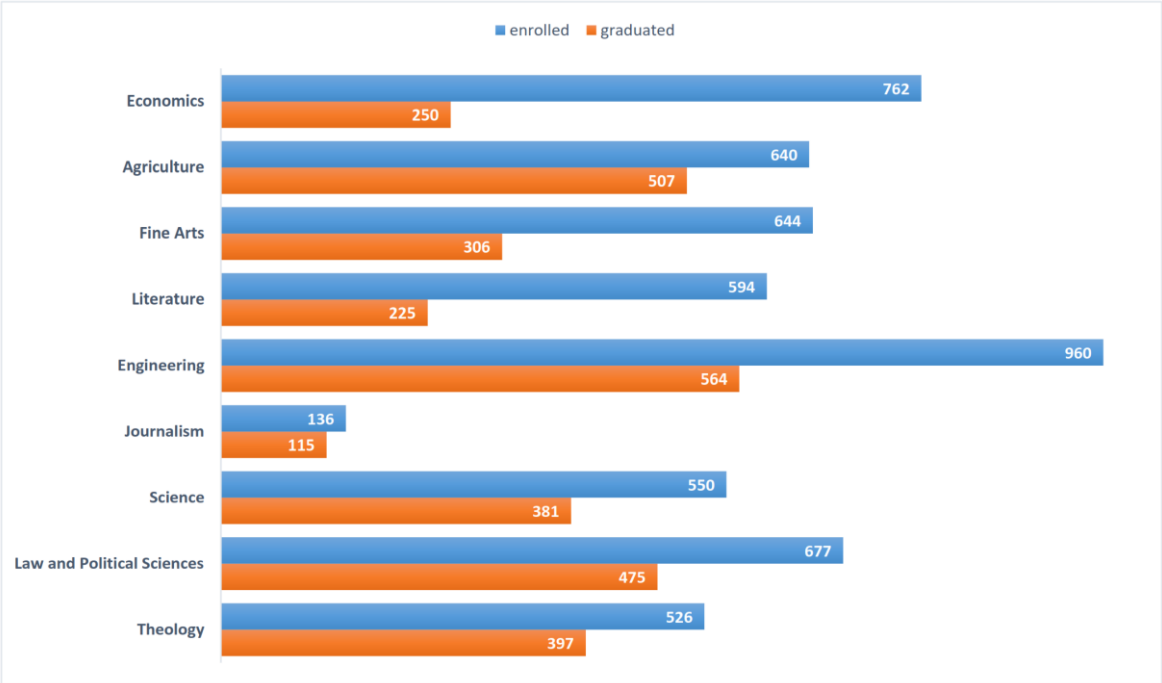

Figure 1-1. Kankor participants admitted into Herat University in 2009-2011 vs. their successful graduation.

High school students and graduates always consider success in *Kankor* as the most important factor influencing their work and social destiny, and families also consider the future of their children to be tied to the *Kankor* results. Competition in *Kankor* is very challenging because the increase in the number of high school graduates is explicitly disproportional to the rate of admission through the *Kankor*: only about one-fourth of applicants pass. Every year this gap inevitably grows larger, leaving more candidates frustrated. Due to the limited seating capacity at universities, and the limited time to complete the test, candidates suffer from anxiety. For Kankor participants to cope with this challenge, those who can afford it attend private Kankor preparation courses. It is very challenging to get admitted into the desired fields of study when about 250,000 high school graduates are applying for around 55,000 slots at the public universities. The former Deputy Minister of Higher Education, Prof. Osman Babury, believes this is due to uninformed selection of field of study or selection of prestigious fields of study and highly selective universities (Mandegar Daily Newspaper 2015). For MoHE to address this challenge, policymakers decided to introduce eligible Kankor participants to semi-higher education institutions. Despite this, the problem continued, and since 2013, a large portion of the eligible participants are introduced to private universities; with a discounted fee – if it is affordable for them.

To confirm the above-mentioned argumentation, the author conducted numerous surveys; the results of the surveys indicate that more than 75% of participants severely lack a proper understanding of *Kankor* examination and procedures. More than 80% believe unfamiliarity with *Kankor* and lack of preparation for it are two of the main reasons participants fail in the *Kankor*. Hence, for Kankor participants to do well, those who can afford it attend supportive courses. These supportive courses are usually more efficient and are run mostly by qualified teachers and university graduates. Additionally, more than 90% of Kankor candidates attended *practice Kankor tests* for self-assessment and to understand the *Kankor* test and procedures.

Proper choice of field of study by Kankor participants according to their skills and abilities is a critical issue which needs to be systematically addressed in Afghanistan. Afghan policymakers, advisors, pedagogists, and sociologists do not have the necessary tools to assess the condition methodologically using the advantages of data mining and recommender systems.

In the context of Afghanistan, there are large amounts of data available for mining purposes in the education domain, but the methods that the educational institutions use to store their data only enable them to achieve basic insights. Their main efforts are to generate (only) basic facts and figures (e.g., total number of students and teachers categorized by gender, location, and some other criteria). However, it turns out that these simple facts and figures do not help educational institutions to improve the educational settings. For instance, they cannot be used to predict suitable field of study for high school graduates, to identify first year university students who are at high risk of attrition, or to recommend proper courses for the university students.

In this thesis, the main goal is to propose and develop practical solutions to improve career decision-making systematically within the context of the existing structures through *'assessment test' to better understand the interest and tendency of Kankor applicants as well as to narrow down more than 100 disciplines; 'predictive models from data mining using the collected Kankor data'; and 'recommender systems approach using high school mark's to identify disciplines which are closest to Kankor candidates' knowledge set, skills and preferences*. The *e-Kankor* is proposed to simulate the *actual Kankor exam* allowing candidates to learn the importance of *timeliness*, *methods and techniques through exam cram* and to estimate the candidates' performance and score prior to taking the *actual Kankor* which is then used in the prediction models as unseen data, as is illustrated in **Figure 1-2**.

The model empowers high school advisors to assist the Kankor candidates and high school students by helping them to understand their academic options, determine the resources needed to reach their academic goals. Advising is a shared responsibility between an adviser and the student (Richard L. and Jessica G. 2014, 1:5), while the final responsibility for making decisions about his or her academic career rests with the student.

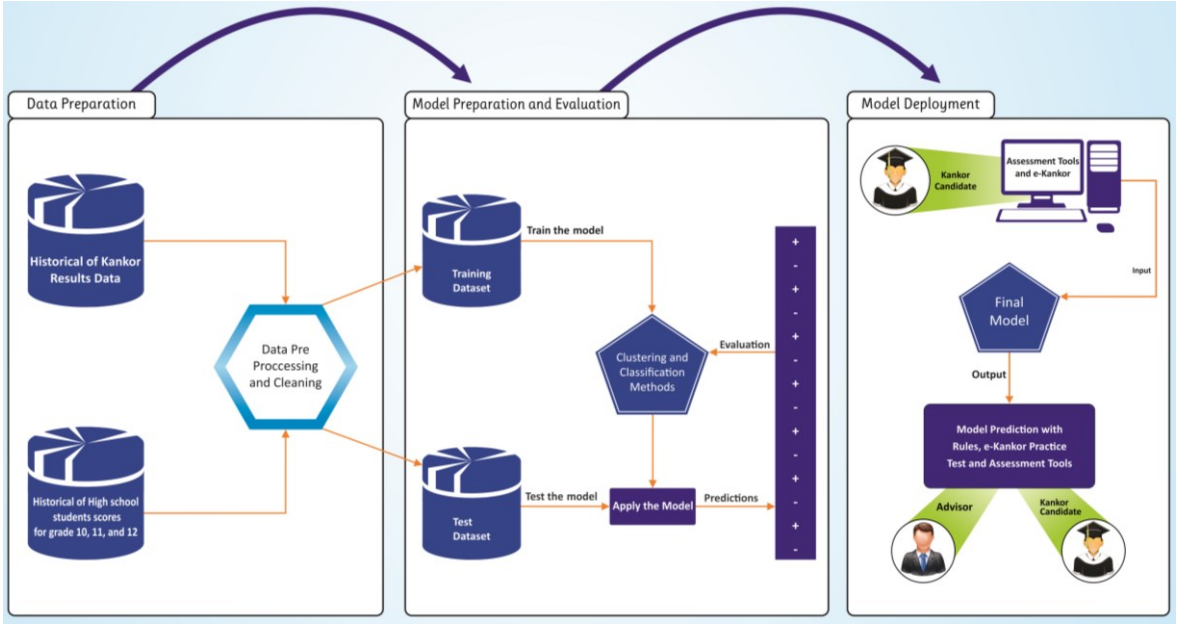

Figure 1-2. The primary prediction model using Kankor data and high school marks.

1.1. MOTIVATION

The author of this research, as a student and later a university faculty member in Afghanistan, observed many challenges in the education settings which could potentially be improved through technological solutions. The following three main observations led the author to choose this topic:

1. To enter higher education, high school graduates need to pass Kankor. They are permitted to take the Kankor twice in their lifetime. If they fail this examination, they will not be able to study in public universities anymore. In Kankor, candidates can opt for five fields of study from among more than a hundred fields of study that are not classified into streams such as sciences and social sciences. This leads participants to make uninformed choices without much knowledge of the requirements and challenges ahead of them. The result is academic failure or weak academic performance, and an increase in the attrition rate in higher education, with adverse economic and social effects for the students and the society. To gain insight into the ratio of admission vs. successful graduation, the author collected the numbers of Kankor participants admitted into different faculties (fields of study) at Herat University in 2009, 2010, and 2011, and the successful graduation rates for the same participants in 2013, 2014, and 2015 respectively. The data shows almost one-third of the participants admitted into Herat University could not complete their higher education studies – dropout, leaving university before completion, taking longer to complete, and change of discipline were among the reasons. In the existing setup, high schools lack any counseling system to advise students on their careers. Also, high school teachers and parents are generally not able to help candidates in this challenging process of making an informed decision concerning their academic careers.

2. In the existing structure the education system lacks student placement into sciences or social sciences in high schools. Furthermore, the current structure of Kankor in Afghanistan does not make it easier for participants to choose fields of study which best suit them. For example, participants who are interested in following a degree in Fine Arts fail in Kankor because half of Kankor questions are mathematics and sciences with higher scoring value. Also, Kankor fails to evaluate candidates' skills and competence for some fields of study. For example, in Computer Science, in addition to mathematics, English language and basic ICT skills are also required. Kankor lacks questions to evaluate these two skills which are required in Computer Science. Despite this, candidates can choose Computer Science through Kankor. As a result, there is great disparity in the knowledge and understanding of individual Computer Science students. The consequence is that Computer Science lecturers are not able to complete the syllabus. This heavily impacts students' learning outcomes and leads to high attrition rates and poor graduation. Based on the survey conducted among Computer Science students, around 90% of the 227 respondents did not have the basic skills and knowledge of programming, database, and operating systems at the time they were admitted into Computer Science at the university.
3. In the existing setup, some fields of study are more general (called faculty) and some are very specific (called department). Usually a faculty comprises of more than one departments. In the third year of the study program students are divided into departments. This division is another major challenge which needs to be addressed methodologically. For example, students at Computer Science faculty of Herat University are given choices to select from the three available departments (Information Database System, Software Engineering, and Network Communication). Making an informed decision about this is usually a major dilemma for students. In many cases, most students pick one department over the others. The result is imbalance in the ratio of students per department, thus making the division of students difficult or entirely impractical. Therefore, students are forced by the administration to choose a department based on the average scores in their first two years, rather than on their actual performance in courses most relevant to the departments. This general mechanism leads to other consequences, such as impacts on students learning outcomes and misallocation of talents. Another major issue that challenges the transparency of the system is that this dilemma also creates the conditions for nepotism. The same dilemma is present in other fields of study across the country.

The author of this research holds a degree in Computer Science and is interested in working with data and data mining. Also, as a faculty member of Herat university, the author is concerned about education in Afghanistan. Interestingly, Computer Science techniques such as data mining and recommendation methods are excellent tools for overcoming the above-mentioned challenges in the Afghan education system. An early solution to these challenges would help high school graduates to make more reasonable decisions regarding their academic careers. This allows students to overcome dilemmas in their choices and the administration officials to better manage talent allocations to appropriate fields of study.

The author picked Kankor as his research case study, with the aim of enabling academic advisors to provide guidance to candidates based on their skills and preferences, as well as improving the Kankor settings through data mining techniques. Moreover, one of the research contributions would lead to the institutionalization of counseling mechanisms as the solution, through advisors whose role is to guide and support students in making informed decisions regarding their academic careers.

1.2. The Problem Statement and Justification of the Research

The increase in enrollment in education and higher education lead to the growth of data in educational settings. In case of Afghanistan, there is no appropriate mechanism available to store and organize the data properly, and then to transform the data into valuable information which can be used to support Kankor participants in their career decision-making process.

On the one hand, there is a lack of academic counseling services inside and outside of educational institutions to advise students on choosing fields of study. On the other hand, no specialization studies are offered at high schools. These are aggravated by unconstructive and uninformed family influence on students. The result is that Kankor candidates do not select their fields of study by following the right criteria during the *Kankor* – they are likely underprepared when choosing a major and do not have sufficient knowledge of the requirements for the various fields of study. Thus, an uninformed selection of field of study leads to their academic failure, lack of interest in continuing higher education, and weak academic performance, with adverse economic and social effects for students and the society.

Proper choice of field of study by Kankor candidates, and their placement into majors according to their skills, abilities, and interests is a critical issue which needs to be addressed systematically in Afghanistan. Currently, Afghan policymakers and sociologists do not have the necessary tools and means to assess the condition methodologically using the advantages of science and technology.

The author's research goal is to design a model based on data mining and recommender systems techniques, and other assessment tools in the situation of a developing country to support advisors to help Kankor candidates make informed decisions concerning fields of study based on the high school score and performance, Kankor results data, and pedagogical feedback. The tool can also be used to direct school students into proper majors if Afghanistan decides in the future to create a specialization system at the high school level.

1.3. Hypothesis

The author's research aims are to verify and confirm the following hypotheses:

1. Through applying data mining techniques on the collected Kankor results data, it is possible to find patterns to help Kankor candidates make better decisions concerning fields of study. Using data mining techniques alongside expert pedagogical feedback, prediction of the proper field of study for Kankor candidates is achievable.

2. It is possible to shape and transform high school students' marks for grades 10, 11, and 12 in a matrix which can be used in collaborative filtering recommender systems methods.
3. It is also possible to adjust the same method and apply it to introduce suitable disciplines to high school students while specialization studies are offered at high schools.
4. Through applying data mining techniques, it is possible to institutionalize advisory organizations inside or outside of educational institutions to guide the students concerning their academic careers.
5. Data mining can make a positive contribution to improving education and higher education systems and assist them to predict the future and make proper decisions.
6. It is also possible to combine other methods such as collaborative filtering recommender systems and assessments tools together with data mining to design a better model with higher accuracy.
7. Rates of attrition or academic failure are higher for students who randomly picked their fields of study, than for students who selected their fields of study based on an informed decision and considering recommendation from experts.

1.4. RESEARCH METHODOLOGY

In this research, both quantitative and qualitative research methods are being used. The author collected historical Kankor data including Kankor candidates' demographic information and their Kankor scores and results. Also, the author collected historical data on High school students including their demographic information and their marks for grades 10, 11, and 12. Finally, the author conducted surveys and distributed questionnaires among university and high school instructors and staff to engage professionals from various fields and to have their feedback and suggested solutions. No such research and data collection using data mining has been conducted in the past in Afghanistan, particularly in the academic field. The following data have been collected by the author:

1. Kankor Results Data (KRD) from 2003-2015 containing more than 1.5 million records of Kankor participants' personal information, high school name and high school graduation year, Kankor score and result. The KRD is used for both descriptive and predictive analytics in chapters 5 and 6.
2. Detailed Kankor Results Data (DKRD) from 2004 to 2006 is a subset of KRD with around 120,000 records of Kankor participants. The DKRD contains additional data e.g. Kankor candidates' scores for Languages, Mathematics, Natural Sciences and Social Sciences as well as whether candidates were successfully accepted into their 1st, 2nd, 3rd, 4th, or 5th choice of field of study. The DKRD is used in chapters 5 and 6 to find out how positive a role Mathematics and Science scores play in a candidate's acceptance, and to assess whether a higher score always plays a role in a candidate's successful admission into his/her first choice of field of study.
3. School Performance Data (SPD) from 2011 – 2013 containing around 6,000 records of high school graduates' information as well as subject-wise scores for 10th, 11th, and 12th grades. The SPD, after matching with KRD, is used in chapters 5 and 6 of

this dissertation for analysis purposes to find out if high school performance plays a positive role in a candidate's successful admission into higher education.

4. Five different questionnaires were prepared and distributed to university and high school students and graduates as well as to heads of faculties, departments, and high schools:
    a. One questionnaire was distributed among heads of faculties, departments, and high schools to have their feedback about specialized study at high school and the impact of counseling on career choice. A total of 227 participants responded to the questionnaire – out of these 227, 44.5% were male and 55.5% female; and 63.4% and 36.6% were high school and university lecturers respectively.
    b. Two other questionnaires: one was circulated among students in public and private universities and a similar one was circulated among students in public and private high schools to know: 1-in which grade specialized studies should be offered and to what extent it will be helpful, 2-how much counseling on career choice positively impacts the proper selection of field of study and what approaches they suggest, 3-whether Kankor is a proper means to identify the candidates' skills and field of study, 4-how much knowledge they have/had about the Kankor test and procedures before taking it, 5-how much they knew about the field that they were admitted into, 6-on what basis they choose field of study in Kankor, 7-for studying which subjects they spent most of their time and why, 8- what methods they recommend for academic counseling for high school students, 9-how useful it is if a computer-based test is developed to support students in selecting a proper field of study. A total of 1,235 participants responded to the questionnaire – out of these 51% were university students and 49% were high school students; 36% were male and 64% were female; and 64% and 36% were from public and private universities and high schools respectively.
    c. Another similar questionnaire was conducted among public and private high school and university students to know 1-how relevant were specialized high school subjects to their field of study, 2-where they take preparation classes for Kankor and how helpful it is for succeeding in Kankor, 3-before Kankor, how much they are/were familiar with the Kankor test and methods, 4-what are the main reasons that students fail in the Kankor, 5-how useful will Kankor simulation be and whether they will use it. A total of 1,310 participants responded to the questionnaire – out of these 53.3% were male and 46.7% were female; and university and high school participants were 40% and 60% respectively.
    d. To find out about the impact of offering specialized studies at schools, another online survey was conducted by the author among Computer Science students. Most participants were university students who had either completed their studies or presently are students. A total of 227 participants responded to the questionnaire – out of these 79.7% and 20.3% were male and female; and 42.7% and 57.3% were current and former students

respectively; 80.2% were from public universities and 19.8% were from private universities.

5. Several interviews and questionnaires were conducted by the author among experts, mainly university professors from different disciplines, to know the role of high school subjects and their coefficient values for various fields of study. A total of 53 either participated in interviews or responded to the questionnaire.
6. This research also relies on the personal observations, findings, and analyses of the author as a student of computer science and later as an active and experienced lecturer and academic member at Herat University in the field of computer science.

To predict fields of study for students either prior to their entering into secondary or high school, or as they get into university, the author took into consideration similar research carried out by other researchers. Other researchers have looked at predicting fields of study for students by applying many approaches mainly from classification methods of data mining and have come up with diverse results. Of the researches based on data mining that have been conducted in the field of education, some examples will be cited.

Data mining is an inter-disciplinary field and the crossroads of various scientific disciplines including statistics, machine learning, database systems, data warehouses, and others. Hence, the author read many books on data mining and related domains, some of which will be cited. Furthermore, proper datasets are crucial for data mining. Prior to building a model, the data need to be pre-processed, cleansed and properly shaped. Hence, the author joined webinars and online courses on ETL (Extract, Load, and Transform) tools to transform and shape the data, and then to use the data for more effective and efficient explorations, visualization and analyses purposes through Business Intelligence and Data Mining Applications.

Finally, the author picked and followed CRISP-DM (Cross Industry Standard Process for Data Mining) methodology, and accordingly structured this research paper based on the 6 phases of CRSIP-DM. Per KDnuggets Poll in 2007 and 2014 (Piatetsky and KDnuggets 2014), CRISP-DM is the most popular methodology for analytics, data mining, and data science projects. It is the leading methodology used by industry data miners.

1.5. Research Contribution

The concept of data mining in educational and other institutions in Afghanistan is quite new. This research paper will be a new initiative in the situation of Afghanistan. If executed completely, in its first phase the proposed and develop models will support an average of 250,000 high school graduates who will attend *Kankor* per year. Moreover, it will help support over 1,000,000 students who are currently in high school who will take *Kankor* over the next 3 years. The system can be used to help the future generations as well.

Statistics and analyses illustrate that a candidate could perform well, get a high score, and still not get admitted into his/her field(s) of choice. Therefore, uninformed choice of field of study can lead to unsuccessful admission. The outcome of this research paper will be a useful guide and roadmap for high school student advisors to support such candidates in determining proper fields of study through a more methodological decision-making process.

The data and survey show almost one-third of the participants admitted into Herat University could not complete their higher education studies. The reason could be lack of interest and motivation, or random selection/admission to the fields of study. The outcome of this research could act as a valuable framework and guideline for the MoHE to enrich the *Kankor* settings (the methods and the process) efficiently, and for the MoE to institutionalize the provision of academic advice to students to guide them in this critical decision.

This research will be an informative and valid English resource for academic and research studies. It is the first of its kind in Afghanistan and so other researchers may base their work upon it. This work opens the door to using data mining to do research about Afghanistan education systems. On the other hand, this will be a useful guide for international donors that are interested in investing in or supporting Afghanistan education systems.

In Afghanistan and other countries which hold *Kankor*, this exam is decisively important for high school graduates who intend to continue into higher education. Choosing a Major through the *Kankor* is an educational challenge with broad social impacts. The fear and stress of *Kankor* is widespread among high school graduates. Moreover, families and the society demand that students do well in *Kankor*, which only adds to the pressure the exam candidates feel. This is the case not only in Afghanistan, but also in other countries including: Iran, Turkey, China, and Japan. Hence, the contribution of this research concept is not just limited to Afghanistan, but can be generalized and adopted in the other countries with *Kankor* examination. Finally, even though not every country has a *Kankor* exam, the results of this research will have the potential to be generalized and applied to other similar contexts.

It is worth mentioning that the author's research differs from similar research conducted in other countries in four important ways: 1) the studies mentioned above dealt with simple cases and scenarios i.e. two to three class values (options) while in Afghanistan there is a complicated system of over 100 choices for fields of study with overlapping admission score. 2) school students in those countries are offered specialized studies at secondary or high school level. 3) academic advice is available to students either in schools or through other offices in the country, which guides the students concerning their academic career. But in the context of Afghanistan none of the above is true. Therefore, more research is required. 4) There are more than 100 fields of study options for the dependent variable (outcome variable). The author applied clustering and descriptive techniques to classify the fields of study systematically prior to developing and training the model using classification methods.

1.6. Further Research Outcome and Contribution

Since no other such studies have been carried out in the past in Afghanistan, this research paper lays the foundation for improving educational settings through applying data mining techniques supporting high school advisors to guide the Kankor candidates on their academic career of choosing proper fields of study. Hence, it opens up whole new range of opportunities and other researchers can base their research upon this study to support MoE in applying data mining techniques to automate division of high school students into proper majors as soon as the law is in place.

Presently, even after passing the *Kankor*, choosing the area of concentration is yet another challenge for students and university administrations. The students who pass *Kankor*, some are directly admitted into departments while others are admitted into faculties. A faculty comprises of one or more departments, while departments are specific areas of concentration. The candidates introduced both to faculties and departments must take the same general courses for the first two years. In the third year, those admitted into departments will continue in their departments, while those admitted into faculties will have to choose a concentration/department. Deciding about this is usually a major dilemma for students:

- In many cases, the numbers of available slots in different departments differ significantly, thus making the division of students into departments very difficult or entirely impractical. As a result, students cannot choose their department. Rather, university officials decide which students can go to which departments based on their average marks for the first two years and not on their performance in courses most relevant to the departments.
- This general mechanism leads to other consequences, such as impacts on students' learning outcomes and misallocation of talents.
- One further major issue that is a challenge for the transparency of the system is that this dilemma creates the conditions for nepotism.

The results from questionnaires and surveys carried out by the author indicate that it is very hard for almost all Kankor candidates to choose an appropriate major. Even students with the option to choose an area of concentration in their third year of university have little idea of what to choose. Therefore, the author suggests that students who have passed the *Kankor* should choose their area of concentration after completion of the first two years of their higher education program. This online article (Freedman 2013) also proposes the same. This does not mean the candidates should not choose a Major through *Kankor* – as too much delaying the choice of a major may have negative consequences (Richard L. and Jessica G. 2014, 1:55). They better choose faculties instead of departments (for example they should choose Faculty of Engineering or Faculty of Computer Science instead of the Department of Civil Engineering or the Department of Software Engineering). Therefore, the author of this dissertation proposes to the MoHE and the Kankor Committee to list only faculties on the *Kankor*, which then the candidates can choose as their major.

This study, through its data mining techniques, will provide a suitable platform for future researchers to further investigate the application of a performance-based approach rather than an average-score-based approach. This modality will allow students to overcome dilemmas in their choices, and the administration officials to better manage talents and allocations to relevant departments upon completion of general courses for the first two years of their program.

The contribution of this research is not limited (only) to public educational institutions, it also benefits private institutions of higher education in Afghanistan. Currently there are more than 109 private institutions of higher education across Afghanistan, with a total number of 128,735 applications. The outcome of this research can also lead to improved decision-

making processes for the students and administrative officials of private higher education institutions in Afghanistan.

1.7. Dissertation Structure

This dissertation comprises seven chapters and is structured according to the CRSIP-DM methodology. Chapter one of the dissertation covers the introduction, motivation, problem statement and justification of the research. Moreover, this chapter outlines the hypothesis, research methodology and contribution of the research. Per the CRISP-DM methodology terms, this chapter, chapter 2 and chapter 3 revolve around the fundamentals of 'business understanding' phase; chapter 4 and 5 rotate around the 'data understanding' and 'data preparation' phases, and finally chapter 6 follows the 'modelling' and 'evaluation' phases.

Chapter 2: This chapter uses comparative methodology to draw lessons from similar educational contexts in the region (Iran, India, Turkey and China) about how Kankor, as a standardized testing system, could be enhanced so that it better serves to evaluate test takers' competencies. Furthermore, this research intends to outline how advising mechanisms can either be created or improved to guide applicants in choosing a field of study which better matches their needs and plans.

Chapter 3: In this chapter, the concepts and applications of data science (data mining, educational data mining, and recommender systems) are described. The chapter also outlines how such concepts can be implemented in the deployment phase of this research. The research provides scenarios and discusses challenges regarding how these data-driven solutions can be used as a resource in Afghanistan's education system. Finally, the research focuses on equipping high school advisors with proper tools, so they can provide suitable guidance to Kankor candidates concerning their academic careers.

Chapter 4: Data preprocessing, cleansing and transformation are very critical and useful prior to any data analysis activities. This chapter discusses missing values, data inconsistencies and other issues and provides numerous methods and algorithms to systematically autofill missing values, fix inconsistencies (such as Unicode variations, extraneous spaces, variations of candidate and province names, duplicate instances), consolidate all the datasets into a master dataset, and audit Kankor dataset with high school dataset and other inconsistencies. These are essential, as they lead to transformation of data into an organized structure and to generation of accurate, valuable and detailed insights. As these steps lead to 'data understanding' and 'data preparation', they may be used by other institutions and organizations in Afghanistan, particularly those active in the education sector.

Chapter 5: To have a better picture of Kankor and high school data in Afghanistan, exploration and visualization techniques are essential before using data mining techniques and recommender systems to support students and educational institutions. The data is assessed and analyzed from numerous aspects including but not limited to: candidates' performance in Kankor, the performance of public and private high schools, candidates' choices of field of study, the contribution of high school scores and several other common

parameters. The broader the understanding of such complex datasets, the better policymakers are placed to reshape education systems as needed.

Chapter 6: This chapter is dedicated to all activities related to 'modelling' and 'evaluation'. It discusses the models developed as part of this research to support policymakers in shaping the education system and empower academic advisors to provide efficient guidance to candidates, so candidates can choose fields of study that closely match their competences, knowledge, and interests. One such method is an "assessment test" and a few other methods which help narrow down the more than 100 available disciplines to ones that best fit candidates' knowledge, tendencies and interests. Two other models 'prediction models of data mining using Kankor data' and 'recommendation approach using high school marks' to identify disciplines that closely match the candidates' competences and skills are also explained in detail. As the main aim of this research is to provide guidance to candidates prior to their actual Kankor exam, a suitable model is designed in order to identify the candidates' scores and performance before they take the actual Kankor exam. One of the appropriate approaches to gauge the candidates scores is to simulate the actual Kankor based on the MoHE rules and regulations[1]. The scores are then used in the 'prediction models' as input data.

Chapter 7: The final chapter covers conclusions, and future work and opportunities. It discusses the accomplishments but also the limitations of the research due to lack of some key data. Finally, this research opens up several opportunities for further investigations a few of which are discussed in this chapter.

[1] You may refer to chapter 2 for details on MoHE's rules and regulations.

Chapter 2

Chapter 2. EDUCATION AND KANKOR IN AFGHANISTAN: A COMPARATIVE RESEARCH

Education systems vary greatly, in entrance policies, teaching methodologies from country to country, even from institutions to institutions. Afghanistan's general education system (elementary, grades 1 to 6; intermediate, grades 7 to 9; and high school, grades 10 to 12) compares closely with the education systems in various countries in the region, such as, Iran, Turkey, China and others. Moreover, *Kankor* is commonly used in Afghanistan and the regional countries.

This chapter of this thesis uses comparative methodology to *draw lessons* from similar contexts in the region about how *Kankor,* as a *standardized* testing system, could be enhanced so that it better serves to evaluate test takers' competencies. Furthermore, the research intends to outline how advising mechanisms can either be created or improved to guide applicants in choosing a field of study which better matches their needs and plans.

At the outset, this chapter provides a review of the education systems in Iran, Turkey and China, and then it proceeds to offer detailed insights into the education system of Afghanistan. It subsequently outlines the educational challenges which the identified countries face with. The research also conducts a needs assessment to understand the specifics of admission into Afghanistan's higher education in comparison with the identified countries in sufficient depth. This step is taken prior to applying data mining techniques, to improve the long-term career decision-making processes.

2.1. EDUCATION IN IRAN

Education and higher education in Iran is centralized. The education is divided into K-12 (comprises the sum of primary, and lower and higher secondary education prior to university) which is supervised by the Ministry of Education. All institutions of higher education, except medical institutions, are under the supervision of the Ministry of Science and Technology. Medical universities are supervised by the Ministry of Health and Medical Education. The data on Iran's education is mainly based on (WES Staff 2017; Sedgwick 2000; WES Staff 2013; Asadzadeh 2015).

2.1.1. Basic Education

Since 2012, basic education, which lasts until grade 9, is compulsory and is free of charge at public schools. It is divided into a six-year elementary education cycle, and a three-year lower secondary cycle. Assessment is carried out to determine students' placement in one of the following three streams at the upper secondary level: academic, technical, and vocational. The students who fail must repeat the year and may take the examination again the following year. If students fail a second time, they must either undertake basic vocational

training or seek employment. Successful students are awarded a Certificate of General Education. Depending on grades achieved in the relevant subjects at the end of the lower secondary cycle, students are eligible to continue their education in the academic, technical or vocational branches of the upper secondary cycle.

2.1.2. Upper Secondary

The upper secondary education is three years in length, grade 10 to 12. It is free of charge at public schools, but It is not compulsory. At this level, students are segmented into the following three streams: academic, technical, and vocational. A student's stream is dependent mainly on his/her examination results at the end of the lower secondary cycle. The academic stream has traditionally been the most popular.

The Academic and Technical Streams: During the first two years, students both in the academic and technical streams follow a common curriculum with the third year focusing on a specialized curriculum.

- Students in the academic stream study one of four subject areas in their final year: Humanities & literature, mathematics & physics, experimental sciences, or Islamic theology.
- Students in the technical stream follow one of three specializations: Technical (industry), business & vocational (service industry), or agriculture.

Upon successful completion of studies and after passing the national examination, students from the academic and technical streams are awarded a certificate of completion of secondary school studies. The graduates either continue to a final pre-university year of education or opt to enter the workforce.

The Vocational Stream: It is more practice-oriented and leads to the award of a skills certificate in the trade/profession studied. Training for skilled and/or semi-skilled employment is provided in 400 areas of specialization. Some vocational students at this level enroll in five-year integrated associate diploma programs at technical institutes.

2.1.3. Pre-University

The pre-university centers are administered by the Ministry of Education. It is an introductory year for students who plan to take the *Kankor*, required for university admission. During the pre-university year, students specialize in a specific field of study (math, experimental sciences, humanities, arts, or Islamic culture). They are graded by continuous assessment and by final examination (accounting for 75 percent of the overall grade). Successful students are awarded the Pre-University Certificate and entitled to sit for the *Kankor* for university admission. The pre-university year originally evolved prior to 2012 and will eventually be incorporated into the new 12-year 6+3+3 system, so that graduates can sit for entrance examinations without first completing an additional year of pre-university study.

2.1.4. Kankor in Iran

The National Organization of Educational Testing (NOET) under the Ministry of Science, Research, and Technology (MSRT), oversees *Kankor*, which is the exam for admission into public universities. The *Kankor*, a 4.5-hour Multiple-choice comprehensive examination, takes place in June every year. The *Kankor* tests the candidates' knowledge in all the subjects taught in high schools, and includes: Persian language and literature, history, Islamic studies, a foreign language, and mathematics. It takes place as five separate exams at the same time, for five different groups: Mathematical Sciences and Technical Sciences, Sciences, Humanities, Foreign Languages and Arts. The *Kankor* is very difficult so that candidates normally spend a year preparing for it. The candidates who fail can repeat the test in the following years until they pass it.

The top candidates usually go into the prestigious fields of study (engineering and medical), and highly selective universities such as: University of Tehran, Amirkabir University of Technology, Sharif University of Technology, etc. Graduates of such universities have a better chance of securing the increasingly limited jobs in the prestigious professions. Many private universities also use *Kankor* for admission purposes.

To solve the problems related to student selection, the Ministry of Science and Higher Education of Iran established a center in 1969 to develop and implement the rules of admission into higher education in collaboration with universities. As the number of applicants for higher education increased and the universities and institutions related to higher education expanded, it was necessary to establish a larger organization specifically tasked with student selection. Therefore, the NOET with three Deputies – Technical and Research, Executive, and Supervision on Universities Affairs – was approved in 1975 and continues to operate till present. The NOET holds various tests to select students for public higher education institutions, including: the *Kankor* as well as a separate *Kankor* exam for entry into national Master and Doctoral graduate programs, and other exams.

In Iran, as in Afghanistan, China and other countries where the university entrance exam is the only criterion for student admission into universities, the number of seats and resources are limited for the many enthusiastic candidates seeking access to higher education. Therefore, the *Kankor* has gone through many phases to improve the situations and settings. A quota system was introduced, a law was passed to help handicapped and volunteer veterans to enter universities, an additional criterion for student selection was introduced to localize the student population, giving priority to candidates who applied to study in their home provinces. Despite attempts made in recent years to reform university selection criteria, the *Kankor* remains an impediment to equal education access.

According to the current legislation, *Kankor* is to be gradually eliminated, and other criteria to be considered for admission into university. One option under consideration is the use of a cumulative grade point average (GPA) of the final three years of high school to admit students.

2.1.5. Higher Education

Admission into university is through *Kankor* and is administered by the Organization for Assessment of the State Education in the Ministry of Science, Research and Technology. The higher education system generally includes the following four prioritized educational degrees: Technicians (2 years), Bachelor's Program (4 years), Master's Program (2 years), and Doctoral Program (4 years).

2.1.6. Summary of Similarities and Differences: Iran's Case

From the literature, it can be seen that Iran's education system in various aspects is similar to Afghanistan's education system – the education and higher education systems are centralized and general education is comprised of K-12 system. Likewise, *Kankor* is the only criterion for admission into universities.

However, the two nations' education systems differ widely in certain areas. For instance, in Iran, at the end of the lower secondary cycle, students are segmented into one of the following three streams at the upper-secondary level: academic, technical and or vocational. Moreover, pre-university/preparatory year is offered for students who plan to take *Kankor*. Finally, all the fields of study are classified into main streams. These opportunities are nonexistent in Afghanistan's context as the existing policy seems to possess deficiencies resulting from lack of appropriate studies.

2.2. EDUCATION IN TURKEY

The Ministry of National Education is responsible for the supervision of public and private education systems, agreements and authorizations under a national curriculum, and prepares and approves textbooks and teaching aids. Public and private universities are directly recognized and overseen by the Council of Higher Education (CoHE). The CoHE is an autonomous institution which is responsible for the planning, coordination and governance of higher education system in Turkey in accordance with the Turkish Constitution and the Higher Education Laws. The data on Turkey's education system is mainly based on (Mihael 2014; Asadzadeh 2014; WES Staff 2003; Kevin and WES Staff 2017).

2.2.1. Compulsory Education

As of 2012, compulsory education lasts 12 years, begins at age 5, and is free of charge at public schools. It includes: elementary, lower secondary, and upper secondary (4+4+4). Upon completion of lower secondary, students can study at general high schools, or vocational & technical high schools. Admission to general high schools is based on the TEOG (Transition from Primary to Secondary Education) examination at the end of grade 8. Those who do not earn sufficiently high scores to get into their school of choice are assigned to schools nearest to their residence, including vocational schools, which generally do not require the TEOG exam for admission.

2.2.2. Teacher Education

Teachers at the pre-school, elementary, and secondary levels are required to have a four-year bachelor's degree to teach. Teacher training curricula incorporate specialization subjects, pedagogical subjects, practice teaching and a teacher certification. Students can earn either a bachelor's degree in education, or complete postgraduate training following a bachelor's degree in a non-teaching discipline. A one-year professional proficiency certificate entitles holders to teach English and selected subjects at the elementary and pre-school levels, whereas elementary and secondary school teachers in subjects like history, geography, mathematics or physics are required to complete a master's degree.

2.2.3. Kankor in Turkey

In addition to Afghanistan, Iran, and China, in Turkey too, students must pass a difficult *Kankor* exam to enter higher education. Every year more than two million students attend *Kankor*, despite an increasing number of private higher education institutions in recent years. The statistics of admission into public universities is about 25% of all applicants.

The *Kankor* includes questions on high school mathematics, physics, chemistry, geography and biology courses, and each applicant can choose a field of study of any academic discipline. Admission into postgraduate education (graduate school), does not take place through *Kankor* but each university chooses the qualified students from among university graduates in accordance with their criteria and educational standards.

2.2.4. Higher Education

Admission into universities is based on students' average high school grades, and the results from the centrally administered YGS-LYS examinations. Some universities may have additional entrance examinations. Universities offer the following degree programs: Bachelor's degree (generally takes 4 years, excepts for one-tier, long-cycle degrees which include: Dentistry, Veterinary, Pharmacy, and Medicine), Master's degree (2 years; non-thesis master's programs generally take 1.5 years), Doctoral degree (4 years, or the duration depends on the program and university). In addition, the Turkish system also includes a two-year associate degree, which is mostly a non-university qualification in technical and vocational fields designed to prepare students for entry into the workforce.

2.2.5. Summary of Similarities and Differences: Turkey's Case

In Turkey, the one-tier, long-cycle degrees (Dentistry, Veterinary, Pharmacy, and Medicine) grant master's degree, while, in Afghanistan they do not. The 12-year compulsory education in Turkey benefits children and increases the literacy rate in the country whereas in Afghanistan only primary and lower secondary education are compulsory. However, the existence of the policy in Afghanistan does not guarantee that this law is abided by, particularly due to socio-economic obstacles. Also, in the Turkish education system, segmentation of students into streams, after assessment and evaluation, supports students in making the right school choices according to their marks and their ideals. Finally, teachers at the pre-school, elementary, and secondary levels are required to have a four-year

bachelor's degree to teach. While, in Afghanistan, per statistics provided by the MoE, 80% of the country's 165,000 school teachers had the equivalent of a high school education or did not complete their higher education studies.

2.3. Education in China

At the national level, the Ministry of Education is the central administration responsible for formulating education policies. Provincial education departments manage the local policy development and implementation mainly under the direct authority of the Ministry of Education and other central government authorities. Local education authorities are primarily responsible for supervising elementary education. The data on China's education system is mainly based on (Gu 2016b; WES Staff 2011; Gu 2016a; Michael 2017, 2016; Sedgwick and Chen 2002).

2.3.1. Basic Education

Basic education is compulsory and comprises of the 6+3 system – six years of elementary school followed by three years of lower secondary school, while some provinces use a 5+4 system. It is free at public schools, and all the children at age of six must attend and complete it. Upon successful completion, students can continue to study at senior secondary level if they pass the senior secondary entrance examination.

2.3.2. Senior secondary Education

The senior secondary education is mainly divided into the following two schools: Regular or academic high schools, three years; and vocational and technical high schools, three to four years.

Students in Regular high schools typically must choose between the science stream and the art stream of study prior to the start of grade 11. To obtain a high school diploma, students must meet a minimum requirement of 144 credits, and pass either the General Examination for High School Students (GEHSS) or the Academic Proficiency Test (APT). Passing the GEHSS or APT is also the prerequisite for being allowed to sit at *Kankor* (commonly known as Gaokao).

The vocational and technical schools offer programs in engineering, agriculture, forestry, medicine, finance, textiles, tailoring, telecommunications and electronics. Approximately 40 percent of all senior middle school students attend these institutions.

2.3.3. Kankor in China

Admission to all higher education institutions in China requires a passing score on the highly competitive National Entrance Examination (commonly known as Gaokao), administered by the Ministry of Education.

The exam lasts about nine hours over a period of two days, depending on the province. Chinese literature, Mathematics, and English language (in most provinces) are required for all students. In addition, students must choose between two streams, the social-science-oriented area, and the natural-science-oriented area. Students who choose the former take an

additional exam on history, politics and geography, while those who choose the latter take an additional exam on physics, chemistry and biology.

For parents and candidates obtaining admission into university is a dream. The Chinese perspective is that people who graduate from universities have a higher chance of gaining government jobs – and so have lots of possibilities and opportunities. Every year more than 10 million candidates participate in *Kankor*, but there are only seven million slots for admission to higher education institutions.

2.3.4. Higher Education

Higher education, structurally, is divided into two sectors: Regular Higher Education and Adult Higher Education. Ninety percent of China's Higher Education Institutions (2,553) are in the Regular Higher Education sector in which over 70% of all undergraduate students are enrolled. The Adult Higher Education programs, in which just under 30% of all undergraduate students are enrolled, follow the curriculum offered by the Regular institutions, but the teaching format is more flexible and diverse, including distance-learning and part-time study. Higher Education Institutions offer bachelors, master and doctoral degree programs. It is worth mentioning that not all these Higher Education Institutions, even in the regular sector, offer degrees; many offer only graduation certificates. The country offers non-degree programs as well.

Admission into a Regular Higher Education institution depends on high school graduation and Gaokao scores, while admission into Adult Higher Education institutions is based on the National Adult College Entrance Examination (also known as 'the adult Gaokao'). Applicants are expected to have academic skills equal to high school graduates at the time of the examination. However, a high school diploma is not required for enrollment in Adult Higher Education programs.

2.3.5. Summary of Similarities and Differences: China's Case

The K-12 education program as well as basic education being compulsory is similar to the Afghanistan education system. Likewise, students' scores in the *Kankor* is used for admission into higher education institutions. The higher the score a student obtains, the higher the likelihood for the student to be enrolled in a prestigious field of study, or highly selective university.

The two systems, on the other hand, also have important differences. In China, all children must complete compulsory basic education, but in Afghanistan the education law is not always enforced/implemented. Finally, in China, students in regular high schools typically choose between the science stream and the art stream of study prior to the start of 11th grade, whereas in Afghanistan, no segmentation policy to different streams exists.

2.4. Education in Afghanistan

In Afghanistan education centers in which modern sciences are taught are called School, and education centers in which Islamic and religious studies are taught are called Madrasa. The documents related to the history of education in Afghanistan show that modern-

style/western-style education in Afghanistan began in 1875 with the establishment of schools for civil and military purposes for the royal family. Subsequently modern schools were re-introduced with the founding of Habibiya School with elementary, lower and higher secondary education levels. The development of many general and vocational schools in Kabul and other major cities led to the formation of Ministry of Education (MoE) in 1922. Next, due to the expansion of higher education institutions in 1970s, the Ministry of Higher Education (MoHE) was formed in 1977 to consolidate the country's institutions of higher education.

Afghanistan's MoE, since formation of modern education and until now, is the central administrative body to manage general education (primary, lower secondary and higher secondary), technical and vocational, and Islamic education across the country.

MoHE is in charge of regulating, expanding and developing Afghanistan's institutions of higher education and sets rules and regulations for assuring the quality of both public and private universities. It is responsible for developing the capacity of lecturers and for establishing a national higher education curriculum as well as special education programs, in-service training, and promoting further education for university academicians. Higher education is provided by public and private higher education institutions. Admission depends on the results of *Kankor*. Universities and other higher education institutions generally offer bachelor and master's degree programs. Recently a limited number of doctoral studies are also offered.

The establishment of the new democracy in Afghanistan in 2001-2002, received substantial international aid to restore the education system through a nationwide rebuilding process. Despite obstacles, numerous private and public educational institutions were established across the country.

Significant progress has been made since the fall of the Taliban in 2001. The children are appreciating easier access to 15,081 general education schools. The increase in enrollment from 1 million pupils, almost all boys to today's enrolment of about 9 million students, 3.5 million of whom are girls, stands out as one of the most significant achievements in terms of quantity. This comprises all students in basic, general, Islamic studies and vocational education. The number of teachers has also significantly increased from 110,000 in 2007 to more than 200,000 in 2015 of which 65,912 are females. The annual number of secondary graduates has risen from about 10,000 in 2001 to more than 266,000 in 2013 and it is estimated to reach 320,000 in 2015. Enrolment at semi higher education institutions has risen from less than 8,000 in 2001 to 461,735 in 2015. Investments have been made to improve quality and relevance of education (CSO 2014). In a nutshell, education growth since 1950 to 2014 - 2015 is reflected in the following figure.

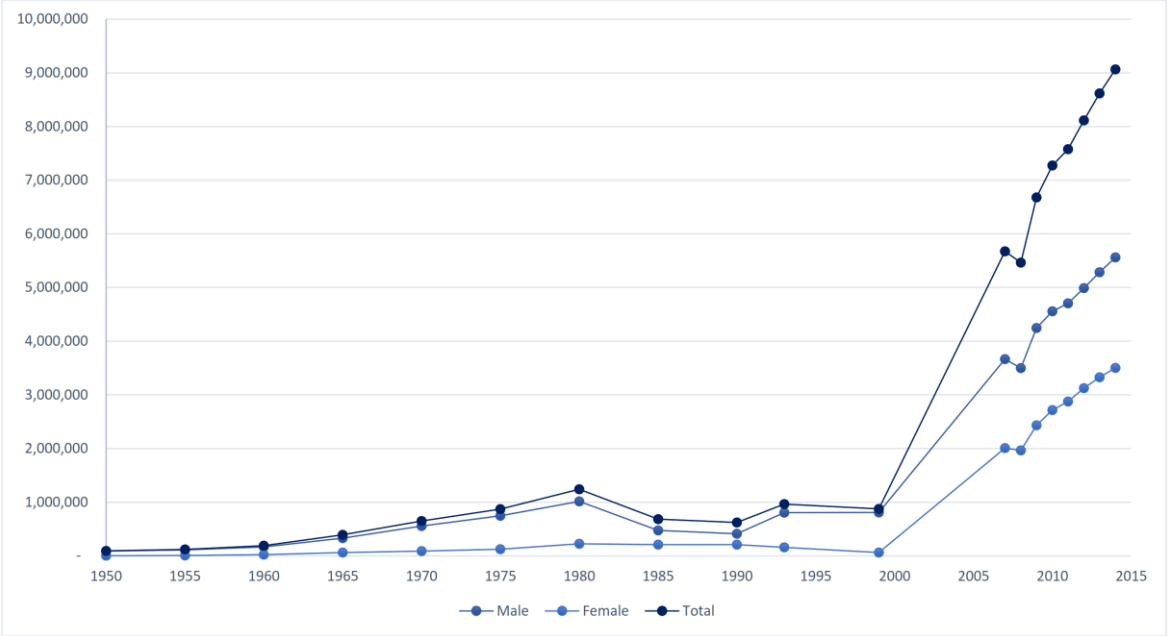

Figure 2-1. Education growth since 1950 to 2014–2015. Since 2007 students' enrollments in teacher training, Islamic education, technical & vocational institutions are also included.

During this period, there has been significant progress and development in Higher education sector. For example, the growth in enrolment from 7,881 students in 2001, to more than 256,140 students in 2014-2015, out of which, 52,832 students are girls. The number of university teachers has also considerably increased from 2,408 in 2007 to 11,381 in 2014-2015, out of which 1,369 are females. The number of public university has risen from 22 in 2007 to 31 in 2014-15. Since MoHE have limited capacity to engage all the Kankor participants, the MoHE as part of its reformed policy of inclusiveness, created private educational institutions where 109 private universities currently fill the gap not met by the public universities. This students and teacher facts and figures cover both public and private universities (CSO 2014). The facts in the next section are based on data generated through EMIS per personal communication with MoE, and Afghanistan Central Statistics Organizations website (http://cso.gov.af/en).

2.4.1. Education System: Structure and Requirements

Afghanistan's education system did not experience massive transformation and changes in terms of structure and style since its early foundation. In fact, the basic and general education with only few changes in some years remained the same. At present, the education system is on the same lines and foundation that was laid in the early years of the twentieth century with only few changes.

Presently, the education system of Afghanistan is divided in six different levels: pre-school (with no serious attention and focus on that), primary school (grades 1 to 6), lower secondary education (grades 7 to 9), higher secondary education (grades 10 to 12), a parallel system of Islamic education, technical and vocational training education, teacher training, and Higher education. Higher education is provided by public universities, and private higher educational institutions. The education is mainly provided in Pashtu and Dari.

The MoE and MoHE share responsibility for the entire education system across the country. The MoE is responsible for basic and secondary education, including teacher training programs and vocational education. On the other hand, the MoHE is responsible for higher education. In a nutshell, enrollment in general education, Islamic education, Teacher training, and Technical & vocational institutes in 2014-2015 is reflected in the following figure.

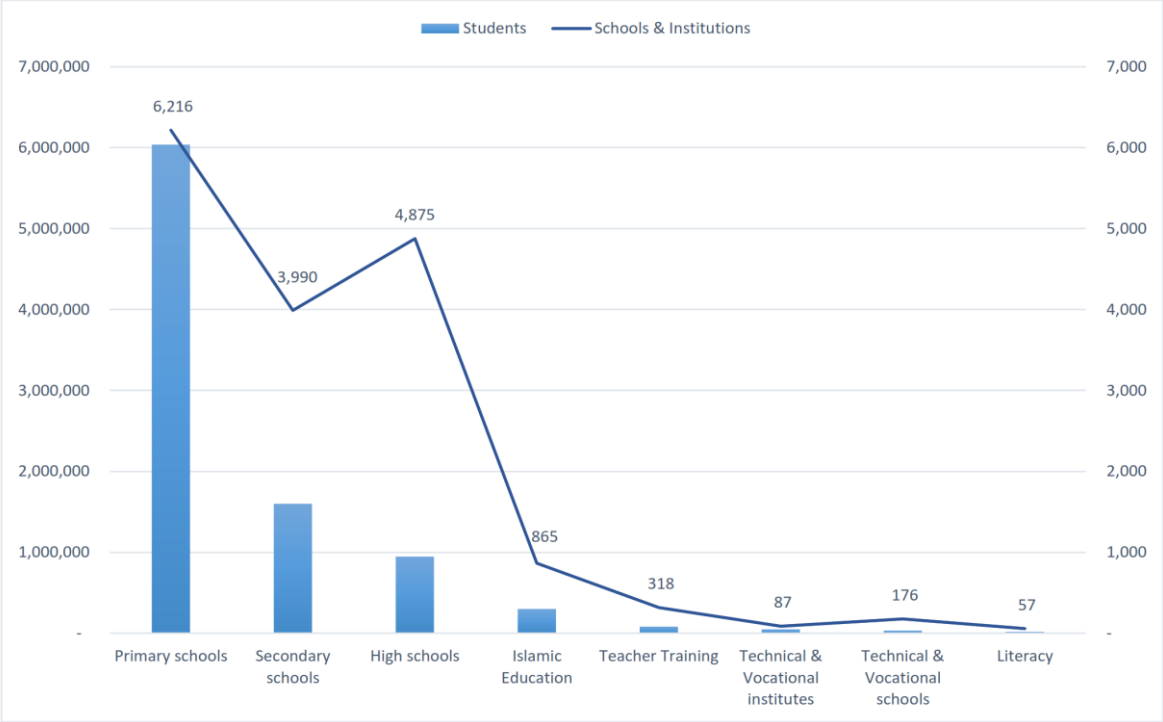

Figure 2-2. Enrollment in general education, Islamic education, Teacher training, and Technical & vocational institutes in 2014-2015.

Pre-school Education

Pre-school education is optional (neither free and nor compulsory) and covers children (3 to 6 years) who have not yet reached the age of entry to primary education. The aims are to foster spiritual development of children, and to prepare them for primary education by reinforcing cognitive/intellectual skills.

Compulsory/Basic Education

Based on the constitutional right in Afghanistan, basic (intermediate) education is compulsory and free of charge at all public schools and requires starting at age seven. All children must complete nine years of compulsory elementary and secondary education. As of 2015 academic year, 6,037,663 and 1,599,485 children were enrolled in 6,216 elementary schools and 3,990 secondary high schools nationwide respectively. The proportion of the figures to the total population is 25.3%.

There are two stages in compulsory education: elementary school and secondary high school. All the provinces across the country use a 6+3 system – six years of elementary school followed by three years of secondary high school. The curriculum comprises and covers subjects such as: Islamic studies, first language (Dari or Pashtu, depending on the region), a

secondary language (Dari or Pashtu, depending on the region), foreign languages, arts, mathematics, natural sciences, social studies, history, geography, and physical education. Upon completion of secondary education (grade 9), students may opt to pursue technical and secondary vocational education, or higher secondary education. Most of the students pursue the latter.

Higher Secondary Education

Higher secondary education is organized into two three-year cycles: The first cycle, lower secondary education (grades 7 to 9), which is compulsory, and the second cycle, higher secondary education (grades 10 to 12). The students have a choice between continuing with an academic path for 3 years that could perhaps lead on to university, or study subjects such as applied agriculture, aeronautics, arts, commerce and teacher training instead. Most students pursue the former. Students are awarded a 12-grade graduation certificate upon successful completion, formerly Baccalaureate certificate. There were 946,558 students attending 4,875 high schools in 2015 academic year, and its proportion to total population was 3.1%.

Technical and Vocational Education

Technical and vocational educational programs are delivered in formal education led by MoE, which train individuals for business and vocational fields to develop and expand knowledge and skills through theoretical and practical training, such that they are prepared for tertiary and higher education. Technical and vocational education is composed of technical and vocational institutes, and technical and vocational schools.

Successful completion of lower secondary education grants admission to technical and vocational schools. The program lasts 3 years and upon successful completion a diploma is awarded to graduates that grant them permission to enroll for a *specialized Kankor* or enter the job market.

On the other hand, high school graduates get admission to technical and vocational institutes through *Kankor*. The program lasts 2 years and upon successful completion, students are awarded an associate degree also called vocational education certificate. Students with equal or greater than 80% overall score can participate in another *specialized Kankor* and upon successful acceptance get admission to higher education.

There were 263 technical and vocational schools and institutes with a total number of 80,830 students and 4,229 teachers in 2015 academic year. The proportion of student's figure to the total population of the country is 0.27%.

Teacher Training Education

Teacher Training Centers in Afghanistan are responsible for training teachers for primary, lower secondary and higher secondary level to improve the learning opportunities across Afghanistan. Teacher training for those at the primary or lower secondary level is overseen by the MoE. Teacher training for those at the higher secondary level or above is overseen by the MoHE. Passing a successful *Kankor* grant applicants admission at Teacher Training Colleges (grade 13-14) under the Teacher Education Department of MoE. Qualified graduates (obtaining >=80% overall score) with the acquisition of vocational training will

be appointed as teachers in primary school and lower secondary school, or they could join the *specialized Kankor* to continue higher education studies. As of 2015 academic year, there are 318 teacher training centers with total number of 3,369 teachers, and 81,212 students. The proportion of this figure to the total country population is 0.27%.

Islamic Education

The Islamic Education centers are educational institutions which offer programs at secondary education level that prepare the students both for tertiary education and work as mullah and preachers in the mosques. Some also work in the jurisprudence area. Islamic education is provided through Madrasas (grades 1 to 12), Dar-ul-Heffaz (schools covering grades 1 to 12 and primarily focusing on Quranic studies, memorization of the Holy Quran, and recitation) and Dar-ul-Ulums (grades 13 to 14 offered in Centers of Excellence where students are provided further Islamic education almost the same format and the same services as in Madrasas; students at the district level attend Madrasas and they can enroll in Centers of Excellence that exist in provincial capitals). As of 2015 academic year there were 865 Islamic education centers, 299,693 students and 9,595 teachers in these centers.

2.4.2. Kankor in Afghanistan

In Afghanistan, high school graduates need to pass the *Kankor* in order to continue higher education in public and private universities.

The *Kankor* is held every year, mainly in the capital and large provinces, usually between December and the end of February. As of 2017, it is scheduled to begin in mid-March, ends in July, and the *Kankor* results are scheduled to be announced in August. Since 2006 the *Kankor* examination process was computerized – the questions bank and paper setting, Optical Mark Recognition (OMR) form correction and assess the answer sheets. Prior to 2006 usually 6 months were required to finish *Kankor* – from registering for the exam until announcing the results.

The *Kankor* exam, a 2-hour exam, comprises 160 Multiple-choice questions about subjects taught at high school mainly in grades 10 to 12 categorized into the following categories: Mathematics, Natural Science, Social Science, and Languages. *Kankor* questions are provided both in Pashtu and Dari, depending on the region. Usually, correct answers are worth one to three points. The maximum number of points is between 320 and 370, but more important is the minimum number of points needed to qualify for a university slot. The Kankor candidates could opt for ten fields of study and for one location for each field of study. They also had three chances to participate in *Kankor* in a lifetime until 2011. Since then, they can opt for only five fields of study and have only two chances in a lifetime.

In general, admission into popular fields of study like medicine, engineering, computer science, economics, and political sciences and into highly selective educational institutions, particularly, in the capital and main cities requires higher scores. However, if a student does not score high enough for any of his/her chosen fields of study, he/she is dropped altogether and is not assigned any field of study, a result called result-less (benatedja).

Statistics and analyses illustrate that a candidate can perform well and get a high score, and still does not get seated as per his/her choice/s. Thus, it is 'first come first served'. It means, if candidates *Kankor score* matches their 2nd, 3rd, 4th, or 5th choice of field of study, they will be admitted only after every other candidate who have marked that field as their first choice, and have a high enough mark for it, have been admitted. If no slots are left in that field after everybody else has been admitted, the candidates will fail in *Kankor*. On the other hand, if there are N seats for Fine Arts, then the first N candidates who's score matches Fine Arts as their first choice will get it, even if their score is high enough to enter a more prestigious field of study like medicine. It can be concluded that the candidates' uninformed choice of field of study might lead to unsuccessful or undesired admission.

The concept of *Kankor* has negative impact on education in general and school instruction in specific. For example, today, many families invest in *Kankor* preparation courses or pay large fees to enroll their children in private high schools, because their main objective is that their children should receive admission. In contrast, the mission of education is not merely preparing students to enter university, but to train students for a successful life. The current structure of *Kankor* essentially prevents high school teachers from offering 'appropriate training' in the classroom, because of the current atmosphere and situations of *Kankor* settings. Thus, the entire training system in high schools is currently affected negatively, because many students do not pay attention to the course content and classroom and only study for admission.

The research conducted among 1,309 public and private high school and university students reveals that more than 86% attended *Kankor pilot tests* which are held by private or public educational institutions. The purpose was to become familiar with *Kankor* test and methods, and to learn through Exam Crams. The Kankor candidates believe a simulator application as an assessment tools (*e-Kankor)* will help the candidates better self-evaluate their competitiveness and self-confidence. A similar research among 632 respondents in public and private universities in Herat province about their preferred field of study also reveals that more than 56% were not admitted into their favorite field of study and more than 50% of students are not satisfied with the field of study in which they have been admitted. Moreover, the author collected data on the number of students admitted into different faculties at Herat University, and the successful graduation rates for same targeted group of students, for 2009, 2010 and 2011 to compare the ratio of enrollment vs. successful graduation. The data shows one-third of the students admitted into Herat public university could not continue Higher education studies.

According to another survey conducted by the author by means of a questionnaire, particularly, among Computer Science students at public and private universities in Herat province, out of 227 respondents (e.g., freshmen, sophomore, and graduates; male and female students) around 90% did not have the basic skills and knowledge of programming, database concept, and operating systems, as echoed in the following figure.

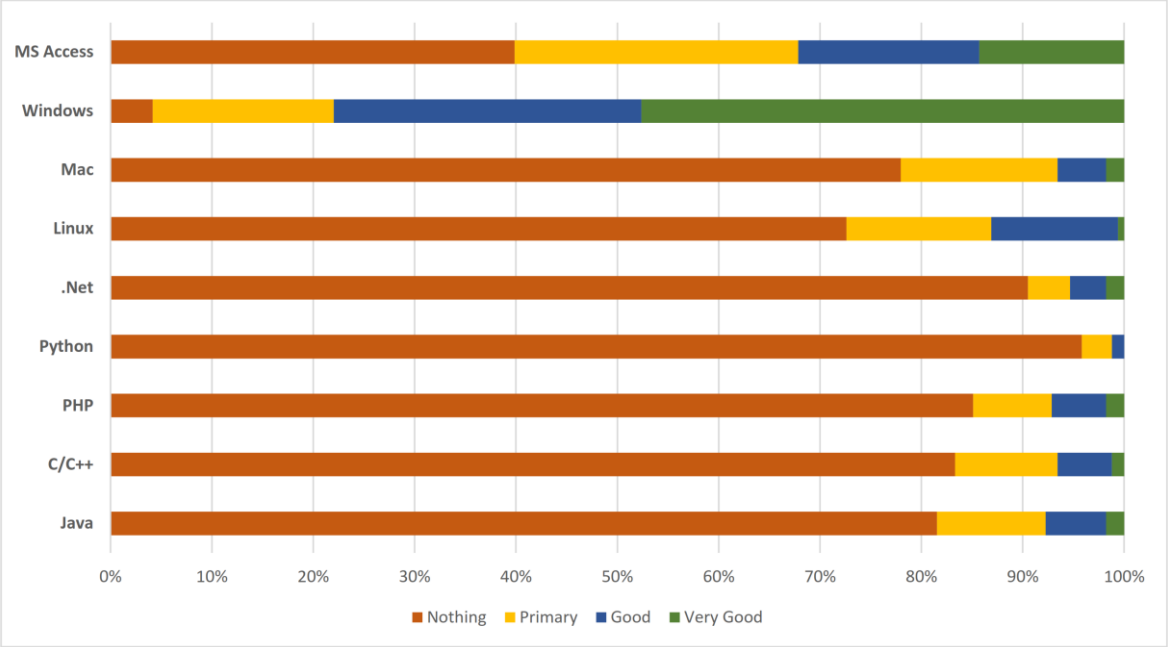

Figure 2-3. Basic ICT Skills of Computer Science students prior of their admission through Kankor.

2.4.3. Higher Education

In addition to the successful completion of higher secondary education, admission to the university also requires a successful *Kankor*, introduced in 1966. The students who fail the examination could take it later in the following year/s. They are permitted to take it twice in their lifetime.

The general structure of Higher education consists of a bachelor's degree that normally takes four years to complete (five years in case of stomatology, and veterinary; seven years in the case of medicine), Master's degrees last two years, and Doctoral degrees usually require three years of study. Presently a limited number of Master and Doctoral degree programs are offered in Afghanistan.

2.4.4. Grading Scale

Both educational and Higher educational institutions in Afghanistan use a 100-point grading scale. In general education, a grade below 40 and in higher education a grade below 55 for a specific subject is deemed unsatisfactory. In higher education, an average overall score of 60 is required to transfer to the next semester.

2.4.5. Afghanistan's Education and Higher Education Systems at a Glance

The following figure gives an overview of the current Afghanistan Education and Higher Education Systems.

1. Officially there is no coordination between primary education and mosque education.
2. Officially there is no pre-primary education under the supervision of Ministry of Education (MoE).
3. Light blue in figure 3 indicates lack of proper monitoring and evaluation by the MoE as well as lack of interest and attention of students to the school subjects.

4. Red in figure 3 depicts the most important focus of both the students and their families. Almost all the students invest in *Kankor preparation courses*. Also, many families pay large fees to enroll their children in private high schools, because their main objective is that their children should receive their admission.
5. Dark blue in figure 3 shows attention and efforts of students only in Higher Education.

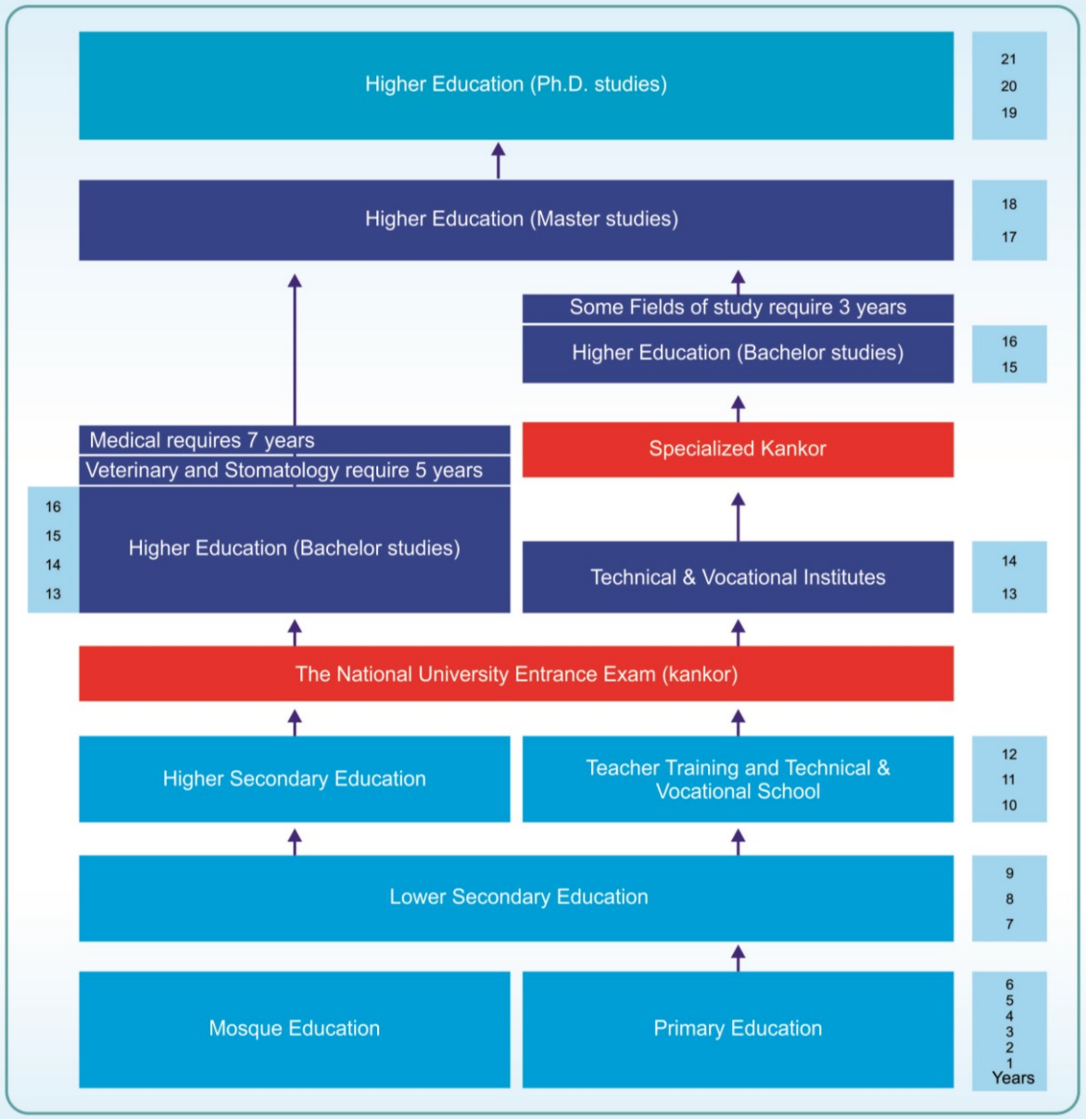

Figure 2-4: Current Educational System of Afghanistan (Visualization by the author).

2.4.6. Summary of Education System in Afghanistan

External aggression and war, political use of educational institutions, radical and rapid change in cultural and traditional issues, intervention of external powers, and low literacy and knowledge rate in the society have negatively impacted the quality and development of education in Afghanistan. The consequences range from weak performance and drop in education both in terms of quantity and quality, to restricted education for females and the fact that most teachers fled to neighboring countries.

Afghanistan's education system (as a base for higher education) is largely uncooperative with the market and Higher education. It does not prepare the students for a career to enter the labor market when they graduate. Neither does it equip students with a strong and quality academic background for entering Higher education. Thus, what is taught at schools and universities in Afghanistan will not be very helpful in meeting the needs of the society, and the students will not be able to easily get into the labor market after graduation.

The following challenges are identified and outlined and should be addressed properly: The severe *Kankor* led to more trust in private high schools because a *Kankor centric curriculum* is taught there. Likewise, because more than 50% of *Kankor* questions focus on mathematics and physics; this leads to less attention by students to non-science subjects. In addition, lack of specialized studies at schools generates more generalists than specialists and leads to weak performance in Higher education and increase in unemployment. And (last not least) lack of qualified teachers at primary, secondary and high schools where most of the teachers do not have a bachelor's degree and are inexperienced.

In Afghanistan, *Kankor* is used to identify the skills and competences of the candidates. The candidates have *two chances* to take *Kankor* in a life-time – the candidates chosen majors are considered their career in higher education. The Higher education is the source of producing qualified and educated manpower. Educated and innovative manpower leads to employment opportunities. Finally, graduates are allocated and are in charge of both the government and non-government sectors. In summary, *Kankor* directly or indirectly influences the skills identification, career development of academic and professional workforce. It is important to note that major problems were only caused by a collection of smaller factors and only countermeasures are required to resolve the given situation. Hence, the author picked *Kankor* as his research case studies using data mining techniques and other disciplines to support policy makers and advisors helping the Kankor candidates concerning their academic careers.

Additionally, the author believes that serious attention at primary and secondary schools could strengthen student skills and would lead toward a far better and qualified Higher education. Therefore, the following main points are to be considered for the improvement of education; and further attention must be paid to them:

- Coordination and building of trust between basic education and mosque education will strengthen education sustainability in the country.
- Development and promotion of public and private kindergartens, particularly in cities can be the founding step for promoting basic education.
- Proper monitoring and evaluation of basic and general education is crucial (since education is the foundation for higher education!)
- Providing high school students with specialized courses with a practice-oriented approach will prepare them for the job market far better. Also, should they choose to continue at the higher education level; they would be more qualified for the courses they choose.
- Advanced teacher training programs to increase teachers' qualifications.

- Quality of higher Education cannot be realized without quality basic and general education. Hence, strong cooperation between the MoE and the MoHE is required.

2.5. SUMMARY

Not only in Afghanistan, but also in other countries including Iran, Turkey, China, and Japan, *Kankor* is very important for high school graduates who are interested in pursuing their dream of higher education. In many countries, there is a *Kankor* exam in some form, even though the terms used to refer to it are different, such as, *college entrance exam, national higher education entrance exam,* and so on.

In most cases, for the top prestigious universities, *Kankor* is required for admission. Competition for admission into desired fields of study and the top ranked higher education institutions is very challenging. Also, in countries where higher education is considered to also be a social status and employment depends on it, *Kankor* exam becomes a method of selecting and filtering the candidates. Therefore, for countries such as Afghanistan, Iran, China and Turkey, where competition for university admission is more intense, the candidates are under enormous psychological pressure. Thus, it can be deduced that the tension associated with *Kankor* is not only a feature of the Afghan testing system but also a common issue for the similar contexts in the aforementioned countries. However, in other countries, researchers from different disciplines have carried out multiple studies to guide the candidates. They have introduced useful methods for *Kankor* candidates to reduce their stress, such as:

- exam crams (test techniques and methods required for exam preparation),
- institutionalization of academic counseling for career decision-making to help the candidates and to make career planning suitable for their interests and skills,
- offering pre-university studies to better prepare the candidates for *Kankor*,
- offering specialized studies either upon entering high school or in the last year of high school.

However, the challenges and problems of *Kankor* candidates in Afghanistan are profoundly deeper in comparison with the problems of candidates in the mentioned countries. A few are as follows:

- lack of specialized studies at secondary and high school levels,
- lack of pre-university courses to prepare the candidates for *Kankor*,
- lack of academic counseling organizations within/outside the education system to support the candidates in this critical career decision-making process,
- lack of proper facilities and systems to train the candidates on the methods and techniques of exam cram,
- lack of qualified teachers at secondary and high school levels,
- lack of proper classification of fields of study into main streams,
- the limitation to attend *Kankor* only twice in a lifetime,
- low literacy or illiteracy of parents,
- and the increasing number of applicants.

These challenges combined together cause the applicants to choose their field of study in complete randomness without systematic preparation and understanding of the consequences. As a result of this wrong choice of field of study, candidates drop out of higher education, or they discover they are not interested in the fields they are studying. This increases attrition rates and decrease students' retention rates and thus negatively impacts the quality of higher education.

Policymakers and top government officials in Afghanistan have inserted reforms at various layers. For instance, a recent change was directed at preventing fraud in the *Kankor* by preventing the leaking of exam questions through printing of separate question forms for each province. Another attempt was to prevent impersonation by implementing biometric system to identify *Kankor* test takers.

The above changes affected the date of commencement for *Kankor*. This rescheduling delayed the start of the university year by one semester. While these changes took place to prevent fraud, which is considered important, scientific research and studies are not carried out to bring inclusive, constructive and systematic reforms in the *Kankor* settings and methods.

The goal of this thesis is to investigate practical solutions and approaches based on data mining, recommender systems and other assessment methods to support policymakers and to facilitate counselling services aimed at guiding of Afghan Kankor candidates so that they can plan their careers in the light of their skills and interests. Reforming Afghanistan's education system, particularly through enhancing *Kankor* methodologies, as outlined in this research, will not only alleviate the associated challenges but will also be a beginning for a multiplier effect in various other dimensions.

Chapter 3

Chapter 3. Elaboration of Data Driven Science and Opportunities in Afghanistan

Initially, the concept and definition of Big Data, and the emergence of big data solutions as powerful tools for managing and analyzing massive volumes of data will be briefly presented. Next, the concept and application of data mining, educational data mining (EDM), and recommender systems will be elaborated. Then the research provides scenarios and discusses challenges on how these data-driven solutions can be used as a resource in the context of education in Afghanistan. Finally, the research focus to equip high school advisors with proper tools, so they can provide suitable guidance to the Kankor candidates concerning their academic career will be briefly explained.

3.1. Big Data

Upon entering the 21st century, there has been a massive increase in data generation, with prevalent mobile technologies, social media, e-commerce, and sensor devices providing real-time information about such vast sources of data as rivers, oceans and other environments, to name just a few, quickly outpacing the capacity of traditional computing techniques. These vast amounts of generated data are too big and complex to be stored, analyzed and processed effectively by traditional data management tools.

Big Data basically refers to the fact that organizations can now collect and analyze data in ways that were simply impossible even a few years ago (Marr, 2016). Another definition from (Prasad 2016) states, "Big Data is any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information where the individual records stop mattering and only aggregates matter." Nevertheless, the term Big Data is quite vague and too generic, it does not just mean very large amounts of data, though that could be one of the factors and characteristics. In another detailed and comprehensive definition from (Buyya, Calheiros, and Dastjerdi 2016), Big Data is identified by the following three main cornerstones known as 3Vs: Volume, Variety and Velocity. Yet, it does not capture all the aspects of Big Data accurately. Later other V's such as Value/Veracity, Variability, and Visibility were added by IBM, Microsoft and others to define the term big data very precisely.

The Characteristics/6Vs of Big Data:

1. *Volume* indicates the scale of data, which means enormous volumes of data are generated by social media, e-commerce, smart phones, GPS devices, sensors, and the

others. IBM indicates that every day 2.5 exabytes[2] of data are created (Sagiroglu and Sinanc 2013).

2. *Velocity* means how rapidly data is generated. According to the InternetLiveStats[3] every minute there are more than 450,000 tweets sent from 308 million active users on Twitter, around 47,000 photos uploaded on Instagram, more than 155,000 calls on Skype, more than 3 million searches on Google, more than 4 million video views performed on YouTube, more than 155 million emails sent, and more than 2 million Gigabytes of Internet traffic generated (InternetLiveStats 2017).
3. *Variety* stands for different forms of data, which means the variety of incompatible and inconsistent data formats and data structures such as images, videos, audios, documents, emails, click streams, search queries, and others. According to (Das and Kumar 2013), "More than 80% of all potentially useful business information is unstructured data, in kinds of sensor readings, console logs and so on."
4. *Veracity* implies the uncertainty of data i.e. biases, noise, and abnormality in data; and focuses on trustworthiness of data sources.
5. *Variability* refers to the complexity of a dataset. In comparison with Variety, it means the number of variables in datasets.
6. *Visibility* emphasizes that you need to have a full picture of data to make informative decisions.

The diagram below (see **Figure 3-1**) summarizes the Douglas Laney's 3Vs, IBM's 4Vs, Yuri Demchenko's 5Vs, and Microsoft's 6Vs of Big Data attributes.

Big data solution technologies enable businesses to store and analyze massive amounts of data effectively and efficiently in real time to gain deeper market insights, and create new value useful for the development of organizations. Data mining is a characteristic of the Big Data solution technologies to help organizations concentrate on the most relevant information in the data. Through data mining, they can discover information within the data, which traditional computation queries and reports may not easily reveal (Reynolds 2016). (Hurwitz et al. 2013) outlines some of the emerging big data applications in areas such as healthcare, manufacturing management, traffic management, etc. which rely on huge volumes, velocities, and varieties of data to transform the behavior of a market. In healthcare, a big data application might be to monitor premature infants to determine when data indicates intervention is needed. In manufacturing, a big data application can be to prevent a machine from shutting down during a production run. A big data traffic management application can reduce the number of traffic jams on busy city highways to decrease accidents, save fuel, and reduce pollution.

---

[2] Data is measured in bytes. An exabyte equals 1,024 petabytes. A petabyte equals 1,024 terabytes, and a tera-bytes equals 1,024 gigabytes.

[3] Internet Live Stats is part of the Real-Time Statistics Project (Worldometers: www.worldometers.info and 7 Billion World: www.7billionworld.com), which collects, visualizes and analyzes large amounts of data from different sources in real time.

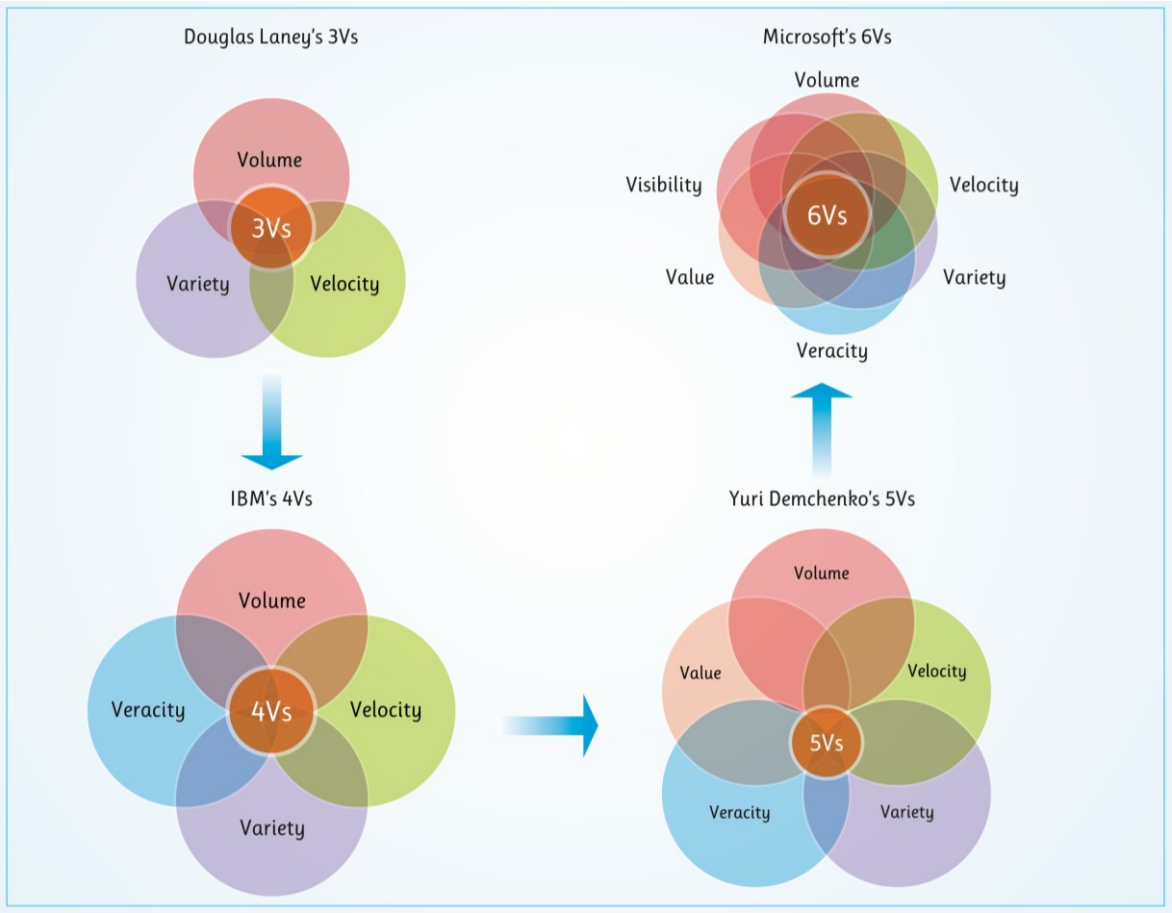

Figure 3-1. The 3Vs, 4Vs, 5Vs, and 6Vs of Big Data, visualization by the author following (Buyya, Calheiros, and Dastjerdi 2016)

Big Data leads to new big data solution technologies and distributed architectures (Hussain and Farah 2015) such as: MapReduce (Dean and Ghemawat 2008), Apache Hadoop (White 2015), Apache Spark (Karau et al. 2015) and Apache Flink, to name a few. These big data solutions can effectively handle and process big data in stream and batch processing in a reliable and scalable fashion. It should be considered that big data, and big data solutions, certainly do not replace the traditional data management tools and computing techniques. Likewise, data mining is not only for big data. Hence, companies and organizations with small and medium sizes of data utilize data mining techniques to produce interesting patterns and insights useful for decision making. On the other hand, Microsoft SQL Server (MacLennan, Tang, and Crivat 2008), Oracle (Tierney 2014), and the powerful Relational Database Management Systems (RDBMSs) are now capable of supporting data mining applications. Similarly, there are plenty of general data mining tools that provide data preparation capabilities, exploration and visualization techniques, and data mining algorithms. Some examples of commercial and open source tools are RapidMiner (Kotu and Deshpande 2014b), Weka (Witten, Frank, and Hall 2011), Tableau, KNIME, IBM SPSS Statistics and Modeler, Microsoft Power BI, powerful packages for R (Williams 2011a; Zhao 2012) and Python (Richert and Coelho 2013) languages, to name a few (Goebel and Gruenwald 1999; Mikut and Reischl 2011).

3.2. DATA MINING

Data Mining, also known as Knowledge Discovery from Data or Knowledge Discovery in Databases, KDD for short (Han, Kamber, and Pei 2011), is the process of mining through large amounts of data for hidden patterns and gaining useful insights, which can project future trends and behaviors. Data mining turns a massive volume of data into useful information and knowledge enabling businesses, decision-makers, and policy-makers to make better decisions.

Data mining is the study of collecting, scrubbing, cleaning, processing, analyzing, and gaining useful insights and making sense from data (Aggarwal 2015). Thus, Data mining is an interdisciplinary field and the cross-road of statistics, database systems, data warehouses, machine learning, artificial intelligence algorithms, and others (Sayad 2011).

Exploration and prediction are the aspects of data mining with numerous of potential to help organizations explain the past by means of univariate and bivariate exploration analysis, and to predict the future by means of classification and clustering techniques (Sayad 2010), as illustrated in (see **Figure 3-2**).

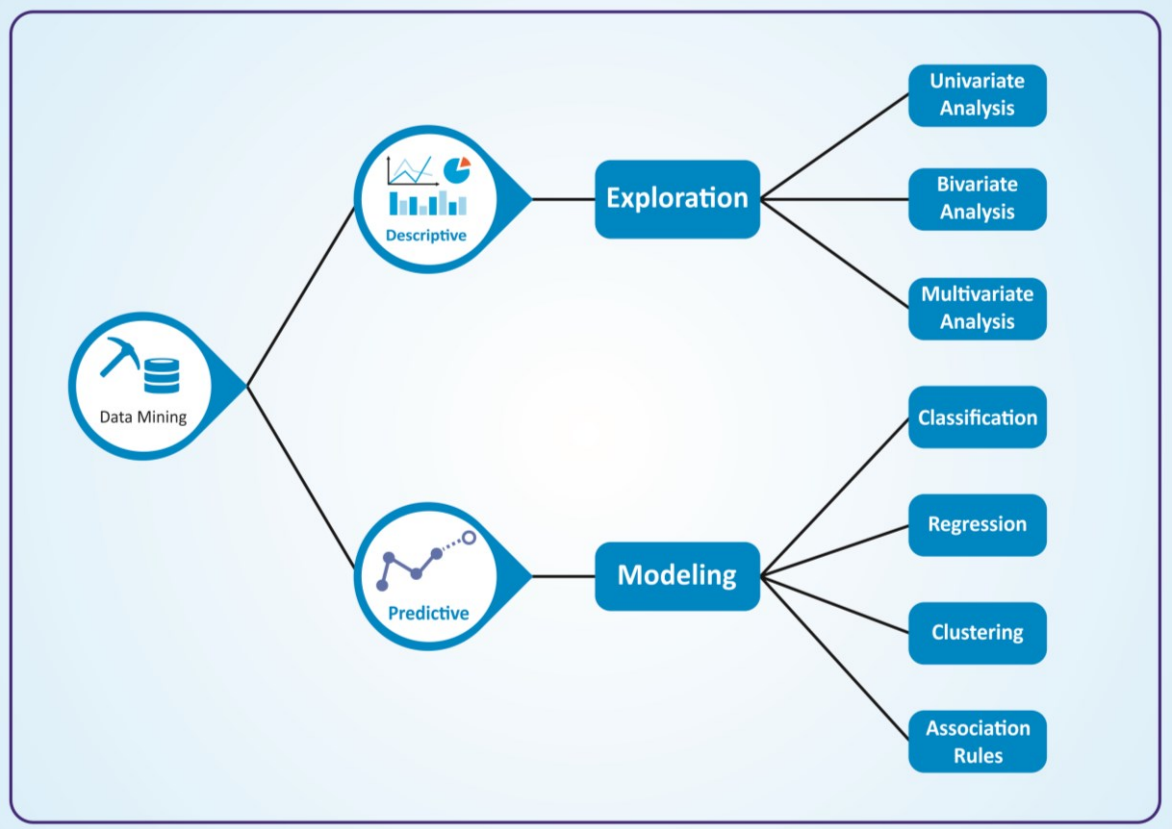

Figure 3-2. The Data Mining Map, visualization by the author following (Sayad 2010)

Prior to building and training a model, it is essential to statistically investigate and visualize the data to have a better picture and understanding of the data. For example, calculating the frequency of the data, and graphing them, to calculate the measures of central tendencies (means, medians, modes), calculating the dispersion of scores (variances, standard deviations), to examine the relationships between variables, and identifying outliers in the distribution of the scores (Ho 2013). Data exploration is to bring important characteristics of

the data into focus for further investigation. There are many methods to describe and explain organizations statistically such as: univariate analysis, bivariate analysis, and multivariate analysis. The univariate analysis explores only one variable/attribute (categorical or numerical) at a time. While bivariate analysis is a special case of multivariate analysis, the simultaneous analysis of two variables, and focus on the concept of relationship between them. The two variables in bivariate analysis can be both numerical, categorical, or numerical and categorical. Finally, the multivariate analysis is like bivariate analysis, where multiple relations between multiple variables are explored simultaneously.

On the other hand, predictive modeling is the process of training and building a model to predict an outcome for new and unseen data. If the outcome of the model is categorical, it is called classification, otherwise, it is called regression. Clustering, also known as unsupervised learning, is the task of organizing observations into clusters/groups so that observations in the same cluster are similar. Finally, association rules are another kind of data mining techniques for discovering interesting relations and associations between variables amongst observations.

Data mining and analytics have successfully been used in industries, and it is still evolving in other disciplines and domains. According to the poll by (KDnuggets 2016), data mining and analytics are applied in a wide range of industries such as: customer relationship management (CRM) or consumer analytics, finance, banking, healthcare, e-commerce, social media, fraud detection, science, advertising, games, education, and many other sectors.

In the previous years, particularly 2008, researchers from a variety of disciplines such as: computer science, statistics, data mining, pedagogy, and education have begun to investigate data uniquely generated from education to improve the education settings and facilitate education research (Cristobal Romero, Ventura, Pechenizkiy, & Baker, 2010).

3.3. Educational Data Mining

A decade ago, researchers from a variety of disciplines including computer science, education, psychology, and statistics together founded the Educational Data Mining (EDM) community. The main purpose of this community is to apply data mining methods in the education contexts and to investigate how data mining methods can improve education and facilitate education research. In the last years, the EDM research community has undergone remarkable growth. Since 2008, an international conference on EDM is held every year to bring together researchers from diverse disciplines to analyze large datasets to answer educational research questions. In 2009, the Journal of Educational Data Mining (JEDM) was established for sharing and broadcasting the research results. Per (C. Romero and Ventura 2007) EDM is increasingly recognized as an emerging discipline.

The EDM community strives to investigate and improve the existing methods of data mining as well as to develop new methods to explore the unique types of data which are generated from educational domains with the aim of improving education. These data are generated from numerous sources, including data from traditional face-to-face classrooms at schools, colleges, and universities; and traces that students leave when they interact with educational

software and online courseware. These sources increasingly provide massive volumes of data, which can be analyzed to easily address questions that were not previously achievable, including differences between student populations, including uncommon student behaviors, providing feedback for supporting instructors, recommendations for students (Cristobal Romero et al. 2010).

The author's investigation and studies (Sherzad 2016) – found that there are vast amounts of data available for mining purposes in Afghanistan. The use of Internet in education, implementation of educational software and online courseware at schools and universities, application of EMIS (Education Management Information System) at MoE, and HEMIS (Higher Education Management Information System) at MoHE means that the amount of data will increase significantly. But the methods that the MoE and MoHE use to store and produce their data only enable them to achieve basic facts and figures (total number of students and teachers based on gender, geographic location, schools and universities) which are not very helpful in decision making to improve the education effectively. For example, one cannot predict proper fields of study (Major) for high school graduates, detect undesirable student behaviors, or identify first year university students who are at high risk of attrition. EDM seeks to use these data to better understand students and the settings in which they learn. The EMD website, defines EDM as follows "Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings, and using those methods to better understand students, and the settings which they learn in" (EDM community website 2017).

The applications of EDM are very broad, ranging from analysis and visualization of data, providing feedback for supporting instructors, recommendations for students, predicting student performance, detecting undesirable student behaviors, grouping students, constructing courseware, to planning and scheduling, and others (C. Romero and Ventura 2010). In the last years according to (Baker and Yacef 2009; Cristobal Romero et al. 2010), EDM has been applied to address a wide number of goals. For example, supporting course administrators and educators in analyzing students' activities in courses, determining how to improve courses (contents, activities, links, etc.), recommending most appropriate courses to students, predicting a student's final grades or other types of learning outcomes (such as retention in a degree program or future ability to learn) based on data from course activities, and the others.

Classification

The idea of classification in data mining is to assign an instance/observation/case to the proper target class or category based on the instance predictors. The aim of classification is to accurately predict the target class for each instance in the data. For example, a classification model can be used to classify students into sciences or social sciences Majors. In real life education, according to (Cristobal Romero et al. 2010), teachers and instructors are classifying their students based on their knowledge, performance, behavior, motivation, and other factors.

Classification is a type of supervised learning algorithm. The simplest type of classification problem is binary class classification. In binary class classification, the target/class attribute has only two possible values such as: sciences or social sciences, pass or fail. While in multiclass classification, the class attribute has more than two values such as: sciences, social sciences, medical sciences, or technical and vocational.

The classification task begins with a dataset in which the class assignments are known. For example, a classification model that predicts Majors can be developed based on observed data for many students over a period. In addition to the historical performance and marks, the data might track students' histories, participation in supportive courses, family literacy and economic status, marital status, and so on. Identifying a Major would be the target, the other attributes would be the predictors, and the data for each student would constitute a case.

Classification has many applications: In banking to analyze the clients' data to know which ones are risky and which are safe; in marketing to analyze the customers' profiles to identify which customers are more likely to buy a new product; in education to analyze students' data to classify them properly into sciences or social sciences Majors. While the classification processes can be manually constructed based on experts' knowledge, it nowadays is a very time-consuming process. Because of that it is more efficient to design classifiers from the existing data to classify new observations in an autonomous fashion.

There are more classification methods, like ZeroR, OneR, Naive Bayesian, Decision Tree, K Nearest Neighbors, Neural Network and others. To build a classification model, initially one should choose one of the classification methods and then a dataset, where all class values are known. The data set is divided into two parts, a training dataset to design and train the classifier model, and a test dataset where all class values are hidden to test the classifier model. If the classifier classifies most cases in the test dataset correctly, it is assumed that it will also work accurately on future data. Otherwise, it is assumed as a wrong model. Typically, designing a classification model is an iterative process. To come across a better classification model, it is common to try different data manipulation, several classification methods with appropriate parameter settings of the classification algorithm.

Clustering

Clustering is a type of unsupervised learning algorithm. It is a technique for extracting information from unlabeled data. Clustering analysis is the process of dividing unlabeled data into groups/clusters (the term 'clusters' and 'groups' are synonymous in the world of cluster analysis) such that similar objects are grouped in one cluster and dissimilar objects are grouped in a different cluster. Typically, measures such as Euclidean Distance, Manhattan Distance, Cosine Similarity, and Correlation Similarity, Jaccard Similarity and the others are used to determine the similarity or dissimilarity of the objects. The definition of similarity and the method in which the objects are grouped vary based on the clustering algorithm being applied. Hence, different clustering algorithms are suited to different types of datasets and different purposes. Typically, designing and training the model – like any other data mining tasks – is an iterative process, where one must try different data manipulation, similarity approaches, and clustering algorithm and settings before a good

clustering model is found. There are many clustering algorithms, and mainly they are classified into the following categories: Partitioning, Hierarchical, Density-based, Grid-Based, Model-Based, and Constraint-based.

Clustering analyses are suitable for many applications where labeled data is difficult to obtain and clustering algorithms can be used to find natural groupings. Clustering is also useful for anomaly detection such that objects which do not fit well into any clusters are outliers. Additionally according to (Wu et al. 2008; Wu and Kumar 2009), clustering can also serve as a useful data-preprocessing step to identify similar clusters on which to build supervised classification models. In case of Afghanistan, there are more than a hundred disciplines (fields of study), and these fields are not classified into main streams i.e. sciences, social sciences, and others. Even sometimes their names vary from province to province and from university to university. On the one hand, this is very challenging for Kankor candidates to know enough about those fields of study to make informed decisions in choosing appropriate fields of study. On the other hand, it is very complex to design a classification model with more than a few hundred values/fields of study in the target attribute. However, clustering techniques can be used to classify the fields of study to reduce the number of possible values for the target attribute prior to designing the classification learning model.

Association Rules

Association rules analyses are data mining techniques for uncovering relations between variables in large transactional databases with the aim of discovering how items are associated with each other. According to (Merceron and Yacef 2008), association rules are widely used to analyze market basket analysis such as if customers buy Item X, they also buy Item Y at the same time – this can be written as X→Y. Typically (Ng 2016; Sayad 2017), there are three common ways to measure association: support, confidence, and lift:

1. The support of a rule shows how frequently the items in the rule occur together. Suppose there is a rule according to which X implies Y such that if customer buys item X he also buys item Y. This can be written as X→Y. Then, support is the ratio of transactions that include the items X and Y to the number of total transactions. Support is defined as follows:

$$Support\ (X \rightarrow Y) = \frac{frequency\ (X, Y)}{N}$$

2. The confidence of a rule denotes the probability of both the antecedent and the consequent appearing in the same transaction, expressed as {X → Y} divided by the number of transactions that include the antecedent. The Item X is known as the antecedent, and Item Y is known as the consequent. Confidence is defined as follows:

$$Confidence\ (X \rightarrow Y) = \frac{frequency\ (X, Y)}{frequency\ (X)}$$

3. One drawback of the confidence measure is that it might misrepresent the importance of an association. This is because it only accounts for how popular Item X is, but not Item Y. If Items Y are also very popular, in general there will be a higher chance that a transaction containing Item X will also contain Item Y, thus inflating the confidence

measure. To account for the base popularity of both constituent items, lift measure is used. The lift is support of a rule over the support of the antecedent times the support of the consequent. It provides information about the improvement, the increase in probability of the consequent given the antecedent. The lift of a rule is defined as:

$$Lift\ (X \rightarrow Y) = \frac{Support\ (X, Y)}{Support\ (X) * Support\ (Y)}$$

For clarification, the author elaborates the concept of support, confidence, and lift with an example. Let I = $\{i_1, i_2, i_3, \ldots, i_p\}$ be a set of p items, and T = $\{t_1, t_2, t_3, \ldots, t_n\}$ be a set of n transactions in the dataset, with each t_i a subset of items I. Table x shows an example of transaction dataset, and table y shows the rules and results of the association measures.

There are various association rules algorithms such as AIS, SETM, Apriori, AprioriTid, AprioriHybrid, FP-growth – each has advantages and disadvantages (Kumbhare and Chobe 2014; Sayad 2017). Like any other data mining learning task, different data manipulation, association algorithms with appropriate parameter settings must be tried before a good association model is found.

In addition to the market basket analysis and e-commerce, association rules have many applications including detecting fraud, intrusion detection, and other domains. Association rules are (Chair et al. 2004; Merceron and Yacef 2008; Zaíane 2002) also useful in EDM and can be extended to Learning Management Systems (LMSs) for analyzing learning data. Most of today's LMSs are very popular and store vast logs of students' activities and behavior implicitly or explicitly while they are browsing courses, taking tests and quizzes, reading the contents, watching videos, following forum threads and topics, performing various tasks, and even communicating with peers. The log data even includes the session duration of entry and exit time, mouse movements, delay taking the tests and quizzes, visiting pages and modules, etc. These data (Mostow and Beck 2006) are very useful for analyzing students' behavior and could create a gold mine of educational data. LMSs such as: Moodle (https://www.moodle.org), Ilias (http://www.ilias.de), Claroline (http://www.claroline.net), Blackboard (http://www.blackboard.com) and the others can offer a great variety of channels and workspaces to facilitate information distribution and communication among participants in online courses. LMSs enable school teachers and university lecturers and administrations to share information with students, produce content materials, prepare assignments and tests, engage in discussions, manage distance classes, and enable collaborative learning with forums, chats, file storage areas, and other services.

MoHE decided to introduce e-learning in Afghanistan higher education. Prior to offering fully online courses; the milestone is to make the environment ready by introducing educational software and courseware to enrich the current education system. This approach enables the educational institutions to collect vast amounts of explicit and implicit data which are very valuable for mining and analyzing students' and other users' behavior (Mostow and Beck 2006). Keeping the application of association mining in mind will be a plus point to improve online educational settings far better. Also other data mining techniques can be used on these collected data to recommend to students the right courses (Zaíane 2002). It also enables the system to predict the students' scores and results while

choosing courses even before any examinations (Maghsoudi et al. 2012). The former factor raises the awareness of the society and the latter one enlightens and strengthens academicians mining the data and the importance of data availability and accessibility.

In addition, in Afghanistan, Kankor candidates can choose 5 different fields of study in order of decreasing priority from 1 to 5. If a candidate's *Kankor score* is not high enough to get them into their 1st choice, then it will be evaluated to see whether they qualify to be admitted into their 2nd, 3rd, 4th, or 5th choices. However, for these 2nd through 5th choices, priority will be given to other candidates who have marked that field as their first choice and have a high enough mark for it.

In most cases there is no clear relationship between the 5 fields of study a candidate chooses in *Kankor*. For example, computer science, engineering, agriculture, law and politics, and literature can be the choices of a candidate. These choices are not relevant academically as some of them belong to the science stream, while others to the social science stream and so on. Currently, there are more than a hundred fields of study that are not organized/grouped into main streams, and candidates must take the *Kankor* to get admitted into any one of these fields. There is even no consensus on the names of the fields of study across the country. Sometimes the name for same field of study varies from university to university. Lack of classification of fields of study can be one of the reasons the candidates' 5 choices are not closely relevant.

No one in Afghanistan has carried out scientific research to evaluate how related the candidates' choices are. Also, no research has been carried out to organize the existing fields of study into main streams. In the former case association techniques can be used to uncover the relationships among the candidates' choices and if their choices are closely relevant. The uncovered relationships might support policymakers to gain insight into how currently the candidates choose their 5 different fields of study. In the latter case clustering approaches combined with other techniques can be applied to group similar fields of study into streams. Such research and studies will support policymakers to systematically improve the *Kankor* settings and methodology and in the long-run to offer *specialized Kankor* for the identified streams. The outcome of the suggested research will not only resolve associated challenges but will be a beginning for a multiplier effect in various other dimensions.

3.4. Recommender Systems

Like data mining and EDM, Recommender Systems (henceforth referred as RSs) are an interdisciplinary field including various scientific disciplines and combining ideas and techniques from areas such as information filtering, user modeling, machine learning, data mining, and human-computer interaction. RSs recommend items to users, such as, what movie to watch, what music to listen to, what book and news to read, etc. Using recommendations in the decision-making process is one of the fundamental elements that people apply when making decisions (Đurović, Dlab, and Hoić-Božić 2016). In real life people use different approaches to make choices and usually rely on friends' and peers' recommendations; even companies rely on recommendation letters in their recruiting process. The RSs are powerful software and techniques which provide proactive and useful

recommendations that are tailored to meet the users' tastes and preferences (Jannach et al. 2010; Ricci et al. 2010). Of course, users' tastes are different and thus different users receive different recommendations matching their preferences. These kinds of recommendations are tailored to individual users, known as personalized recommendations. The users' preferences can be found out explicitly (ask the users to rate items, such as, ratings, likes or dislikes, reviews) or implicitly (make inferences from users' behavior, such as the users' purchasing a product, reading news and articles, watching a movie, marking an item as favorite/wishlist, viewing information about an item, and other users' actions which can be interpreted as the users being interested in the item). Additionally, there are non-personalized recommendations which are simple aggregates, not very complicated to design e.g. popular movies and news, top ten books, etc.

The study of RSs is new in comparison to the traditional information retrieval methods – which are very substantial to the highly rated websites including Netflix (https://www.netflix.com), YouTube (https://www.youtube.com), Spotify (https://www.spotify.com), Last.fm (https://www.last.fm), Amazon (https://www.amazon.com) and others (Ricci et al. 2010). Nowadays many other organizations are designing and deploying RSs as part of the services they provide to their audiences/clients. RSs emerged as an independent research area in the mid-1990s (Adomavicius and Tuzhilin 2005; Ricci et al. 2010). The interest in RSs has dramatically increased over the last decade, and still remains high. To give an example, ACM Recommender Systems (RecSys), established in 2007, is one of the conferences and workshops dedicated to the field. Since its establishment RecSys has received a total of 1279 (only) long paper submissions, and after peer review 329 of them were accepted (a 26% acceptance rate) for publication (ACM RecSys 2017).

Although RSs have shown their effectiveness in e-commerce (Linden, Smith, and York 2003) and other domains, they have recently gained attention both in education and higher education since the educational environment is no longer limited to face-to-face classes (Santos and Boticario 2001). The field of education is an emerging and very promising application area of RSs. Several initiatives such as special issues in journals (Recommender Systems to support the dynamics of virtual learning communities aka Rs4vlc, Recommender Systems in E-Learning Settings aka Rsels), books (Cristobal Romero et al. 2010; Santos and Boticario 2001) and workshops (Educational Recommender Systems aka EdRecSys, Recommender Systems for Technology Enhanced Learning aka RecSysTEL) intend to set up the basis for this specialization in the field. The applications of RSs in education are commonly to support students in their learning process. Therefore, nowadays RSs are immensely integrated in LMSs, Intelligent Tutoring Systems (ITSs) and other educational courseware that are commonly used at all levels of education (Đurović 2016; Đurović, Dlab, and Hoić-Božić 2016; Thai-Nghe, Horvath, and Schmidt-Thieme 2011).

The RSs are typically classified into the following categories, based on how recommendations are made, and each category has advantages and drawbacks, briefly explained as follows:

- **Content-based recommendations:** Recommend items to user x similar to items previously rated highly by the user x. For instance, in a movie recommendation application, if a user has rated a movie highly which belongs to the animation or comedy genre, then the system learns to recommend other movies from these genres in the future. The system does not recommend movies outside the user's content profile, such as, movies from other genres which have not been rated by the user. This is one of the drawbacks of the content-based recommendation, known as, overspecialization. For content-based recommendation, the goal is to create both an item profile consisting of feature-value pairs and a user profile summarizing the preferences of the user, based on their tastes and preferences. Creating the item profile is another limitation of content-based recommendation. Whether extracting the features in an autonomous fashion or manually, sometimes finding the appropriate features, such as image feature extraction, is hard. Furthermore, if two different items are represented by the same set of features, they are indistinguishable e.g. a well-written article and a badly written one, if they happen to use the same terms (Adomavicius and Tuzhilin 2005). Finally, the cold-start problem for new users is another issue in content-based recommendation. The following diagram (see **Figure 3-3**) shows the overall principle behind content-based recommendation:

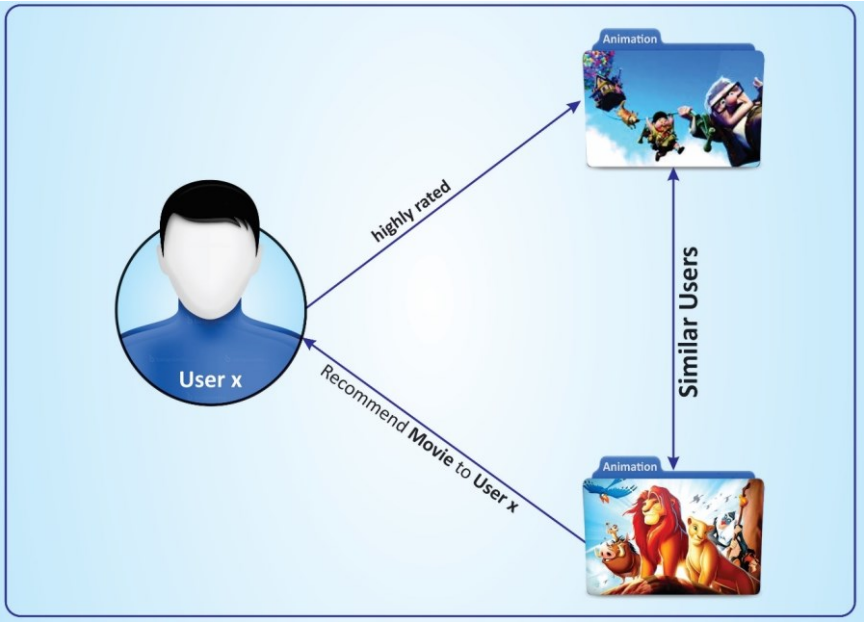

Figure 3-3. The core idea behind the content-based recommendation (visualization by the author[4]).

- **Collaborative recommendations:** The user will be recommended items that other users with similar tastes and preferences liked (rated the same items similarly) in the past. The process of identifying similar users and recommending what similar users like is called collaborative filtering (Leskovec, Rajaraman, and Ullman 2014). There

[4] The movie folder icons are downloaded from the following URLs:

- http://www.iconarchive.com/show/movie-folder-icons-by-lajonard/Animation-icon.html
- http://www.iconarchive.com/show/movie-genres-folder-icons-by-limav/Animation-icon.html

are various approaches to compute the user similarity or the item similarity, such as, K-Nearest Neighbor, Pearson Correlation, Cosine Similarity and others. The collaborative techniques are the most popular and widely used in recommender systems, but suffer from the cold-start (new users have no history and new items have no ratings) and sparsity (most users did not rate most items) problems due to insufficient data. The following diagram (see **Figure 3-4**) shows the overall principle behind user-user collaborative recommendation:

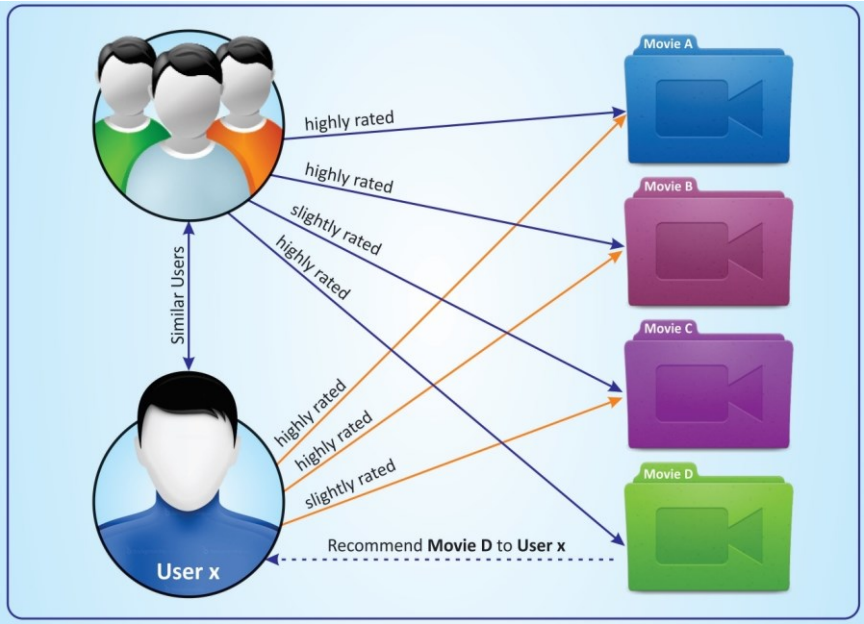

Figure 3-4. The core idea behind the collaborative filtering recommendation (visualization by the author[5]).

- **Demographic:** Demographic methods use descriptions of the user to learn the relationship between a specific item and the types of users who like it (Montaner, López, and Rosa 2003). In this approach, recommendations are based on the demographic profile of the user. Typically, items can be recommended for different demographic niches, by combining the ratings of users in those niches. For instance, recommendations may be customized according to the country, language, or age of the user.
- **Hybrid approaches:** These methods combine the above-mentioned techniques i.e. collaborative and content-based methods. In some cases, hybrid approaches can be more effective as advantages of one technique can overcome some of the disadvantages of the other technique. For instance, collaborative techniques suffer from new item problem, but in the content-based method the new item is not a problem since recommendation is calculated based on the item features. Hybrid methods can be implemented in numerous ways e.g. making content-based and collaborative recommendations separately and then combining them, adding content-

[5] The movie folder icon is downloaded from the following URL:

- http://www.softicons.com/folder-icons/aqua-lion-icons-by-lukeedee/movies-folder-icon

based capabilities to a collaborative approach (and vice versa), or unifying the approaches into one model (Adomavicius and Tuzhilin 2005).

In RSs data is mainly about items to recommend, users who receive the recommendations, and users' transactions (relations between users and items). It is important to note that the data and knowledge sources available for recommender systems can be very diverse, and ultimately whether they can be exploited or not depends on the recommendation technique (Ricci et al. 2010). For example, content-based methods try to recommend items similar to those that a user liked in the past – keywords/features are used to describe the items, and the users' profiles are built to show the type of items the users like. But, collaborative filtering methods do not need the item's profile or the user's profile. They are based on users' activities (ratings, purchases, and other preferences) and predict what users will like based on their similarity to other users. Typically, in a recommendation system application there are two main entities, referred to as users and items. Another entity comprises of users' ratings for items, referred to as transactions/utility matrix.

- **Items:** Items are the objects that are recommended to the user, such as movies to watch, jobs to apply for, or even courses to enroll. Keywords/features are used to describe the items e.g. genre, director, actors, year of release, etc. can be used to describe a movie.
- **Users:** Users are the objects that receive the recommendations. To personalize the recommendations, RSs exploit a range of information about the users depending on the recommendation technique e.g. in collaborative approach, users are modeled as a simple list containing the ratings provided by the user for some items.
- **Transactions:** Transactions represent the relations between users and items i.e. any one user may rate/purchase many items, and on the other hand, any one item may be rated/purchased by many users. Ratings are the most popular form of transaction data – they can be in various forms e.g. numerical ratings such as ratings of movies on a 1-5 stars scale; ordinal ratings such as strongly agree, agree, neutral, disagree, strongly disagree; binary ratings such as like or dislike, and other forms.

3.4.1. Measuring Similarity

In this section, it is briefly explained how to measure similarity of users or similarity of items in the utility matrix using Jaccard similarity, Cosine similarity, and Pearson correlation. The following data and examples are from the Mining Massive Data course book (Leskovec, Rajaraman, and Ullman 2014).

The following users/movies utility matrix (see Table 3-1), represents the users' ratings of movies on a 1–5 scale, with 5 the highest rating. Blanks represent the situation where the user has not rated the movie. The movie names are HP1, HP2, and HP3 which stand for Harry Potter I, II, and III; TW for Twilight; and SW1, SW2, and SW3 for Star Wars I, II, and III. The users are represented by capital letters A through D. For example, User A in Table 3-1, has rated the three movies HP1, TW, and SW1 with rating values of 4, 5, and 1 respectively.

Users/Movies	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

Table 3-1: A utility matrix represents ratings of movies on a 1–5 scale.

Suppose r_x and r_y the vector of user x's and user y's ratings. The need to compute the similarity metric $sim(x, y)$ to capture the intuition that $sim(A, B) > sim(A, C)$. In **Table** 3-1, both users A and B highly rated HP1 in common, while users A and C rated TW and SW in common with completely dissimilar ratings, such as, user A seems to like the TW movie while user C does not.

Jaccard similarity: The Jaccard similarity, also known as Intersection over Union, often called Jaccard index, is used for comparing the similarity and diversity of sample sets. It measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets, described as follows:

$$sim(x, y) = \frac{|r_x \cap r_y|}{|r_x \cup r_y|}$$

Per the utility matrix (see **Table** 3-1) user A and user B rated the movie HP1 in common, and they rated a total of 5 movies, therefore, $sim(A, B) = \frac{1}{5}$. On the other hand, users A and C rated the TW and SW1 in common, and overall, they rated 4 movies, so, $sim(A, C) = \frac{2}{4}$.

The intuition was to capture that $sim(A, B) > sim(A, C)$, but the result is $sim(A, B) < sim(A, C)$, which is in contrast with the intended intuition. Thus, it can be said that Jaccard is not appropriate to consider weights as it ignores the value of the rating.

Cosine similarity: The Cosine similarity is a measure of similarity between two non-zero vectors, and calculates the cosine of the angle between them. The cosine of 0° is 1, and the cosine of any other angle is less than 1. A larger (positive) cosine implies a smaller angle and therefore a smaller distance, and thus represents close similarity. The formula for Cosine similarity is described as follows:

$$sim(\boldsymbol{x}, \boldsymbol{y}) = \cos(\boldsymbol{r_x}, \boldsymbol{r_y}) = \frac{r_x . r_y}{||r_x||.||r_y||} = \frac{\sum_{i=1}^{n} r_{xi} . r_{yi}}{\sqrt{\sum_{i=1}^{n} {r_{xi}}^2} \sqrt{\sum_{i=1}^{n} {r_{yi}}^2}}$$

To compute the Cosine similarity, some value should be inserted for the blank/unknown ratings. The simplest method is to treat them as zero, as illustrated in **Table** 3-2.

Users/Movies	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4	0	0	5	1	0	0
B	5	5	4	0	0	0	0
C	0	0	0	2	4	5	0
D	0	3	0	0	0	0	3

Table 3-2. The modified utility matrix, zeros inserted for the blank/unknown ratings.

The result is $sim(A,B) = 0.38 > sim(A,C) = 0.32$, which slightly meets the intuition that $sim(A,B) > sim(A,C)$. However, $sim(A,B)$ is slightly greater than $sim(A,C)$. It is because of treating the blank ratings as negative ratings by inserting zeros for them. One method to fix this problem is to use the centered Cosine, also known as Pearson correlation.

Pearson correlation: The Pearson correlation, or centered cosine similarity, is a measure of the linear correlation between two variables X and Y. It has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

In this approach, the ratings are normalized for a given user, by subtracting from each rating the average rating of that user. It makes zero as the average rating for every user, turns low ratings into negative numbers and high ratings into positive numbers. The following **Table** 3-3 illustrates the average ratings for each user, where **Table** 3-4 shows the normalized ratings by subtracting from each rating the average rating of that user.

Users/Movies	HP1	HP2	HP3	TW	SW1	SW2	SW3	Average Ratings
A	4			5	1			3.33
B	5	5	4					4.67
C				2	4	5		3.67
D		3					3	3.00

Table 3-3. The average ratings for each user in the utility matrix has been calculated.

Users/movies	HP1	HP2	HP3	TW	SW1	SW2	SW3	Average Ratings
A	2/3	0	0	5/3	-7/3	0	0	0.00
B	1/3	1/3	-2/3	0	0	0	0	0.00
C	0	0	0	-5/3	1/3	4/3	0	0.00
D	0	0.00	0	0	0	0	0.00	0.00

Table 3-4. The ratings are normalized for each user, by subtracting from each rating the average rating of the user

Let's consider S_{xy} as items rated by both users x and y, and r_x, r_y … as the average ratings of x, y in the formula described as follows:

$$\boldsymbol{sim(x,y) = \frac{\sum_{s \in S_{xy}} (r_{xs} - \overline{r_x})(r_{ys} - \overline{r_y})}{\sqrt{\sum_{s \in S_{xy}} (r_{xs} - \overline{r_x})^2} \sqrt{\sum_{s \in S_{xy}} (r_{ys} - \overline{r_y})^2}}}$$

The cosine of the angle between user A and user B is described as follows:

$$sim(A,B) = \frac{\left(\frac{2}{3}\right) x \left(\frac{1}{3}\right)}{\sqrt{\left(\frac{2}{3}\right)^2 + \left(\frac{5}{3}\right)^2 + \left(\frac{-7}{3}\right)^2} \sqrt{\left(\frac{1}{3}\right)^2 + \left(\frac{1}{3}\right)^2 + \left(\frac{-2}{3}\right)^2}} = 0.09$$

The cosine of the angle between user A and user C is described as follows:

$$sim(A,C) = \frac{\left(\frac{5}{3}\right) x \left(\frac{-5}{3}\right) + \left(\frac{-7}{3}\right) x \left(\frac{1}{3}\right)}{\sqrt{\left(\frac{2}{3}\right)^2 + \left(\frac{5}{3}\right)^2 + \left(\frac{-7}{3}\right)^2}\sqrt{\left(\frac{-5}{3}\right)^2 + \left(\frac{1}{3}\right)^2 + \left(\frac{4}{3}\right)^2}} = 0.56$$

Per the centered cosine similarity measure, user A and user C are much further apart than user A and user B, and neither pair is very close. Both these observations make intuitive sense, given that users A and C disagree on the two movies they rated in common, while user A and user B give similar scores to the one movie they rated in common. It is because after normalization the blank/unknown ratings are considered as average ratings. Also, the centered cosine handles tough-raters and easy-raters.

Now we predict the unknown ratings from the similarity metric to provide recommendations for any given user. Suppose r_x is the vector of user x's ratings, and consider N to be the set of k users most similar to user x who have also rated item i. Once the set of K neighbors to user x are found, then prediction for item s of user x can be made either using option 1 or option 2. Of course, many other techniques and tricks are possible.

- **Option 1:** The simplest prediction is to take the average ratings of all the users for item i in the neighborhood N, and consider it as the estimate of the rating of user x for item i. There might be a range of similarity values within the neighborhood of user x, some users who are highly similar to the user x and a few users who are not that similar to the user x. But option 1 is very simple, and it ignores the actual similarity values between users. Therefore, option 2 is used to weight the average ratings by the similarity values.

$$r_{xi} = \frac{1}{k} \sum_{y \in N} r_{yi}$$

- **Option 2:** It is a weighted average overcoming the simplicity issue of option 1. It looks at the neighborhood N and for each user y, in the neighborhood N, it rates y's rating for item i by the similarity of x and, x and y, and then it just normalizes it by taking the sum of the similarities, and that gives a rating estimate for user x and item i. In the formula below, S_{xy} stands for $sim(x, y)$.

$$r_{xi} = \frac{\sum_{y \in N} s_{xy} \cdot r_{yi}}{\sum_{y \in N} s_{xy}}$$

The techniques that have been illustrated and used so far for finding similar users are called user-user collaborative filtering. It is important to note that any of the above-mentioned techniques can be used on columns of the utility matrix to find similar items, is called item-item collaborative filtering.

With the emergence of LMSs and other educational software in education and higher education, educational institutions are offering RSs as part of their services to better meet the needs of the individual student. In Afghanistan, according to the second national strategic plan of the MoHE, LMSs are planned to be implemented into higher education systems to enrich and enhance educational settings. The pilot phase of introducing and implementing

LMSs in education has already been implemented in a few major universities that are equipped with the necessary infrastructure and have IT centers, mainly, in faculties of Computer Science, Journalism, etc. which deal with internet and IT. The author proposes that policymakers collect LMSs data with the aim of automating some of the strategies and enhance the educational settings through data mining and RSs applications.

Moreover, it is possible to transform high school students' marks for grades 10, 11, and 12 in a matrix which then can be used in RSs. Let's consider the following scenario: item, user and transaction are the three required entities/objects in the collaborative recommendation. It is possible to consider high school students as the user object, high school subjects as the item object, and high school students' marks as the transaction object. Then preprocess and normalize these data so they become acceptable for RSs.

In addition to the collaborative filtering approaches, other techniques can be used/combined to enhance the data and recommendations. For example, experts in the field can streamline the classification of more than a hundred fields of study into main streams e.g. sciences and social sciences, and then determine the importance/relevance of each high school subject for the fields of study or for the identified streams.

Furthermore, it is possible to take the high school students' socio-demographic attributes such as age, gender, marital status, family literacy and economic status, high school status (public or private), high school location (urban or rural), whether the student attended supportive courses, etc. These students' attributes can then be assigned proper weights representing their importance.

Also, the application of RSs might be integrated in the *Kankor practice test application* (*e-Kankor*). The *e-Kankor* simulates the *actual Kankor* for the Kankor candidates to become familiar with the *actual Kankor* test and methods through exam cram techniques. In *e-Kankor* application, the Kankor candidates might play the role of the user object, the fields of study might play the role of the item object, the test result and the candidate's choice of fields of study might play the role of the transaction object. Then through RSs it might be feasible to suggest reasonable fields of study to a candidate based on the similarity of his/her performance to that of other candidates, and to even answer other research questions. The *e-Kankor* will generate data explicitly and implicitly while the Kankor candidates are taking the tests by analyzing the questions that the candidate answers, etc. These data could be useful for data mining applications – enabling policymakers and high school advisors to systematically review and analyze the candidates' learning processes through their actions and behaviors.

3.5. SUMMARY

More than ever, data is turning into a precious resource for any nation or organization. Data driven solutions like Knowledge Discovery from Data (KDD), Data Mining, Educational Data Mining, Data Science and Big Data are shaping how our world functions. It is the data that enables organizations to explore and explain the past and predict the future through improved business intelligence. Data in the developing world is taking a more robust role in determining their path towards prosperity by creating opportunities for entrepreneurship,

innovation, and better hopes of independence for a future generation that is more optimistic and futuristic. The developing world applies state-of-the-art architectures to their most complex problems, be it social, economic, political, or environmental and at all levels: local, national, regional and global.

In this chapter, the research provided scenarios (and discussed gaps and challenges) on how data can be used as a resource in the context of education in Afghanistan. Due to lack of data culture, the real power of data is not understood at all. However, the archived data collected from various sources in the education system can play a vital role for establishing the culture of data. Below are some examples of data which is currently disregard as value-less, but which can prove to be cornerstones for informing decision making through business intelligence.

- **Match High School Data with Kankor Data:** To find out the relation between high school marks and the success rate of candidates in the *Kankor* exam, it is required to merge and compare the candidates' high school marks with the Kankor data. They currently exist as two separate datasets without any mechanism for establishing a relationship between the Ministry of Education and Ministry of Higher Education to understand what the data trends might inform them. A proposed solution is to bring these two datasets together using the following common attributes: First Name, Father's Name, Province, High School Name and Graduation Year. This comparison will demonstrate not only the correlation of the two datasets but will also serve as valuable intelligence in policy discussions. For instance, no government or non-government entity has the knowledge to show whether or not the use of high school grades in lieu of *Kankor* exam can be justified. This might be the beginning of an enlightened discussion for those who are proposing to eliminate *Kankor* exam for higher education admission.
- **Named Entity Recognition:** The Kankor participants data since 2003 consist of 65,000 unique male and female "First Names"; more than 29,000 unique "Family Names"; more than 61,000 unique "Father's Names", and finally more than 65,000 unique "Grand Father's Names". These names represent almost all the common names used in Afghanistan. These names plus province, district, and village names together with high school and educational institution names could be used as a rich dataset in Text Mining, mainly, Named Entity Recognition (NER) to recognize 'Person Names', "Locations" and "Educational Institutions" within unstructured Persian/Dari text.
- **Autofill Missing Gender Values using Identical Names:** Out of 1.5 million records of Kankor data from 2003 to 2015, for around 200,000 records, the gender label is missing. This situation poses a challenge for data quality and should be addressed systematically through a structured process. There are many Kankor applicants with the same "First Name" e.g. more than 11,000 were named "Zabiuallah". It is assumed that the gender for at least one of the identical names is identified. That gender value is used as a reference to autofill the gender for the identical names whose gender value is missing. This process is explained in detail in the next chapter.

- **Autofill Missing High School Geographical Location:** As in the case of gender issue, for 660,547 records the location of high schools is missing which naturally causes a large amount of data to have gaps. The algorithm used to autofill the missing gender values from existing data can also be used to autofill the missing location of high schools. This process is explained in detail in the next chapter.
- **Popular Male and Female Names in Afghanistan:** The Kankor participants' "First Names" represent almost all the common names used in Afghanistan. These names can be used to find popular male and female names.

Furthermore, there is a large amount of data available for mining purposes in the education sector. But the methods that the educational institutions use to store their data only enable them to achieve basic insights. However, these basic insights do not help policy makers and educational institutions to improve the educational settings such as to predict the right fields of study for high school graduates, to identify first year university students who are at high risk of attrition or failure, or to recommend suitable courses for university students.

It can be concluded that there are numerous opportunities for the application of educational data mining in Afghanistan. Among other things, it can be used to predict suitable disciplines for high school graduates, help policy makers in shaping the education system, evaluate how related the candidates' five choices of disciplines in the Kankor are, classify more than a hundred disciplines into clusters and allow early warning systems to identify university students who are at high risk of attrition.

Education Management Information System (EMIS) at Ministry of Education (MoE) and Higher Education Management Information System (HEMIS) at Ministry of Higher Education (MoHE) can be utilized together to provide the data for Data Mining applications.

The main focus of this research is to equip high school advisors with proper tools, so they can provide suitable guidance to the Kankor candidates concerning their academic career. Therefore, as part of this research more than **1.5 million records of participants** who attended the Kankor exam between 2003-2015 as well as high school data for **6,000 high school graduates** was collected. High school graduates' information included biographical information as well as their marks for grades 10, 11 and 12. In chapter 4, various techniques are used to cleanse and transform the collected data and make it ready for further analysis activities. In chapter 5, the processed data is used to gain a better understanding of the data prior to employing data mining and recommender system techniques. Chapter 6 of this research focuses on a comprehensive analysis of findings, solutions, recommendations as well as key trends involving various direct and indirect stakeholders, institutions, technologies, methodologies and an overall summary of the research outcomes.

Chapter 4

Chapter 4. DATA PREPROCESSING AND CLEANSING

Data preprocessing and cleansing are very critical and useful for data mining purposes (Larose and Larose 2015) and even prior to any data analysis activities. The author of this thesis identified and corrected inconsistent data, identified and addressed outliers and smoothed out noisy data, filled in missing values, resolved inconsistencies, removed duplicated records, added additional calculated columns and measures and then combined data from multiple sources into a coherent and consolidated dataset – described in the following sections of this chapter in detail. These are essential, as they lead to transformation of data into an organized structure and generation of accurate, valuable and detailed insights.

4.1. OVERVIEW

Inconsistent Latinization and Anglo-Saxonization of names is a major issue with the naming conventions and inconsistent use of labeling is exacerbated by unnecessary leading and trailing of spaces as well as blank spaces in the middle of a string, unwanted rows and columns, and other unwanted irregularities in the data which negatively impact data analysis activities and generation of accurate reports.

Prior to performing any data analysis activities, such as exploring and visualizing the data, building a data mining model, or even generating reports, it is often required to preprocess and cleanse the data. Data preprocessing and cleansing is to manipulate the data when it is not in optimal shape and includes tasks such as removing unnecessary leading and trailing spaces as well as blank spaces in the middle of a string, removing unnecessary rows and columns, sorting and filtering columns, splitting or merging columns, adding custom columns, aggregating or summarizing data, converting data types, replacing inconsistent values, pivoting or unpivoting data, combining multiple data sources into a big uniform master dataset, and several other related transformation activities.

Sometimes, transformation tasks are relatively straightforward and there are specific features that do the job such as quickly removing duplicates by using *Remove Duplicates* transformation; or, using *Trim* transformation to clean up the leading and trailing spaces. At other times, custom methods are required to do the job such as, to remove inessential blank spaces in the middle of the text or to identify the missing values.

The basic steps for transformation and cleansing activities are as follows:

- create a backup copy of the original data;
- import and load the data for preprocessing;
- ensure that the data is in a tabular format with all columns and rows visible;
- Keep columns and rows of interest and exclude the columns and rows that do not require preprocessing and transformation;

- first perform straightforward tasks, which do not require column manipulation e.g., trimming, sorting, filtering, replacing, removing duplicates, and others;
- and then perform tasks that do require column manipulation and or custom method.

Data transformation and cleansing is a tedious, repetitive, and time-consuming process and may be the hardest part of data science. But it is essential prior to any data analysis activities (Osborne 2012). There are powerful tools with capabilities to automate the transformation and cleansing tasks – sometimes on their own and sometimes in combination with some other tools. These types of tools include:

- Microsoft Excel predefined formulas such as LOOKUP(), VLOOKUP, HLOOKUP, INDEX(), MATCH(), OFFSET(), INDIRECT(), AGGREGATE(), LEFT(), LEN(), TRIM(), CLEAN(), SUBSTITUTE(), helper columns, other Excel features, and many more. While most Excel users tend to use formulas, the complexity of these formulas varies depending on the user's experience (Puls and Escobar 2015; Winston 2016) and these formulas start to slow down the calculation speed with large datasets.
- Visual Basic for Applications (VBA) and Macro can help one to create powerful and dynamic data extraction and transformations (Cook 2015). VBA and Macro techniques tend to be used by advanced users due to the discipline required to truly master them (Puls and Escobar 2015).
- Structured Query Language (SQL) is another powerful language for manipulating and querying data that is often used to make intelligent personal or business decisions, and it can be extremely useful for selecting, sorting, grouping, and transforming data (Linoff 2015). The reality, however, is that this language is also typically only used by advanced users.

All these tools have something in common and for many years, they have essentially been the only tools available for cleaning and transforming data into something useful. Despite their usefulness, many of these tools also have the following two serious weaknesses (Puls and Escobar 2015):

1. They require time to build a solution,
2. and time to master the techniques.

Nowadays there are powerful and easy-to-use data cleansing tools that solves these two weaknesses such as Power Query. The Power Query experience is made possible by M, a formula language for data transformation – M stands for mashup queries. It was designed to be an easy-to-use data transformation and manipulation tool, even for less technical users through its most intuitive Graphical User Interface (GUI). To cleanse the same data source or similar ones periodically – the applied steps can be used to automate the entire process. With Power Query, it is very easy to maintain the transformation, as it shows each step of the process, which can be reviewed or updated later.

Once the steps for transformation and cleansing the data are set up, the applied steps are recorded and can be repeated over and over every time the data changes – It saves one a massive amount of time (Christopher Webb and Limited 2014). Power Query uses the

principle of defining the transformation and manipulation process once and applying it anytime it is needed.

Power Query is essentially an ETL (Extract, Transform and Load) tool similar to *Talend Open Studio for Data Integration*, *IBM Infosphere Information Server*, *SQL Server Integration Services (SSIS)* and other ETL tools. Its function is to *extract* data from various sources, *transform* it as desired, and then *load* it into one of four places: Excel tables, the Power Pivot Data Model, Power BI, Connections only (Puls and Escobar 2015). Connections only are simply helper queries that can be used by other queries. Microsoft first launched Power Query as an Excel Add-In, and then integrated this incredible product into Excel 2016 (Excel team 2015). Microsoft has also released a standalone Power BI Desktop application for sourcing and modeling data for free, which exists together with Power Query, Power Pivot, DAX (Data Analysis Expressions) formulas and Power View in a standalone application, removing almost all the Microsoft Excel constraints (Allington 2015; Collie and Singh 2016). It is an all in one product and provides a richer and interactive way to further analyze and visualize information in addition to transforming and handling big datasets more efficiently.

This chapter details the preprocessing and transformation of the Kankor and High School data collected by the author of this thesis, explained in Chapter 1, Research Methodology section. Furthermore, it provides a deep dive into the capabilities of Power Query and shows how to unlock data wrangling capabilities either exposed by the GUI or tailored and customized M language. The goal as illustrated in the following diagram is to pull the collected data into Power Query, *preprocess and transform* them as desired, and finally *load* them into Power BI or other business intelligence tools for further analysis activities:

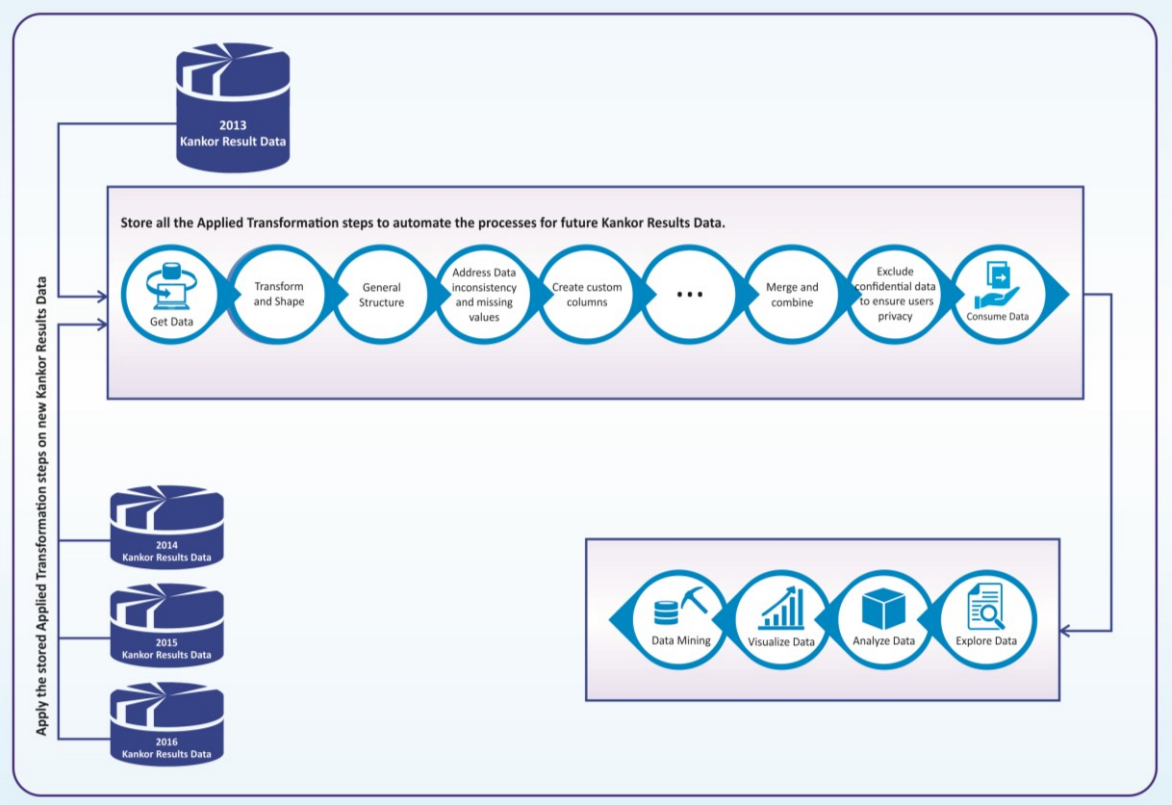

Figure 4-1. This diagram shows the plan for cleansing of Kankor data prior to further data analysis activities.

4.1.1. Remove Unnecessary Spaces

Study of *Kankor and other* datasets reveals that values for the candidates' personal information such as First Names and Father's Names as well as High School Names, Geographical Location of High Schools, Higher Education Institutions and other attributes contain leading and trailing spaces and even blank spaces in the middle that are unnecessary. Moreover, sometimes the values contain Zero Width Space characters (*Unicode character set values 200B, 200C, 200D and FEFF*) and other inessential characters.

The unnecessary spaces and characters can sometimes cause unexpected results when data is sorted, filtered, searched or matched. To remove these nonessential spaces and characters, a combination of text transformations such as, *Trim*, *Clean*, *Replace*, and *Substitute* is used.

The built-in *Trim* formula in Microsoft Excel can be used to remove extraneous spaces from data whether spaces are at the beginning, middle or at the end of the text – except single spaces between the words. However, the built-in Power Query *Trim* transformation only removes spaces from beginning and end of the text, and not the additional spaces in the middle. Hence, the following custom function has been used to remove blank spaces in the middle of text. This custom function initially was introduced by Ken Puls (Puls 2015). The author of this thesis has taken the code, modified it, and called it *fxCustomTrim*, as shown below.

```
fxCustomTrim = (txtInput as text) as text =>
    let
        delimiter = " ",
        /* Returns a list containing parts of a text value
         * that are delimited by the 'delimiter'.
        */
        splitText = Text.SplitAny(txtInput, delimiter),
        // Selects the items that match a condition.
        removeBlankItems = List.Select(splitText, each _ <> ""),
        /* Returns a text value that is the result of joining
         * all text values with each value separated the 'delimiter'.
        */
        result = Text.Combine(removeBlankItems, delimiter)
    in
result
```

Table 4-1. This Power Query function removes leading, trailing and extraneous blank spaces in the middle of text.

It is a slightly different version that can remove leading and trailing spaces as well as blank spaces in the middle of text strings, as represented in the following figure for further clarification.

Province Names	Length	Comment
بادغیس	8	Unnecessary leading spaces
بادغیس	6	No additional spaces
بادغیس	10	Unnecessary trailing spaces
مزار شریف	15	Unnecessary leading and trailing spaces
مزار شریف	12	Unnecessary spaces in the middle
مزار شریف	9	No additional spaces

Inessential Spaces were Removed

Province Names	Length	Comment
بادغیس	6	Additional spaces removed from the values
بادغیس	6	Additional spaces removed from the values
بادغیس	6	Additional spaces removed from the values
مزار شریف	9	Additional spaces removed from the values
مزار شریف	9	Additional spaces removed from the values
مزار شریف	9	Additional spaces removed from the values

Figure 4-2. Extra spaces were removed from beginning, end of the text and even additional spaces in the middle.

4.1.2. Rectify Unicode Characters

Moreover, study of *Kankor and other datasets* shows that some values for the candidates First Names, Father's Names, High Schools and their Locations and other attributes are indistinguishable by humans. For the example both of the following names in Dari/Persian بادغیس and بادغیس represent Badghis province. Both names look the same and are indistinguishable by human. However, they are completely different names for computers when data is used for comparison, filtering or sorting. In the former بادغیس for the letter Yeh (ی) *Arabic Letter Farsi Yeh* with *Unicode 06CC* is used. However, in the latter بادغیس for the letter Yeh (ي) *Arabic Letter Yeh* with a completely different *Unicode 064A* is used.

The *Arabic letter Yeh* always has two dots underneath (except *Alef Maskura* with *Unicode 0649*), while the *Arabic Letter Farsi Yeh* has dots only in initial and medial positions – and has no dots underneath when used at the end of a word.

Furthermore, these two names کابل and کابل represent Kabul province, the capital of Afghanistan also are not distinguishable by humans. However, for machines they are

considered completely different names when the data are used for comparison or other preprocessing and transformation purposes. In the former کابل for the letter Kaf (ک) *Arabic Letter Keheh* with *Unicode 06A9* is used. However, in the latter كابل for the letter kaf (ك) *Arabic Letter Kaf* with a completely different *Unicode 0643* is used. It is worth mentioning that the *Arabic letter Kaf* when appears at the very end of the text string can be differentiated by human, otherwise, it is not.

This issue has been solved by using the built-in Power Query *Replace* transformation to make all the Unicode consistent prior to performing other transformation steps. For further clarification purposes, these scenarios are illustrated in the following figure. Essentially, the goal is to transform and cleanse data from the top table represented in red color to bottom table represented in green color in the figure.

Province	Province in English	1st char (unicode)	1st char (unicode)2	1st char (unicode)3	1st char (unicode)4	1st char (unicode)5	1st char (unicode)6
بادغیس	Badghis	U+0628 = ب	U+0627 = ا	U+062F = د	U+063A = غ	U+06CC = ی	U+0633 = س
بادغیس	Badghis	U+0628 = ب	U+0627 = ا	U+062F = د	U+063A = غ	U+064A = ي	U+0633 = س
غزنی	Ghazni	U+063A = غ	U+0632 = ز	U+0646 = ن	U+06CC = ی		
غزني	Ghazni	U+063A = غ	U+0632 = ز	U+0646 = ن	U+064A = ي		
نیمروز	Nimruz	U+0646 = ن	U+06CC = ی	U+0645 = م	U+0631 = ر	U+0648 = و	U+0632 = ز
نيمروز	Nimruz	U+0646 = ن	U+064A = ي	U+0645 = م	U+0631 = ر	U+0648 = و	U+0632 = ز
کابل	Kabul	U+06A9 = ک	U+0627 = ا	U+0628 = ب	U+0644 = ل		
كابل	Kabul	U+0643 = ك	U+0627 = ا	U+0628 = ب	U+0644 = ل		
کاپیسا	Kapisa	U+0643 = ك	U+0627 = ا	U+067E = پ	U+06CC = ی	U+0633 = س	U+0627 = ا
کاپیسا	Kapisa	U+06A9 = ک	U+0627 = ا	U+067E = پ	U+06CC = ی	U+0633 = س	U+0627 = ا
کاپیسا	Kapisa	U+0643 = ك	U+0627 = ا	U+067E = پ	U+064A = ي	U+0633 = س	U+0627 = ا
کاپیسا	Kapisa	U+06A9 = ک	U+0627 = ا	U+067E = پ	U+064A = ي	U+0633 = س	U+0627 = ا

Unicode were Rectified

Province	Province in English	1st char (unicode)	1st char (unicode)2	1st char (unicode)3	1st char (unicode)4	1st char (unicode)5	1st char (unicode)6
بادغیس	Badghis	U+0628 = ب	U+0627 = ا	U+062F = د	U+063A = غ	U+06CC = ی	U+0633 = س
بادغیس	Badghis	U+0628 = ب	U+0627 = ا	U+062F = د	U+063A = غ	U+06CC = ی	U+0633 = س
غزنی	Ghazni	U+063A = غ	U+0632 = ز	U+0646 = ن	U+06CC = ی		
غزنی	Ghazni	U+063A = غ	U+0632 = ز	U+0646 = ن	U+06CC = ی		
نیمروز	Nimruz	U+0646 = ن	U+06CC = ی	U+0645 = م	U+0631 = ر	U+0648 = و	U+0632 = ز
نیمروز	Nimruz	U+0646 = ن	U+06CC = ی	U+0645 = م	U+0631 = ر	U+0648 = و	U+0632 = ز
کابل	Kabul	U+06A9 = ک	U+0627 = ا	U+0628 = ب	U+0644 = ل		
کابل	Kabul	U+06A9 = ک	U+0627 = ا	U+0628 = ب	U+0644 = ل		
کاپیسا	Kapisa	U+06A9 = ک	U+0627 = ا	U+067E = پ	U+06CC = ی	U+0633 = س	U+0627 = ا
کاپیسا	Kapisa	U+06A9 = ک	U+0627 = ا	U+067E = پ	U+06CC = ی	U+0633 = س	U+0627 = ا
کاپیسا	Kapisa	U+06A9 = ک	U+0627 = ا	U+067E = پ	U+06CC = ی	U+0633 = س	U+0627 = ا
کاپیسا	Kapisa	U+06A9 = ک	U+0627 = ا	U+067E = پ	U+06CC = ی	U+0633 = س	U+0627 = ا

Figure 4-3. Different Unicode for the letter Yeh and Kaf were rectified.

4.1.3. Fix Inconsistencies for Attributes with Limited Values

Study of *Kankor and other datasets* also show that there are attributes with limited set of possible values such as Gender attribute with only male and female identifiers; High Schools' Geographical Location attribute with only 34 possible values of Afghanistan provinces. But almost all these attributes suffer from data inconsistencies. This situation poses a challenge for data quality and completeness and requires to be fixed thoroughly through a structured process.

Inconsistent use of labelling for Gender attribute was common. For example, m, male, boys, man, 1, and other equivalent Persian identifiers were all used to represent the *male* gender of the candidates. Overall, multiple different labels were used to indicate the male and female genders of candidates. Generating accurate gender-wise reports are not practical from such messy data unless they are prepressed and cleansed. The inconsistency of these identifiers has been fixed with correct labelling, using the Power Query *Replace* transformation operation.

Furthermore, graduates from high schools across the country are classified into 34 geographical locations. For geographical identification purposes, the classification comprises all the locations by province. Since province names are not standardized, e.g., the province of *Balkh* is also identified by its other common names, such as, *Mazar-e-Sharif* same as *Balkh* in the line above, sometimes even with various formats and typos such as, *Mazar-e Sharif*, *Mazar-i-Sharif*, *Mazar*, and others. While all these names are used to identify the same province, they are used interchangeably at will.

The built-in transformations in Power Query are not good for resolving this abnormality efficiently for several reasons:

- This attribute represents geographical location (province) of high schools across the country. Since Afghanistan comprises of 34 provinces, there are 34 possible values for this attribute. As mentioned above provinces were written in several formats and variations. Therefore, manually repeating the *Replace* transformation is time-consuming and inefficient.
- Also, this method cannot detect and handle new inconsistent identifiers for *Balkh* province or other provinces in the future.

This requires further research and should be addressed systematically through a structured and efficient process, explained in the following section.

4.1.4. Fix Inconsistencies of Province Attribute

As already mentioned above, study of Kankor data shows the province names that represent geographical locations of high schools are not unified such that for the same province, different common names and formats have been used. These inconsistencies do not let organizations to generate accurate reports e.g. they cannot easily produce reports to indicate the number of candidates graduated for a given province, or to compare candidates' province-wise, or to answer other types of questions which require geographical data.

It is not efficient to perform Power Query *replace* transformation manually to rectify the existing inconsistencies, as elaborated above. On the other hand, it is not practical to add a new column directly into the original source data for several reasons:

- Suppose there are tens of thousands of candidates who attended Kankor in Balkh province, and an average of five different common names and formats were used to represent the province name. One is required to look at each record and correct or replace the inconsistent formats manually. Inspecting and replacing manually is a time-consuming process because there are 34 provinces, and there is an average of five different inconsistent variations and formats for each one.
- Moreover, when there are other new inconsistent province names in the future – this method cannot detect and handle them automatically and it requires manual processing and replacing to make the values for the provinces uniform.

An approach, named, *self-referencing join*, is carried out to address this dilemma systematically. This approach was introduced by Matt Allington (Allington 2016). The method used by the author of this thesis to overcome the identified challenge efficiently, as explained in the following steps:

1. The consolidated Kankor dataset was imported into Power Query, named, *Table 1*.
2. Another reference query from *Table 1* was created. All other columns except the Province column representing location of high schools were removed. Then *Remove Duplicates* transformation was used to narrow down 1.5 million records of Kankor data only to their unique common province names. The result from this step was outputted to Microsoft Excel spreadsheet, named, *Table 2*.
3. A new column was added at the end of *Table 2*, named, *Corrected Province Name*. At this stage it is possible to enter correct values for the *Corrected Province Name* column, as shown in the following figure.

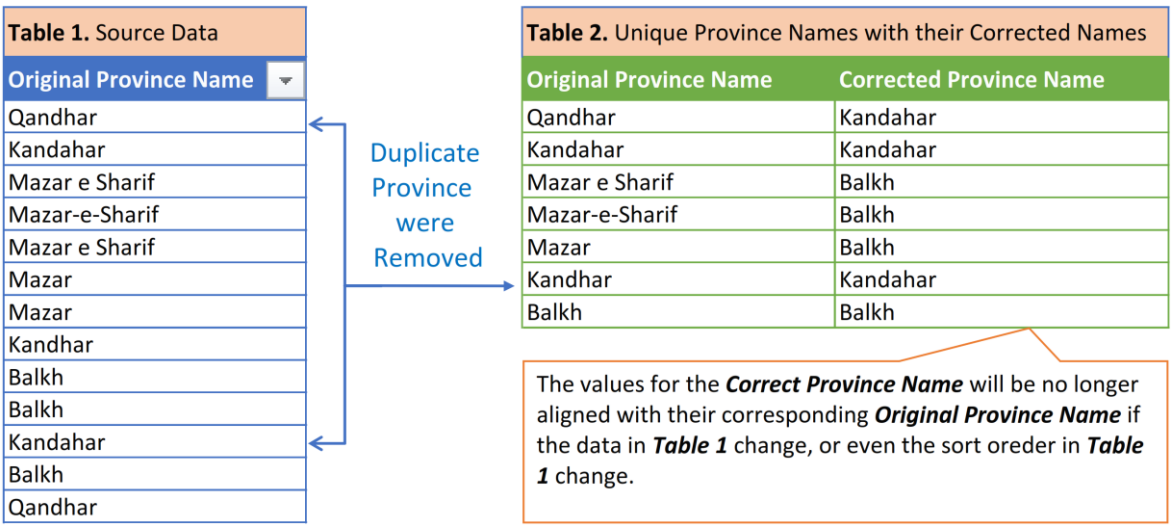

Figure 4-4. adding a new column, named Corrected Province Name, and then proper values were entered instantly.

However, the problem is that the values entered for *Corrected Province Name* are not logically linked to their corresponding *Original Province Name*. If the data or the sort order in *Table 1* changes, then the values for the *Corrected Province Name* are no longer aligned

with their corresponding *Original Province Name*, as is illustrated in the following figure in red color for further clarification.

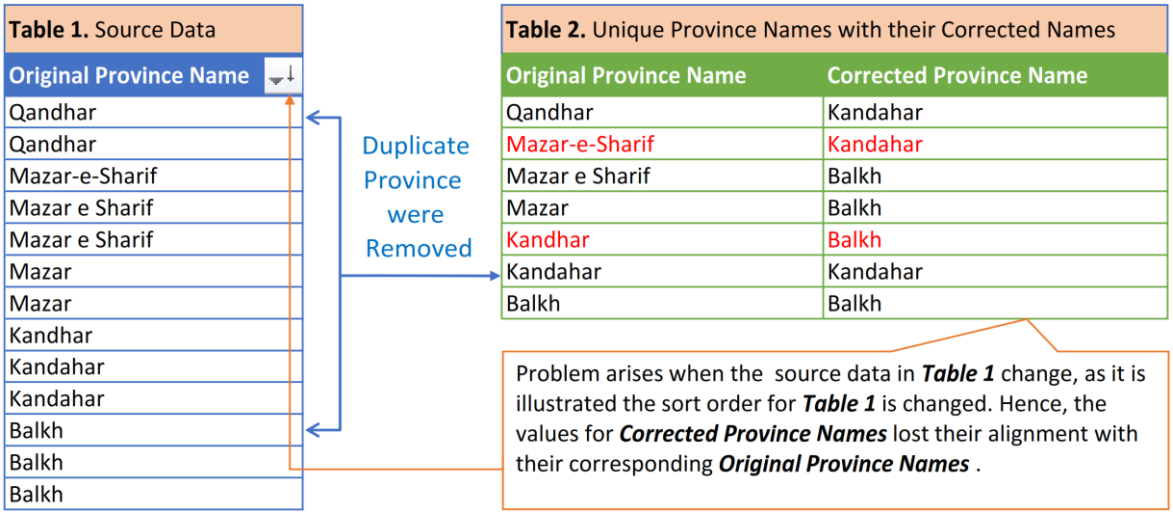

Figure 4-5. The values for the Correct Province Name lost their alignment with their corresponding Original Province Name as soon as the order of the source data in Table 1 changed.

Self-referencing join is carried out to keep the position/reference of values in the *Corrected Province Name* column aligned with their corresponding *Original Province Name*; regardless of whether the data or the order of data in *Table 1* change, as described by the following steps.

4. A self-join is a query in which a table is joined and compared to itself. Hence, from *Table 2* including *Corrected Province Name* another *connection only query* was created, named *Table 2 connection only*.
5. The query for *Table 2* that was created in *step 2* was edited and merged with *Table 2 connection only*. The merge adds a new column (contains all *Table 2 connection only* columns) to *Table 2*. From this new column, the *Corrected Province Name* column was selected, and the option *Use original column name as prefix* was deselected so that the *Corrected Province Name* has the same name as in its original *Table 2*. With these settings, the data was expanded, and the query was loaded back into Microsoft Excel spreadsheet.

The data outputted to the Microsoft Excel spreadsheet from *step 5* looks the same as in *step 3*. However, there is one important difference – the values for *Corrected Province Name* will be properly linked with their corresponding *Original Province Name*, even if the data or the sort order for *Table 1* change.

Now that the steps are set up, one can gradually add the correct values for the *Corrected Province Name* for each corresponding inconsistent value of *Original Province* Name as well as change the order of the data in *Table 1* without any issue, as illustrated in the following figure.

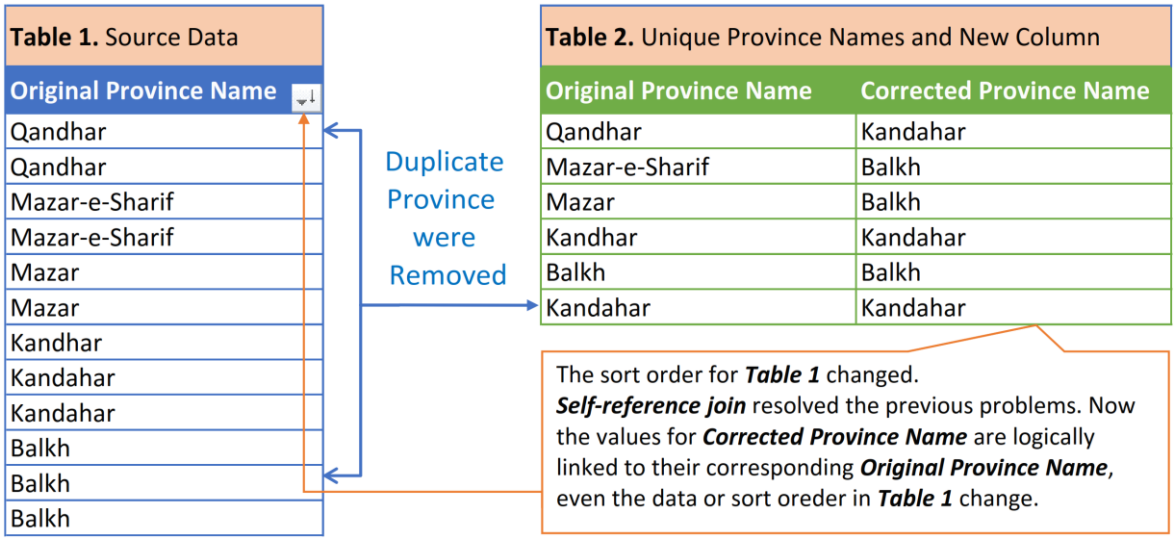

Figure 4-6. The values for the Corrected Province Name are now logically linked and aligned with their corresponding Original Province Name, even if the sort order of Table 1 changes.

Also, new data (e.g. Kabol, Kabul, Kabul and Mazar-i-Sharif) can be added to the source data, named *Table 1*, even the order of source data can be changed (e.g. the order changed from *descending* to *ascending*), and it will all work as expected, as shown in the following figure for further clarification purposes.

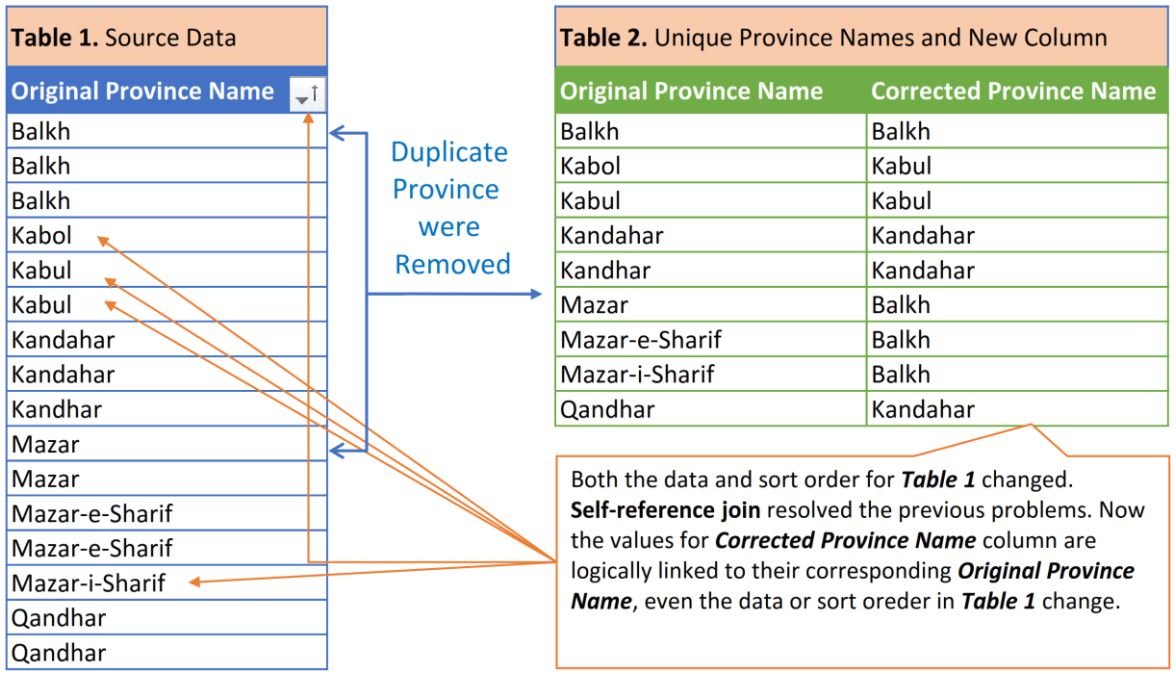

Figure 4-7. New data added to the source data as well as the order of source data changed from descending to ascending, and the Corrected Province Name are aligned with their corresponding original values.

The approach described above will work just fine if the following two conditions are maintained (Allington 2016).

1. There is a unique ID for each record,
2. Duplicate data are not accidentally loaded.

In our case, since initially the *Remove Duplicate* transformation is carried out on *Table 2* – This transformation operation guarantees uniqueness of the records.

4.1.5. Audit Data to Remove Duplicate Candidates

Duplicate data are one the biggest problems in any data analyst's life and a common problem which require to be addressed prior to data analyses. Duplicate data significantly leads to inaccurate facts and figures, reports and insights as well as it negatively impacts informed decisions and accuracy of predictive methods. Thus, it deemed efficient to make sure that that data is accurate, complete and duplicate free.

Study of Kankor data reveals that for some years there are two datasets. The first dataset includes all the candidates who attended Kankor regardless of their final results, success or failure, as shown in the following figure.

First Dataset - Original Kankor Data

Kankor Code	First Name	Last Name	Gender	Kankor Score	Result
F00000001	عبدالوحید	وحیدی	Male	290	Computer Science
F00000002	سوزان	رحیمی	Female	195	Resultless
F00000003	سلیمان	کریمی	Male	205	Resultless
F00000004	مریم	امیری	Female	285	Economics
F00000005	منیژه	سادات	Female	210	Resultless
F00000006	نرگس	رضایی	Female	190	Resultless
F00000007	فرید	عزیزی	Male	310	Medicine
F00000008	فرهاد	فرهادی	Male	200	Resultless

Figure 4-8. First dataset contains all the Kankor candidates' info and results announced for the first-round.

The second dataset includes only a small portion of the candidates from the first dataset. These candidates either failed or announced result-less in the first round of Kankor but succeeded in the second round after reconsideration. For some of these candidates their scores and results updated and got admission into semi or higher education institutions, as shown in the following figure.

Second Dataset - Second Capacity

Kankor Code	First Name	Last Name	Gender	Kankor Score	Result
F00000003	سلیمان	کریمی	Male	205	Social Sciences
F00000005	منیژه	سادات	Female	210	Teacher Training
F00000008	فرهاد	فرهادی	Male	200	Pedagogy

Figure 4-9. Second dataset contains only the candidates that their Kankor results have been updated.

These two datasets need to be merged and compared using the candidates' unique Kankor code. Then the candidate's Kankor score and result from the second dataset must replace the

Kankor score and result in the first dataset only for the candidates that matched in both datasets, as this scenario is illustrated in the following figure.

Desired Output					
Kankor Code	First Name	Last Name	Gender	Kankor Score	Result Updated
F00000001	عبدالوحید	وحیدی	Male	290	Computer Science
F00000003	سلیمان	کریمی	Male	205	Social Sciences
F00000002	سوزان	رحیمی	Female	195	Resultless
F00000004	مریم	امیری	Female	285	Economics
F00000005	منیژه	سادات	Female	210	Teacher Training
F00000006	نرگس	رضایی	Female	190	Resultless
F00000007	فرید	عزیزی	Male	310	Medicine
F00000008	فرهاد	فرهادی	Male	200	Pedagogy

Figure 4-10. First Dataset was merged with second dataset and the Kankor score and results were updated from the second dataset.

There are numerous tools available to address the identified scenario such as Relational Database Management Systems (RDBMSs); Microsoft Excel built-in formulas such as LOOKUP, VLOOKUP, HLOOKUP, or INDEX/MATCH formulas; VBA and others which have their limitations. For instance, **Microsoft Excel formulas** are great. Using them, however, often means adding thousands of formulas to the workbook, which increases the file size and calculation time. Moreover, Excel cannot handle more than 1,048,576 rows in a sheet.

Here is the SQL statement in MySQL to audit and merge both the datasets to update the result from the second one, but for years with two datasets other preprocessing and transformation steps should also be carried out in addition to this statement:

```
UPDATE firstdataset
       JOIN seconddataset USING ("Kankor Code")
SET    firstdataset.result = seconddataset.result;
```

Table 4-2. This SQL statement merges the two datasets and updates the content of the result column from the second dataset.

The Power Query merge feature creates a dynamic process that runs faster and is more intuitive than Excel LOOKUP formulas and other tools. It basically uses SQL joins and supports *inner*, *outer*, *left*, *right*, *full*, and *anti* joins. The following Power Query custom code is used to address the identified Kankor scenario efficiently, as reflected in the above figure.

```
let
    Source = firstDataset,
    #"Merged with secondDataset" = Table.NestedJoin(Source,{"Kankor
Code"},secondDataset,{"Kankor Code"},"NewColumn",JoinKind.LeftOuter),
    #"Expanded Result Column from secondDataset" = Table.ExpandTableColumn(#"Merged
with secondDataset", "NewColumn", {"Result"}, {"NewColumn.Result"}),
```

```
    #"Updated Result Column" = Table.AddColumn(#"Expanded Result Column from
secondDataset", "Result Updated", each if [NewColumn.Result] = null then [Result] else
[NewColumn.Result]),
    #"Removed old Result Columns" = Table.RemoveColumns(#"Updated Result Column",
{"Result", "NewColumn.Result"})
in
    #"Removed old Result Columns"
```

Table 4-3. This Power Query code merges the two datasets and updates the result column from the second dataset.

The above Power Query code initially merges the first dataset with the second one through *LEFT OUTER JOIN* and it returns all records from the left table (first dataset), and the matched records from the right table (second dataset). The Result column is NULL from the right side, if there is no match. The tricky part is to update the result from the second dataset by adding a new conditional column with the following rule: If the Result in the second dataset is null then take the Result from the first dataset, otherwise, take the Result from the second dataset.

This is not the end of the story of duplicate data. A close inspection of the Kankor datasets reveals that for some years some of the datasets contain duplicate data. For example, the dataset for the year 2011 contains 5,502 duplicate records; the year 2012 contains 18,443 duplicate records; the year 2013 contains 1,058 duplicate instances; and the year 2014 contains 1,866 duplicate rows.

The following techniques were used to remove duplicate records from these datasets:

- Examination of Kankor data shows that sometimes the same candidate was recorded twice in a dataset with same Kankor score and Kankor result. One of the reasons can be that the Kankor score and result was not satisfactory for the candidate and he/she has claimed for reconsideration. But after the revision process, the candidate did not qualify. However, his/her result was added by administrators once more. The author of this thesis performed the following condition to remove such kind of duplicate records: If two records in the same dataset have the same Kankor Code, First Name, Kankor Score and Kankor Result then remove one of the records.

But sometimes after the revision process some of the candidates are entitled to a higher or lower Kankor score. This change in the Kankor score changes their admission from one field of study to another field of study or from one institution to another. In such cases the above technique does not work for removing the duplicate records. Therefore, the author of this thesis took a completely different approach into consideration, explained through the following steps:

- If two records in the same dataset had only the same Kankor Code, then both records were marked as duplicate in a new calculated column, named status.
- The "failed" and "result-less" labels in the Kankor Result column were replaced with values that appear at the very bottom of the dataset when the data is sorted by this column. It is because sometimes there are candidates who either failed in the Kankor or were announced result-less. After reconsideration, some of these candidates

become qualified and get admission into some field of study, even though their Kankor score does not change. Also, there are candidates with good Kankor scores who were announced failed or result-less. After reconsideration, even though their Kankor score does not change they become qualified and get admission into semi or higher education institutions. In the next steps, it is explained that all the duplicate instances are sorted by multiple columns and then the first instances of each record will be kept, and other instances will be deleted. These values must be changed prior to the deletion process.

- Then the data of the dataset were sorted by the following five attributes in the same priority order: initially by Status in ascending order, then by Kankor Code in ascending order, next by First Name in ascending order, followed by Kankor Score in descending order, and finally by the Kankor Result in ascending order. The Kankor score sorted in descending order because after in the reconsideration process the candidates may receive the same score or a higher score.
- Next the first instances of all those records marked as duplicate were kept and the duplicate ones were excluded from the dataset.
- Finally, the "failed" and "result-less" labels were replaced back to their original labels.

4.1.6. Audit and Match High School Data with Kankor Data (String Matching)

To find out the relation between high school marks and the success rate of candidates in the Kankor exam, it is required to merge and compare the candidates' high school dataset with the Kankor dataset.

Since there are no common keys for these two datasets, the only solution is to use the combination of following common attributes forming a composite unique key for comparison: First Name, Father Name, Province, High School Name, and Graduation Year.

Afghanistan consists of 34 provinces, hence, the values for the 'Province' attribute are limited, and the existing irregularities and variations were rectified with a structured and systematically approach independent of blank spaces in the middle of the strings, described above in this chapter. Also, the values for the 'Graduation Year' are numeric and relatively straight-forward to be preprocessed and cleansed.

However, the values for the candidates' First Names, Father's Names and High School Names contain lots of anomalies and are written in different variations. Therefore, adding and comparing the Kankor dataset with the High School dataset through traditional approaches such as exact string matches does not produce a result-set including all the common instances in both datasets; since the data (First Names, Father's Names, and High School Names) we are trying to match together is not an exact match for the following reasons.

Single Spaces Issue

Every language has its own unique rules. In Dari/Persian language, an alphabet character may take different shapes depending to its position in the string and the character which comes before or after it. For instance, when the character Aleph (ا) or (alef or alif;

Persian/Dari: الف) follows the character Re (ر) such as in زهرا, it is written separately and not connected with its previous character. But when the character الف follows the character Fe (ف) such as in فاطمه, it is connected with its previous character.

When a character is unconnected to its previous character, that does not mean that there should be a space between the two characters. In fact, spaces are not allowed between unconnected characters within one word. An example is فاطمه (there is no space between its unconnected characters). If there is only one space between the unconnected characters of a word (for example in فا طمه in which there is one additional space after the فا), the built-in trim method cannot remove that space, because the trim method is designed to remove two or more connected spaces. Thus, the names فاطمه and فا طمه will not be identical, even after the trim transformation is carried out because the built-in trim function cannot remove the single unnecessary space in the middle of strings, as it is illustrated for further clarification in the following diagram:

Name	Length	Variations	Length	Matched?	Comment
فاطمه	5	فاطمه	6	FALSE	One unnecessary trailing space
فاطمه	5	فا طمه	6	FALSE	One unnecessary space in middle
فاطمه	5	فا طمه	7	FALSE	Two unnecessary spaces in middle
فاطمه	5	فاطمه	7	FALSE	Two unnecessary leading spaces
ذبيح الله	9	ذبيح الله	11	FALSE	One unnecessary leading space
ذبيح الله	9	ذ بيح الله	10	FALSE	One unnecessary space after 1st char
ذبيح الله	9	ذبيح الله	10	FALSE	Two unnecessary spaces in middle
ذبيح الله	9	ذبيحالله	8	FALSE	Two words are connected

Additional Spaces were Removed Except Single Space in Middle

Name	Length	Variations	Length	Matched?	Comment
فاطمه	5	فاطمه	5	TRUE	Trailing space was trimmed
فاطمه	5	فا طمه	6	FALSE	Space in middle could not be trimmed
فاطمه	5	فا طمه	6	FALSE	Only one of middle spaces were trimmed
فاطمه	5	فاطمه	5	TRUE	Leading spaces were trimmed
ذبيح الله	9	ذبيح الله	9	TRUE	Leading spaces were trimmed
ذبيح الله	9	ذ بيح الله	10	FALSE	Space after 1st char could not be trimmed
ذبيح الله	9	ذبيح الله	9	TRUE	Only one of middle spaces were trimmed
ذبيح الله	9	ذبيحالله	8	FALSE	Nothing happened

Figure 4-11. The Trim method cannot remove single unnecessary spaces from the middle of Dari/Persian text.

Unicode Character Issue

Moreover, study of Kankor data reveals that some string values (names) are indistinguishable such as "یاسین", "یاسین", "یاسین" and "یاسین". In the first "یاسین" for the letter Yeh (ی) Arabic Letter Farsi Yeh with Unicode 06CC is used. However, in the second "یاسین" for the letter Yeh (ي) Arabic Letter Yeh with a completely different Unicode 064A is used,

as it is presented for further clarification in the following diagram. Unicode can be rectified using the Power Query *Replace* transformation, as explained above in this chapter.

Names	1st char (Unicode)	2nd char (Unicode)	3rd char (Unicode)	4th char (Unicode)	5th char (Unicode)
یاسین	06CC	0627	0633	06CC	0646
یاسین	064A	0627	0633	064A	0646
یاسین	06CC	0627	0633	064A	0646
یاسین	064A	0627	0633	06CC	0646

Figure 4-12. In Dari/Persian same name can be written using different Unicode characters.

To audit and compare two datasets and the comparison involves string values, the identified challenges will be a major issue and further research is needed to handle the problem of unnecessary spaces in the middle of strings, even when there is only one occurrence of unnecessary blank space. The 1.5 million Kankor participants data since the year 2003 include of almost 65,000 unique male and female First Names; almost 61,000 unique Father's Names; and more than 20,000 unique High School Names that are written in multiple variations. Therefore, the value of these attributes cannot be rectified using basic preprocessing and transformation steps. Even the *self-referencing join* technique used to rectify inconsistencies of the Province attribute, explained above in this chapter is not effective nor efficient to rectify these values. The author of this thesis proposes the following four approaches addressing the issues explained above in this section:

1. Exact String Matching,
2. Tailored Method,
3. Stripping All Spaces from String Values,
4. and Approximate String Matching.

It is worth mentioning that prior to choosing and performing any of the suggested approaches it is strongly recommended to "remove unnecessary spaces" using the *fxCustomTrim* method and to "rectify Unicode characters" using the Power Query *Replace* transformation as explained above in this chapter. These two steps help a lot fixing nearly most of the issues available in the data as well as increase accuracy of the suggested methods.

4.1.6.1. Exact String Matching

This technique is the simplest version of string matching. However, in our case the exact string matching ignores all those instances that are the same in both datasets but either the candidates First Names, Father's Names or High School Names are written in different variations e.g. contains additional unnecessary spaces such as فاطمه without any additional spaces and فا طمه with an additional space as described in the above. In conclusion, exact string matching will not match the former name "فاطمه" without any extraneous spaces to later name "فا طمه" with an extraneous space.

4.1.6.2. Tailored Method

All the existing trim methods are designed to remove all the leading and trailing spaces from a given string. Additionally, depending on the programming language, they can only remove additional spaces (more than one successive spaces) from the middle of a given string.

However, in the case of Dari/Persian there are cases one additional space is present in the names that cause the names to not match and the trim methods cannot detect and trim those additional spaces from the names. Therefore, a personalized and tailored-made method is required to be designed addressing the following issues properly:

- identifying the single unnecessary blank spaces in the middle of string values (names) properly and then remove them,
- identifying names with more than one words that are connected and then separate the connected words properly,
- and finally addressing the issue of the Unicode within the string values properly.

It is worth mentioning that designing such a tailored-made method, addressing the *first step* and *second step*, is complex and requires contribution of expertise from linguistic and other disciplines.

4.1.6.3. Stripping All Spaces from String Values

In this method, initially all the present blank spaces (including both necessary and unnecessary spaces) must be entirely removed from First Names, Father's Names, and High School Names in both datasets. Then, through exact string matching almost all instances from both datasets that are equal and are written in several variations will be matched precisely, as shown below (see **Figure 4-13**).

This method works great in comparison to other proposed approaches. Also, this method is efficient because it performs both the preprocessing steps and matching process very quickly and does not consume computer resources much in comparison to the Approximate String Matching technique.

However, it suffers from the following weakness: Dari/Persian language includes all Arabic characters and a few more. Most male and female names in Afghanistan are Arabic or have Arabic root and are written the same in both languages. A very small number of names, however, are written a bit different such as "اسحق" in Arabic style and "اسحاق" in Dari/Persian style, "اسمعیل" in Arabic style and "اسماعیل" in Dari/Persian style and "عبدالرحمن" in Arabic style and sometimes "عبدالرحمان" in Dari/Persian style – these cases are very rare. The current method is not able to match such names unless they are preprocessed and transformed prior to the matching process.

Name	Length	Variations	Length	Matched?	Comment
فاطمه	5	فاطمه	6	FALSE	One unnecessary trailing space
فاطمه	5	فا طمه	6	FALSE	One unnecessary space in middle
فاطمه	5	فا طمه	7	FALSE	Two unnecessary spaces in middle
فاطمه	5	فاطمه	7	FALSE	Two unnecessary leading spaces
ذبيح الله	9	ذبيح الله	11	FALSE	One unnecessary leading space
ذبيح الله	9	ذ بيح الله	10	FALSE	One unnecessary space after 1st char
ذبيح الله	9	ذبيح الله	10	FALSE	Two unnecessary spaces in middle
ذبيح الله	9	ذبيحالله	8	FALSE	Two words are connected
ياسين	5	ياسين	5	FALSE	Different Unicode for letter Yeh
ياسين	5	ياسين	5	FALSE	Different Unicode for letter Yeh

All Spaces were Removed

Name	Length	Variations	Length	Matched?	Comment
فاطمه	5	فاطمه	5	TRUE	All spaces were stripped
فاطمه	5	فاطمه	5	TRUE	All spaces were stripped
فاطمه	5	فاطمه	5	TRUE	All spaces were stripped
فاطمه	5	فاطمه	5	TRUE	All spaces were stripped
ذبيحالله	8	ذبيحالله	8	TRUE	All spaces were stripped
ذبيحالله	8	ذبيحالله	8	TRUE	All spaces were stripped
ذبيحالله	8	ذبيحالله	8	TRUE	All spaces were stripped
ذبيحالله	8	ذبيحالله	8	TRUE	All spaces were stripped
ياسين	5	ياسين	5	TRUE	Unicode for letter Yeh were unified
ياسين	5	ياسين	5	TRUE	Unicode for letter Yeh were unified

Figure 4-13. Unicode were unified, spaces were removed from Names, and then Exact Sting Matching has been used for comparison.

4.1.6.4. Approximate String Matching

Approximate string matching, conversationally referred to as 'fuzzy string searching', is a technique of finding strings that match a pattern approximately (rather than exactly). The following approaches are proposed by the author of this thesis:

Fuzzy Lookup in Power Query

The author of this thesis wrote a customized method in Power Query, named *fxBasicFuzzyLookup*, as the code shown below.

```
/* The 'fxBasicFuzzyLookup' receives two string parameters,
 * and calculates the similarity of the two given strings.
 * It then returns a score determining how the strings are similar.
 * @strWord1 the first string
 * @strWord2 the second string
*/
```

```
(strWord1 as text, strWord2 as text) as number =>
let
    delimiter = " ",
    // rectifies Unicode for strWord1 and strWord2
    fixWord1UnicodeYeh = Text.Replace(strWord1, "ي", "ى"),
    fixWord2UnicodeYeh = Text.Replace(strWord2, "ي", "ى"),
    fixWord1UnicodeKaf = Text.Replace(fixWord1UnicodeYeh, "ك ", "ک"),
    fixWord2UnicodeKaf = Text.Replace(fixWord2UnicodeYeh, "ك ", "ک"),
    // strips all spaces from strWord1 and strWord2
    stripWord1Spaces = Text.Remove(fixWord1UnicodeKaf, delimiter),
    stripWord2Spaces = Text.Remove(fixWord2UnicodeKaf, delimiter),
    word1 = stripWord1Spaces,
    word2 = stripWord2Spaces,

    //finds intersection of word1 and word1
    intersection = List.Intersect({Text.ToList(word1), Text.ToList(word2)}),

    // calculates total characters of both word1 and word2
    totalChars = Text.Length(word1) + Text.Length(word2),

    // calculates and returns similiarity score
    return = Number.Round(2 * List.Count(intersection) / (totalChars), 2)
in
    return
```

Table 4-4. Tailored fxBasicFuzzyLookup method calculates and returns similarity score between two string values.

The *fxBasicFuzzyLookup* accepts two string parameters from a user. Initially it unifies the Unicode characters for both string parameters, and it subsequently removes all the necessary and unnecessary spaces from both string parameters. The method then calculates similarity for the two given string parameters and it then returns a similarity score in range from 0 to 1 to the user. The similarity score provides a measure of how confident the method is in its matching of inexact data and can be used as threshold to exclude instances less than or equal to the threshold. The algorithm used to design this method is inspired by Jaccard similarity algorithm.

The following figure demonstrates the output of *fxBasicFuzzyLookup* represented in **Table 4-4** approximately matched Word1 with Word2 and returned a score showing their similarity.

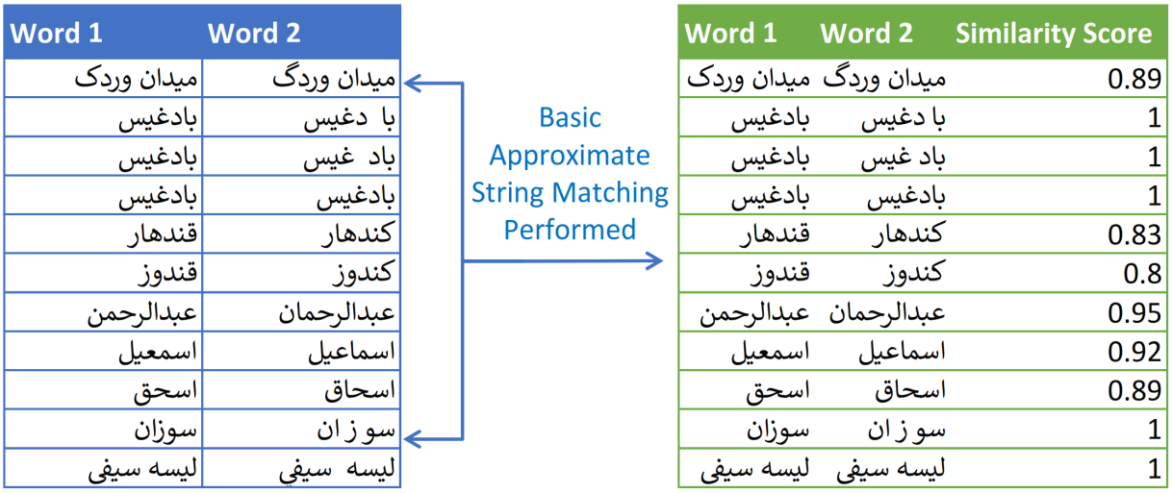

Figure 4-14. Output of the fxBasicFuzzyLookup comparison.

This method is not a great solution for the following two reasons:

- It is very time and resource consuming.
- It returns a perfect similarity score for two names with equal number of similar characters, even if the order of the characters differs; for example, in 'رامز' and 'مزار'.
- There are names with equal number of characters and all the characters but one are equal and their order is also equal for example in جمال الدین (Jamaluddin) and کمال الدین (Kamaluddin). Even though these names are completely different names the fxBasicFuzzyLookup method often produces very high similarity score which is not correct, as pictured in the following figure.

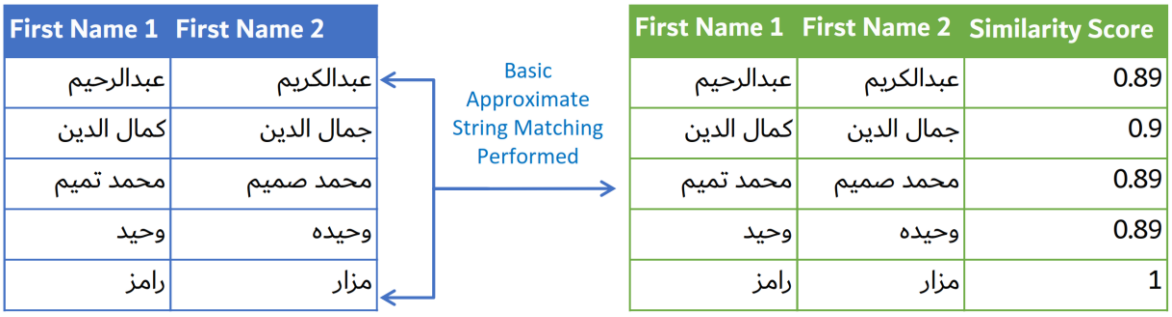

Figure 4-15. The fxBasicFuzzyLookup often produces high similarity score for completely different names.

It is worth mentioning that concatenation of all the common attributes that are needed for comparison such as First Names, Father's Names, High School Names and others greatly improve accuracy of *fxBasicFuzzyLookup* method. The identified *second and third problems* associated with the *fxBasicFuzzyLookup* method can be addressed almost fully; since there is no chance or the chance is really very small that two candidates have the same First Name and Father's Name and graduate from same high school in the same year. However, the *first issue* remains a problem.

Fuzzy Lookup in Microsoft Excel

In addition to the Power Query approach, there is an Add-In named Fuzzy Lookup, for Microsoft Excel, developed by Microsoft Research. This Add-In performs fuzzy matching of textual data in Microsoft Excel.

The project description claims (Microsoft Research 2014), “It can be used to identify fuzzy duplicate rows within a single table or to fuzzy join similar rows between two different tables. The matching is robust to a wide variety of errors including spelling mistakes, abbreviations, synonyms and added/missing data. For instance, it might detect that the rows ‘Mr. Andrew Hill’, ‘Hill, Andrew R.’ and ‘Andy Hill’ all refer to the same underlying entity, returning a similarity score along with each match”.

Based on the (Microsoft Research 2014), the Fuzzy Lookup Add-in allows one to join data from multiple tables into one. Therefore, prior to fuzzy lookup, the data ranges in Excel must be converted to Excel tables.

The following figure presents a sample of High School Data and Kankor Data to be used for the basis of fuzzy lookup. High School Dataset is the name of the table on the top and Kankor Dataset is the name of the table on the bottom.

First Dataset: High School Data

First Name	Father Name	High School Name	High School Location	High School Graduation	GPA Grade 10	GPA Grade 11	GPA Grade 12
فرید احمد	عبدالعزیز	لیسه سیفی	هرات	2013	85	90	95
فرید احمد	عبدالنا صر	لیسهانقلاب	هرات	2015	60	65	70
منیژه	عبد ال کر یم	لیسه مریم	کابل	2015	80	85	90
سو ز ا ن	عبدالرحیم	لیسه مر یم	کابل	2014	65	70	60
عبد ا لر حمن	جمال ا لدین	لیسهسیفی	هرات	2014	90	95	85
عبدالسلام	کمال الدین	لیسه سیفی	هرات	2015	75	80	90

Second Dataset: Kankor Data

First Name	Father Name	High School Name	High School Location	High School Graduation	Kankor Score	Kankor Result
عبد الوحید	سلطانی	لیسه سیفی	هرات	2015	260	Science
فر ید احمد	عبد الناصر	لیسه انقلاب	هرات	2015	210	Resultless
عبدالرحمن	جمال الد ین	لیسه سیفی	هرات	2014	305	Medicine
سوزان	عبد ا لر حیم	لیسه مریم	کابل	2014	205	Resultless
سلیمان	سلیمانی	لیسه انقلاب	هرات	2015	290	Law & Politics
شقایق	شقایقی	لیسه مریم	کابل	2015	285	Law & Politics
منیژ ه	عبدالکریم	لیسهم ر یم	کابل	2015	295	Stomatology
فرید احمد	محمد حسین	لیسه سیفی	هرات	2015	285	Engineering

Figure 4-16. Sample of High School Data (top) and Kankor Data (bottom).

There is no primary and foreign key to join these two datasets together, as explained earlier in this section. Therefore, the only solution is to match the ‘First Name’, ‘Father’s Name’, ‘High School Name’, ‘High School Location’ and ‘High School Graduation’ attributes in High School Dataset respectively to the ‘First Name’, ‘Father’s Name’, ‘High School Name’, ‘High School Location’ and ‘High School Graduation’ attributes in Kankor Dataset and create a results list that shows ‘GPA Grade 10’, ‘GPA Grade 11’ and ‘GPA Grade 12’ from High School Dataset together with all attributes from Kankor Dataset.

The challenge is that not all of the values for the common attributes in both datasets are spelled and arranged the same way; therefore, the previous attempt with traditional methods such as exact string matching to complete this task was not very successful. For example, consider the First Name ‘سو ز ا ن’ on row 4 of High School Dataset with a length of 8 characters due to the unnecessary single spaces in the name. In looking at row 8 of Kankor Dataset, it exposes that what should be the matching name is arranged as “سوزان” with a

length of 5 characters; thus, an exact match would not match the two names together. Likewise, the High School Location ‘لیسهانقلاب’ on row 2 of High School Dataset where the words are connected is not exact matches to ‘لیسه انقلاب’ on row 2 of Kankor Dataset which the words are not connected.

The author of this thesis matched the data in these two datasets using the Fuzzy Lookup Add-in in Microsoft Excel. The common attributes mentioned earlier were selected for matching between the two datasets. The required attributes as well as the system-generated Similarity score were selected to be included in the results list. With these settings, the Fuzzy Lookup Add-in analyzed the data in the two tables and provided the results list pictured in the following figure.

First Name	Father Name	High School Name	High School Location	High School Graduation	Kankor Score	Kankor Result	GPA Grade 10	GPA Grade 11	GPA Grade 12	Similarity
عبد الوحید	سلطانی	لیسه سیفی	هرات	2015	260	Science				0.00
فر ید احمد	عبد الناصر	لیسه انقلاب	هرات	2015	210	Resultless				0.00
عبدالرحمن	جمال الد ین	لیسه سیفی	هرات	2014	305	Medicine	90	95	85	0.80
سوزان	عبد ا لر حیم	لیسه مریم	کابل	2014	205	Resultless	65	70	60	0.98
سلیمان	سلیمانی	لیسه انقلاب	هرات	2015	290	Law & Politics				0.00
شقایق	شقایقی	لیسه مریم	کابل	2015	285	Law & Politics				0.00
منیژ ه	عبدالکریم	لیسهم ر یم	کابل	2015	295	Stomatology	80	85	90	0.81
فرید احمد	محمد حسین	لیسه سیفی	هرات	2015	285	Engineering				0.00

Figure 4-17. Results list produced by Microsoft Excel Fuzzy Lookup Add-In.

A close inspection of the data in above figure reveals that the Fuzzy Lookup tool worked well, though not flawlessly. Particularly, it did not find matches for candidate with name ‘فرید احمد’ on row 2 of High School Dataset and ‘فر ید احمد’ on row 2 of Kankor Dataset. There are extraneous spaces within the candidate’s First Name and Father’s Name as well as the name of the High School in both datasets were written in different formats.

In order to improve the accuracy of this approach, the author of this thesis concatenated the common attributes and then removed all necessary and unnecessary spaces from both datasets using the following Microsoft Excel’s built-in functions TEXTJOIN() and SUBSTITUTE(), as shown below.

```
=SUBSTITUTE(
    SUBSTITUTE(
        SUBSTITUTE(
            TEXTJOIN(
                  "", TRUE,
                  HighSchoolDataset[@[First  Name]:[High  School  Graduation]]
            ), " ", ""
        ), "ي", "ى"
    ), "ك", "ک"
)
```

Table 4-5. Common attributes were concatenated and then all spaces were removed.

Then the fuzzy lookup with same settings as previous except this time the calculated column (concatenation of all common attributes) was chosen for matching. Upon doing so, as shown in the following, the Fuzzy Lookup tool matched all of the data in the two tables with better and even perfect similarity score and does so without any errors.

First Name	Father Name	High School Name	High School Location	High School Graduation	Kankor Score	Kankor Result	GPA Grade 10	GPA Grade 11	GPA Grade 12	Similarity
عبد الوحید	سلطانی	لیسه سیفی	هرات	2015	260	Science				0.00
فرید احمد	عبد الناصر	لیسه انقلاب	هرات	2015	210	Resultless	60	65	70	1.00
عبدالرحمن	جمال الد ین	لیسه سیفی	هرات	2014	305	Medicine	90	95	85	1.00
سوزان	عبد ا لر حیم	لیسه مریم	کابل	2014	205	Resultless	65	70	60	1.00
سلیمان	سلیمانی	لیسه انقلاب	هرات	2015	290	Law & Politics				0.00
شقایق	شقایقی	لیسه مریم	کابل	2015	285	Law & Politics				0.00
منیژ ه	عبدالکریم	لیسهم ر یم	کابل	2015	295	Stomatology	80	85	90	1.00
فرید احمد	محمد حسین	لیسه سیفی	هرات	2015	285	Engineering				0.00

Figure 4-18. Concatenation and removing spaces from common attributes improved accuracy of Fuzzy Lookup.

Microsoft Excel's Fuzzy Lookup is robust. It also considers the order of the similar characters in addition to the total number of characters and the number of similar characters in given strings. Therefore, it can be concluded that the accuracy of Microsoft Excel's Fuzzy Lookup is better than Power's Query. It is worth mentioning that concatenation, removal of all blank spaces and rectification of Unicode of common attributes prior to matching increase both the performance and the accuracy of the method.

In the end, in this case and for other similar cases the author of this thesis recommends either to use Fuzzy Lookup technique with the mentioned rectified settings or to use Strip All Spaces technique with the exact string match.

4.1.7. Autofill Missing Gender Values Using Identical Names

Out of 1.5 million records of Kankor data, for the 234,266 records the gender label is missing. This situation poses a challenge for data quality and completeness and should be addressed systematically through a structured and efficient process. There are many Kankor applicants with the same First Names e.g. more than 11,000 male candidates were named 'Zabiuallah' and more than 9,000 female candidates were named 'Fatemeh' attended Kankor. It is assumed that the gender for at least one of the identical names is identified. That gender value is used as a reference to autofill the gender for the identical names whose gender value is missing. For further clarification purposes, this scenario is illustrated in the following diagram (see **Figure 4-19**).

Name	Gender
Tofaan	Female
Tofaan	Male
Tofaan	Male
Tofaan	
Fatemeh	
Fatemeh	Female
Fatemeh	
Zabiullah	
Abdul Rahman	Male
Abdul Rahman	
Maryam	Female
Zabiullah	
Maryam	
Zabiullah	Male
Maryam	Female
Tofaan	

Missing Gender Values were Autofilled

Name	Gender
Tofaan	Female
Tofaan	Female
Tofaan	Female
Tofaan	Female
Tofaan	Female
Fatemeh	Female
Fatemeh	Female
Fatemeh	Female
Zabiullah	Male
Zabiullah	Male
Zabiullah	Male
Abdul Rahman	Male
Abdul Rahman	Male
Maryam	Female
Maryam	Female
Maryam	Female

Figure 4-19. Incorrect autofill of missing gender values from existing data using identical names.

The following Power Query code is used to autofill missing values for the gender column using identical names of Kankor candidates:

```
let
fxPreprocess = (txtInput as text) as text =>
    let
        delimiter = " ",
        // Rectifies Unicode for txtInput
        fixUnicode = Text.Replace(txtInput, "ي", "ى"),
        // Strips all spaces from txtInput
        stripSpaces = Text.Remove(fixUnicode, delimiter)
    in
stripSpaces,
    // Import the messy dataset into Power Query
    Source = Excel.CurrentWorkbook(){[Name="MessyDataset"]}[Content],
    /* Apply the 'fxPreprocess' method on the Name attribute.
     * This method initially fixes Unicode character issue,
     * then it removes all necessary and unnecessary space.
    */
    #"Fixed Unicode & Removed Spaces" = Table.TransformColumns(Source,{{"Name", each
fxPreprocess(_) }}),
    /* Group all the records by Name attribute;
     * then find the minimum value of Gender over each group.
     * Also, select All Rows for grouping and name it 'All Data'
     * in order to expand this new grouped attribute later.
    */
    #"Grouped Records" = Table.Group(#"Fixed Unicode & Removed Spaces", {"Name"},
{{"Gender", each List.Min([Gender]), type text}, {"All Data", each _, type table}}),
    /* Expand the new grouped attribute 'All Data'.
     * This picks the first Gender label for each Group,
     * and automatically fills for all other Names in that Group.
    */
    #"Autofilled Missing Gender" = Table.ExpandTableColumn(#"Grouped Records", "All
Data", {"Gender"}, {"Gender.1"}),
    // Remove the origianl Gender attribute
    #"Removed Extra Column" = Table.RemoveColumns(#"Autofilled Missing
Gender",{"Gender.1"})
in
    #"Removed Extra Column"
```

Table 4-6. This power query code auto fills missing gender values using identical names.

It is worth mentioning that in some rare cases there are First Names used both for men and women, something which is the case with many English names. In particular, in some instances a man's first name, such as, *Tofaan*, might has been used for a woman. In the Kankor data consider the following scenario: A female Kankor candidate, named *Tofaan*,

has a female label in the gender column. There are records of male candidates with the same names, but the labels for their gender column are missing. Because the names are identical, the approach described above may take the gender value of the female candidate and autofill for the missing values for the male candidates, which is false.

A close inspection of the data in **Figure 4-19** exposes that the method worked well, though not perfectly. There four candidates named *Tofaan*. Two of them were labeled male, one of them labeled female and the label for the other one is missing. The method picked gender label from the very first candidate named *Tofaan* with female label and auto filled all other candidates in the same group with female gender label. This situation raises a challenge for the labeling of the missing values for the gender column, as illustrated in the above diagram. To address this predicament, the names are grouped and sorted such that identical names with the maximum number of occurrences of male/female values appear at the top. This power query code is entirely similar with the previous one. Except this time MIN aggregation has been replaced with MAX aggregation while grouping all the candidates by their Name. This approach guarantees that the pattern will autofill the missing values with labels from the top, shown in **Figure 4-20**.

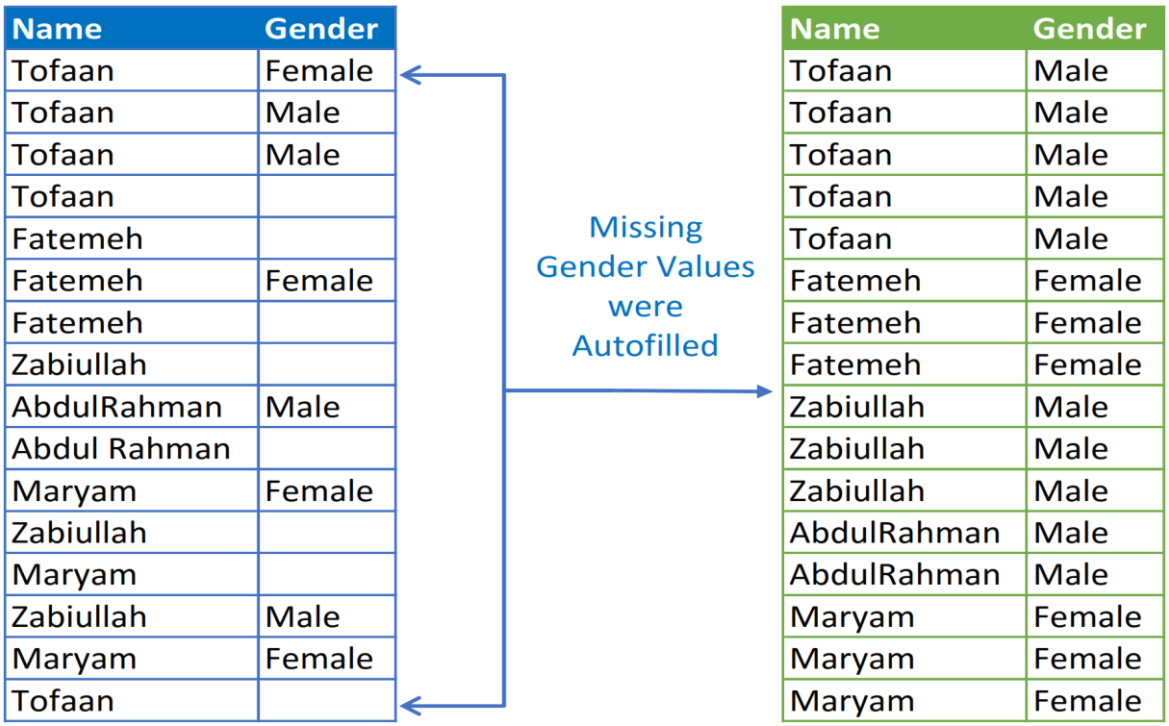

Figure 4-20. Autofill missing gender values based on the maximum number of occurrences.

However, inspection of the above data in **Figure 4-20** shows that the above approach overwrites the gender values already entered for the gender column by the administrators. For example, consider the four candidates named *Tofaan,* the very first one is already labeled female by the administrators. Since the method picks the maximum number of occurrences of existing values for each group, therefore, it picked male label and auto filled that for all the candidates with *Tofaan* name, even the ones already labeled by the administrators.

To address this predicament, another approach is suggested to guarantee that the algorithm will autofill the missing values properly. This approach divides the dataset into two datasets.

One dataset contains source data, named *first dataset*. The other dataset contains identical names of those with the maximum number of occurrences of gender values, named *second dataset*. Then these two datasets are merged for comparison and based on the maximum number of occurrences of existing values in the *second dataset* and the missing gender values in the *first dataset* are auto filled. This approach ensures the values already entered by the administrators are not overwritten, as illustrated in the following figure.

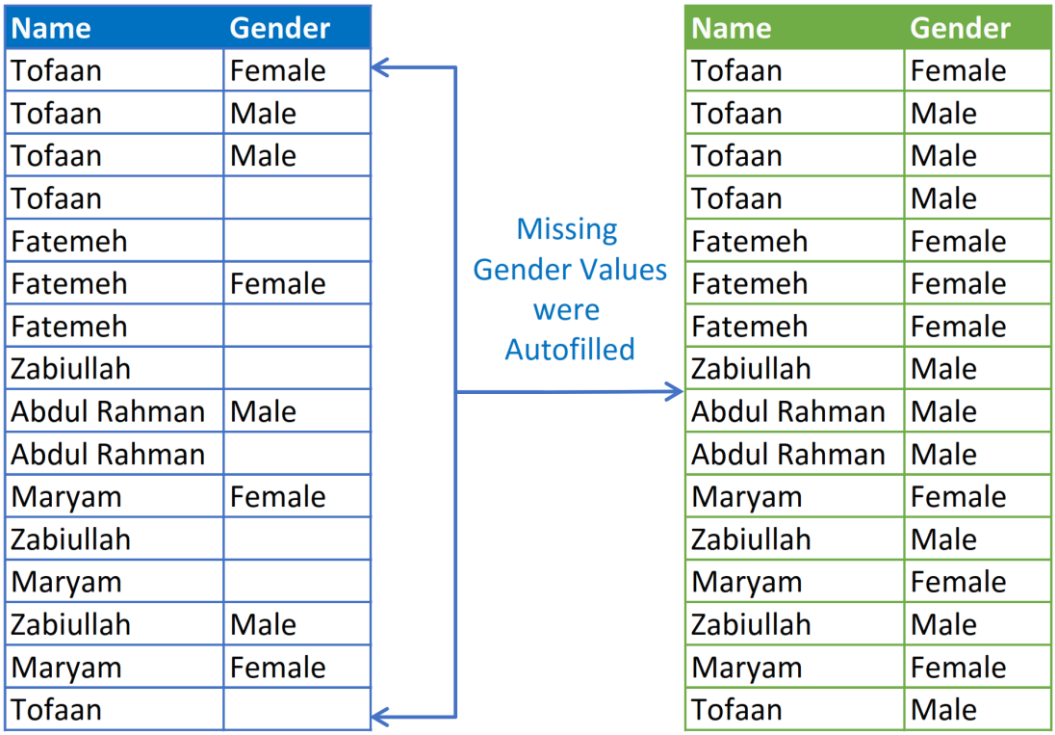

Figure 4-21. Autofill missing gender values from existing data using a separate query and merge feature.

It, however, must be considered that when performing the above approach, leading and trailing spaces must undergo the trimming operation in order to not only unify data cleansing but also to address data quality gaps hierarchically. Furthermore, as detailed in the order of operations, certain alphabetical letters are represented by more than one Unicode character, even though in the human eye, they may appear similar or even identical. The following Power Query code is used to form the *second dataset* including the unique names with maximum number of occurrences of gender values. It initially groups all the instances by Name and Gender attributes, then removes instances with missing gender values, and finally extracts from the identical names those with the maximum number of occurrences of gender values.

```
let
    Source = MessyDataset,
    #"Grouped By Name and Gender" = Table.Group(Source, {"Name", "Gender"}, {{"Count",
each Table.RowCount(_), type number}}),
    #"Removed Records With Missing Gender" = Table.SelectRows(#"Grouped By Name and
Gender", each ([Gender] <> null)),
    #"Grouped By Max Occurrences" = Table.Group(#"Removed Records With Missing
Gender", {"Name"}, {{"AllData", each Table.Max(_, each [Count]), type record}}),
```

```
    #"Expanded Max Occurrences Group" = Table.ExpandRecordColumn(#"Grouped By Max
Occurrences", "AllData", {"Gender"}, {"Gender"})
in
    #"Expanded Max Occurrences Group"
```

Table 4-7. This power query code creates the second dataset contains unique names with maximum number of occurrences of gender values.

The following Power Query code is used to autofill the missing gender values properly. It initially merges the *original source dataset* with *second dataset* that created from the above Power Query. Then a new calculated column with the following condition added to the *original source dataset*: If the value of 'Gender' attribute in the *original source dataset* is null, then get the value from the Gender attribute of the *second dataset*. Otherwise, do not overwrite the gender value in the *original source dataset*.

```
let
    // 'first dataset' with missing gender is merged with 'second dataset'
    Source = Table.NestedJoin(MessyDataset,{"Name"},#"second dataset",{"Name"},"Second
Dataset Columns",JoinKind.LeftOuter),


    // Add 'Gender' column from the 'second dataset' into 'first dataset'
    #"Expanded Gender From Second Dataset" = Table.ExpandTableColumn(Source, "Second
Dataset Columns", {"Gender"}, {"Second Dataset.Gender"}),


    /* A conditional column added, named 'Autofilled Gender'.
     * If the 'Gender' attribute in 'original source dataset' is null,
      * then get the value from the 'Gender' attribute of the 'second dataset'.
     * Otherwise, do not overwrite the value.
    */
    #"Add Conditional Column" = Table.AddColumn(#"Expanded Gender From Second
Dataset", "Autofilled Gender", each if [Gender] = null then [Second Dataset.Gender]
else [Gender]),


    // Remove 'Gender' attribute from both 'original source dataset' and 'second
dataset'
    #"Removed Other Columns" = Table.RemoveColumns(#"Add Conditional
Column",{"Gender", "Second Dataset.Gender"}),


    // Rename 'Autofilled Gender' to 'Gender'
    #"Renamed Conditional Column" = Table.RenameColumns(#"Removed Other
Columns",{{"Autofilled Gender", "Gender"}})
in
    #"Renamed Conditional Column"
```

Table 4-8. This Power Query code merges the source dataset with second dataset and auto fills the missing values and ensures the existing values are not overwritten.

A question about the Power Query solution to autofill the missing values was asked by the author of this thesis on the Power BI section of MrExcel Forum, and the solution was suggested by a member of the Forum, named MarcelBeug. The idea then was revised and

expanded by the author of this thesis to auto fill the missing gender values correctly. The very last technique works very well and ensures the existing values recorded by the administrators are not overwritten. The author of this thesis took this Power Query technique and wrote it using Structured Query Language (SQL) and made the solution available for MySQL users. The SQL statement is very complex, therefore, it is broken down into multiple separate SQL statements for a better understanding, as shown below.

```
-- Step 1: Group by name and gender, and count the participants
CREATE view step1 AS
  SELECT name, gender, Count(name) AS CountRows
  FROM   users GROUP  BY name, gender;

-- Step 2: Exclude records missing gender value
CREATE view step2 AS
  SELECT * FROM   step1 WHERE gender IS NOT NULL;

-- Step 3: Exclude similar names with less occurrences
CREATE view step3 AS
  SELECT step2.name,
         Min(step2.gender) AS Gender,
         step2.countrows
  FROM (SELECT name,
                 Max(countrows) AS MaxOccurances
          FROM   step2
          GROUP BY name) step22
         INNER JOIN step2
                 ON step2.name = step22.name
                    AND step2.countrows = step22.maxoccurances
  GROUP BY step2.name,
            step2.countrows;

-- Autofill Gender by merging the source dataset with above step3 result-sets
CREATE view finalstep AS
  SELECT users.id,
         users.name,
         CASE
           WHEN users.gender IS NULL THEN step3.gender
           ELSE users.gender
         end AS Gender
  FROM   users
         INNER JOIN step3
                 ON users.name = step3.name;
```

Table 4-9. This SQL auto fills missing gender values and ensures the existing values are not overwritten.

It is worth mentioning that sometimes high school names include information that shows the high school is only for males or females. Therefore, still there is room to improve the accuracy of this technique to tell the algorithm to consider high school names as part of its evaluation while auto filling the missing gender values.

Additionally, in cases that a name is used both for male and female candidates, another alternative is to calculate frequency distribution of male and female for that name, and then tell the algorithm to auto fill the missing values for the candidates with that name based on their frequency distribution.

4.1.8. Autofill Missing High School Geographical Location Values

Out of 1.5 million records of Kankor data since 2003, for 660,547 records the location of high schools is missing. This is a big issue and a dilemma which require to be addressed systematically prior to data analysis activities.

The technique used to autofill the missing gender values from existing data can be used to autofill the missing location of high schools, as shown in the following figure for further clarification.

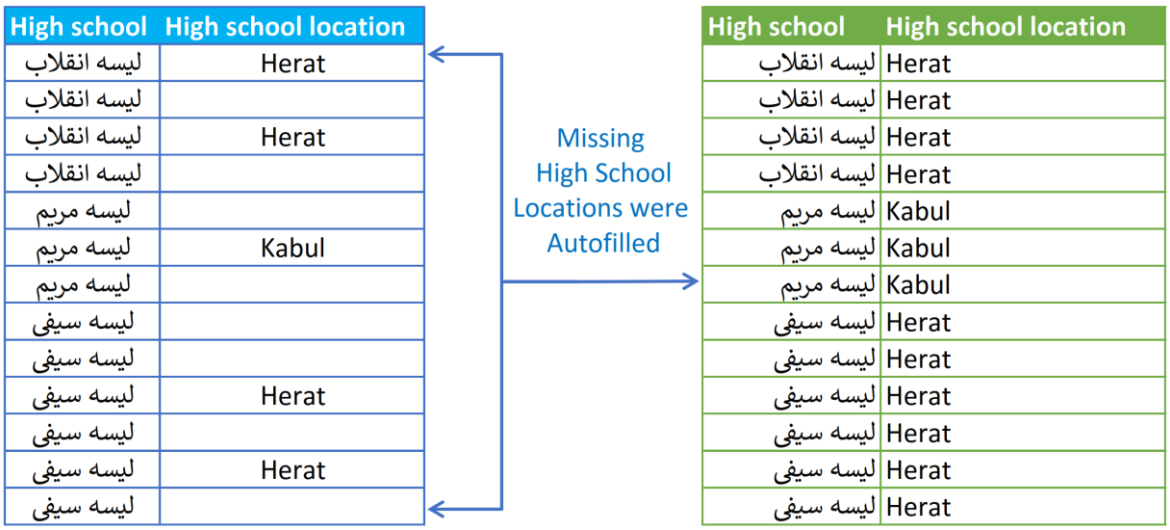

Figure 4-22. Autofill missing high school location based on maximum number of occurrences of existing values.

It is worth mentioning that in rare cases high schools share the same name across the country. Hence, there is a chance for two high schools to be located in different locations (say, one is located in Herat province and the other in Kabul province) but have the same name. In such a situation, the above-mentioned approach does not work and there is the possibility to misallocate values – In such a situation, eyeballing by the administrators cannot allocate the missing values properly.

4.1.9. Combine Data from Multiple Files and Sources

A common task that most analysts have faced at one time or another is to combine data from multiple files or multiple sources into one big uniform master table. Without a solid knowledge of Microsoft Excel's advanced functions and Macro and VBA programming, this

task typically entails opening each file, copying the data, and then pasting the data into a single workbook, which is very complex and time-consuming.

Apart from using Microsoft Excel and VBA programming, there are multiple other methods using which one can consolidate data from various files or sources. However, the learning curve is steep and out of reach for users without a solid knowledge. The Power Query feature is a strong data analytical tool that will allow one to connect with numerous data sources or multiple files from a single folder and consolidate dispersed data into a single master table effortlessly. The data sources can be all in one format such as Microsoft Excel, Microsoft Access, even in CSV formats, or can be stored in multiple sources and formats.

An example of this is combining Kankor datasets from multiple years, where the schema and structure of the datasets unified by the author of this thesis, but the data differ for each year. Presently, there are 13 separate files for Kankor results data for the years 2003-2015, one file per year. It is deemed required to combine these files to produce rich and detailed reports and other analyses. To start the process of combining these multiple files, first it must be ensured that all the files are in a single folder (aka, directory) since Power Query has the capability to read files from a given folder, as illustrated in the following figure.

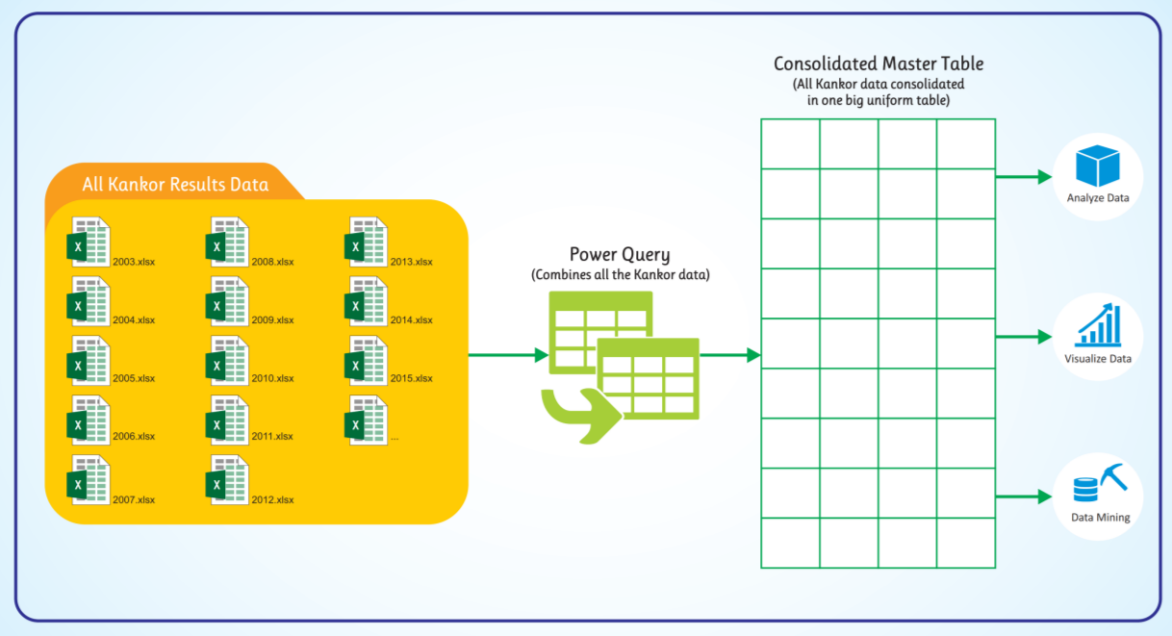

Figure 4-23. Combine all Kankor data for the years 2003-2015.

To ensure compatible loading and combining is performed consistently across all files, the following steps were considered:

- All Microsoft Excel workbooks were moved into a folder, named “All Kankor Results Data”;
- Each of the Microsoft Excel workbooks only has one Worksheet that carries Kankor data for one year;
- The Kankor data in each Worksheet converted into Excel Table;
- Then Power Query is trained to read the content of “All Kankor Results Data” folder and to consider only one of the Workbook as the template (it is called ‘key’ in Power Query) for transformation, as shown below;

```
let
    Source = (#"Sample File Parameter" as binary) => let
    /*
     * One of the Excel file is chosen as the template,
     * Power Query calls it, 'key'.
     * Power Query then use it to combine all the other Excel files with it.
    */
    Source = Excel.Workbook(#"Sample File Parameter", null, true),

    /*
     * Power Query reads 'Ranges', 'Tables' and 'Name Ranges' from Excel.
     * Therefore, all rows with 'Kind' = 'Table' were kept, others excluded.
    */
    #"Keep Excel Tables" = Table.SelectRows(Source, each [Kind] = "Table"),

    /*
     * Presently, each Excel contains only one sheet,
     * and each sheet contains only one dataset as 'Excel Table'.
     * Just for safety precautions keep only one 'Table'
    */
    #"Kept First Row" = Table.FirstN(#"Keep Excel Tables", 1),

    // Keep the 'Data' attribute and remove others
    #"Remove Metadata" = Table.SelectColumns(#"Kept First Row",{"Data"}),

    // Expand the content the 'Data' attribute
    #"Expanded Data" = Table.ExpandTableColumn(#"Remove Metadata", "Data",
{"KankorCode", "FirstName", "FamilyName", "Gender", "School",
"GraduationYear", "Province", "Score", "Field of Study", "Institution"},
{"KankorCode", "FirstName", "FamilyName", "Gender", "School",
"GraduationYear", "Province", "Score", "Field of Study", "Institution"})
in
    #"Expanded Data"
in
    Source
```

Table 4-10. The fxTransformTemplate function receives a Microsoft Excel file and preprocesses it for consolidation.

- Finally, the Power Query is set to automatically parse and transform the template transformation steps from above and combine the data from each Workbook and produce a comprehensive output in one master table which is ready for further transformation or analysis activities, as shown below.

```
let
    // Read content of the 'All Kankor Results Data' folder
    Source = Folder.Files("path\All Kankor Results Data"),

    // Covert 'Extension' metadata to lowercase
    #"Lowercase Extension" = Table.TransformColumns(Source,{{"Extension",
Text.Lower, type text}}),

    // Only keep Excel files and exclude other contents in the folder
    #"Keep Excel Files" = Table.SelectRows(#"Lowercase Extension", each
[Extension] = ".xlsx" or [Extension] = ".xls"),

    // Call the method trained to parse and transform the template Excel file
    #"Invoke fxTransformTemplate" = Table.AddColumn(#"Keep Excel Files",
"Transform File from All Kankor Results Data", each
fxTransformTemplate([Content])),

    // Keep only 'File Name' and 'Data' attribute and exclude other
attributes
    #"Remove Metadata" = Table.SelectColumns(#"Invoke
fxTransformTemplate",{"Name", "Transform File from All Kankor Results
Data"}),

    // Expand all the content of the 'Data' attribute
    #"Expanded All Data" = Table.ExpandTableColumn(#"Remove Metadata",
"Transform File from All Kankor Results Data",
Table.ColumnNames(fxTransformTemplate(#"Sample File"))),

    // Change 'attributes' data type properly
    #"Changed Type" = Table.TransformColumnTypes(#"Expanded All
Data",{{"Name", type text}, {"KankorCode", type text}, {"FirstName", type
text}, {"FamilyName", type text}, {"Gender", type text}, {"School", type
text}, {"Province", type text}, {"Field of Study", type text},
{"Institution", type text}, {"GraduationYear", Int64.Type}, {"Score", type
number}})
in
    #"Changed Type"
```

Table 4-11. This Power Query code imports each Microsoft Excel file and passes it as a parameter to the fxTransformTemplate to be preprocessed and finally consolidates all of them into a uniform master table.

Power Query records all the transformation and cleansing steps. Once the steps are setup, the imported data can always be refreshed, and the recorded steps will be applied when the original files change.

Furthermore, when the Kankor results data for the *year x* is announced and the data is ready for preprocessing, there is no need to perform all the transformation and preprocessing steps

manually on the data for the *year x*. Since all the applied steps of transformation are recorded, the data for the *year* x is automatically cleansed when it is moved to the "All Kankor Results Data" folder.

4.1.10. Transformation of Persian Labels to their English Equivalent Terms

The Kankor data are recorded in Dari/Persian language, for clarification purpose a sample of the Kankor data is illustrated in the following figure.

KankorCode	First Name	Last Name	Gender	Province	Kankor Score	Result
F00000001	عبدالوحید	وحیدی	مذکر	هرات	290	کامپیوتر ساینس
F00000002	سوزان	رحیمی	مونث	کابل	195	بی نتیجه
F00000003	سلیمان	کریمی	مذکر	هرات	205	بی نتیجه
F00000004	مریم	امیری	مونث	کابل	285	اقتصاد
F00000005	منیژه	سادات	مونث	بلخ	210	بی نتیجه
F00000006	نرگس	رضایی	مونث	جلال آباد	190	بی نتیجه
F00000007	فرید	عزیزی	مذکر	هرات	310	طب
F00000008	فرهاد	فرهادی	مذکر	بلخ	200	بی نتیجه

Figure 4-24. Sample of Kankor data recorded in Dari/Persian.

The values for the gender, province, fields of study, educational institutions and other columns with limited possible values/labels are significant for reports, visualization and other analysis activities. To make the output globally readable and understandable it is deemed efficient to convert these values to their English equivalent terms, as demonstrated in the following diagram for further clarification.

KankorCode	First Name	Last Name	Gender	Province	Province in English	Kankor Score	Result	Result in English
F00000001	عبدالوحید	وحیدی	Male	هرات	Herat	290	کامپیوتر ساینس	Computer Science
F00000003	سلیمان	کریمی	Male	هرات	Herat	205	بی نتیجه	Resultless
F00000002	سوزان	رحیمی	Female	کابل	Kabul	195	بی نتیجه	Resultless
F00000005	منیژه	سادات	Female	بلخ	Balkh	210	بی نتیجه	Resultless
F00000004	مریم	امیری	Female	کابل	Kabul	285	اقتصاد	Economics
F00000006	نرگس	رضایی	Female	جلال آباد	Jalalabad	190	بی نتیجه	Resultless
F00000007	فرید	عزیزی	Male	هرات	Herat	310	طب	Medicine
F00000008	فرهاد	فرهادی	Male	بلخ	Balkh	200	بی نتیجه	Resultless

Figure 4-25. Transformation of Dari/Persian values to their equivalent English terms.

For some columns with a few limited values such as gender, the *replace* transformation is used to do the conversion. However, for some other columns with numerous limited values, lookup tables are used. For example, values for the province column explained above were reduced to 34 distinct values, or fields of study to more than 100 unique values, after the issues of different common names, typos, and formats were addressed. It is not an efficient approach to apply the *replace* transformation 34 times or more than 100 times. It is deemed efficient to use lookup tables including the equivalent English terms for the province and

other columns, as presented in the following figure. Then audit and merge the original source dataset with the lookup tables using proper join types.

Province Lookup Table	
Province	**Province in English**
هرات	Herat
کابل	Kabul
بلخ	Balkh
جلال آباد	Jalalabad

Result Lookup Table	
Result	**Result in English**
کامپیوتر ساینس	Computer Science
اقتصاد	Economics
طب	Medicine
بی نتیجه	Resultless

Figure 4-26. Sample of lookup table for Province and Result attributes.

The following Power Query code illustrates the process of conversion of Dari values to their equivalent English terms using the *replace* as well as the *merge* transformations.

```
let
    // Load 'source data' dataset
    Source = MasterTable,

    // Replace 'male' value from Dari to English
    #"Replaced Male Gender" =
Table.ReplaceValue(Source,"مذکر","Male",Replacer.ReplaceValue,{"Gender"}),

    // Replace 'female' value from Dari to English
    #"Replaced Female Gender" = Table.ReplaceValue(#"Replaced Male
Gender","مونث","Female",Replacer.ReplaceValue,{"Gender"}),

    // Merge the the 'Province Lookup' table with 'source data'
    #"Merged with Province Lookup" = Table.NestedJoin(#"Replaced Female
Gender",{"Province"},ProvinceLookup,{"Province"},"NewColumn",JoinKind.LeftOuter),

    // Expand 'Province in English' attribute from 'Province Lookup' table into
'source data'
    #"Province Equivalent Expanded" = Table.ExpandTableColumn(#"Merged with Province
Lookup", "NewColumn", {"Province in English"}, {"Province in English"}),

    // Merge the the 'Result Lookup' table with 'source data'
    #"Merged with Result Lookup" = Table.NestedJoin(#"Province Equivalent
Expanded",{"Result"},ResultLookup,{"Result"},"NewColumn.1",JoinKind.LeftOuter),

    // Expand 'Result in English' attribute from Result Lookup' table into 'source
data'
```

```
    #"Result Equivalent Expanded" = Table.ExpandTableColumn(#"Merged with Result
Lookup", "NewColumn.1", {"Result in English"}, {"Result in English"}),


    // Change the order of the attributes
    #"Reordered Columns" = Table.ReorderColumns(#"Result Equivalent
Expanded",{"KankorCode", "First Name", "Last Name", "Gender", "Province", "Province in
English", "Kankor Score", "Result", "Result in English"})
in
    #"Reordered Columns"
```

Table 4-12. Power Query code to transform Dari values to their equivalent English terms using replace and merge transformations.

4.1.11. Categorize Provinces into Regions for Regional Analysis

There is a total of 34 different provinces representing the geographical locations of high schools and higher education institutions. It is deemed efficient to categorize these provinces into their respective regions, so they are easier for analysis.

To accomplish the above identified scenario, a lookup table including province names and their corresponding regions is required. The following Wikipedia page (https://en.wikipedia.org/wiki/Provinces_of_Afghanistan) covers all the 34 provinces of Afghanistan plus other details of the provinces, such as, their center, population, area, number of districts, and others. As mentioned earlier in this Chapter, Power Query provides an intuitive interface to connect with and import data from a wide variety of internal data sources such as Microsoft Excel, Text Files and Folders and external data sources such as Websites and other sources including SharePoint, Microsoft Azure, Hadoop, and Facebook, to name just a few. The Following is the Power Query code used to get and parse all 34 provinces of Afghanistan with their respective regions:

```
let
    Source =
Web.Page(Web.Contents("https://en.wikipedia.org/wiki/Provinces of Afghanistan")),
    #"Selected Proper Data" = Source{2}[Data],
    #"Removed Other Columns" = Table.RemoveColumns(#"Selected Proper Data", {"Map #",
"ISO 3166-2:AF[5]",
                                        "Centers", "Population (2015)[6]", "Area
(km²)", "# Districts"}),
    #"Renamed Regional Column" = Table.RenameColumns(#"Removed Other Columns",{{"U.N.
Region", "Region"}})
in
    #"Renamed Regional Column"
```

Table 4-13. This power query code connects to Wikipedia and gets proper data.

The next step is to audit and compare this dataset imported from Wikipedia with the consolidated Kankor Master Table using Power Query merge transformation operation with the proper type of joins as previously described in this Chapter. The following diagram represents a sample output of mapping and categorizing high school geographical locations (provinces) into their corresponding regions:

Kankor Year	High School	High School Location
2012	High school 2	Balkh
2012	High school 9	Daykundi
2012	High school 1	Daykundi
2012	High school 5	Daykundi
2012	High school 6	Daykundi
2012	High school 3	Samangan
2012	High school 8	Daykundi
2012	High school 4	Ghor
2012	High school 2	Daykundi
2013	High school 9	Balkh
2013	High school 2	Balkh
2013	High school 5	Balkh
2013	High school 4	Balkh
2013	High school 5	Herat
2013	High school 2	Herat
2013	High school 7	Balkh
2013	High school 3	Balkh
2013	High school 9	Balkh
2014	High school 1	Kabul
2014	High school 6	Kabul
2014	High school 7	Kabul
2014	High school 4	Daykundi
2014	High school 5	Balkh
2014	High school 3	Daykundi
2014	High school 2	Takhar

Province	Region
Badakhshan	North East Afghanistan
Badghis	West Afghanistan
Baghlan	North East Afghanistan
Balkh	North West Afghanistan
Bamyan	Central Afghanistan
Daykundi	South West Afghanistan
Farah	West Afghanistan
Faryab	North West Afghanistan
Ghazni	South East Afghanistan
Ghor	West Afghanistan
Helmand	South West Afghanistan
Herat	West Afghanistan
Jowzjan	North West Afghanistan
Kabul	Central Afghanistan
Kandahar	South East Afghanistan
Kapisa	Central Afghanistan
Khost	South East Afghanistan
Kunar	North East Afghanistan
Kunduz	North East Afghanistan
Laghman	East Afghanistan
Logar	Central Afghanistan
Maidan Wardak	Central Afghanistan
Nangarhar	East Afghanistan
Nimruz	South West Afghanistan
Nuristan	North East Afghanistan

Kankor Year	High School	High School Location	Region
2012	High school 2	Balkh	North West Afghanistan
2013	High school 9	Balkh	North West Afghanistan
2013	High school 2	Balkh	North West Afghanistan
2013	High school 5	Balkh	North West Afghanistan
2013	High school 4	Balkh	North West Afghanistan
2013	High school 7	Balkh	North West Afghanistan
2013	High school 3	Balkh	North West Afghanistan
2013	High school 9	Balkh	North West Afghanistan
2014	High school 5	Balkh	North West Afghanistan
2014	High school 2	Balkh	North West Afghanistan
2012	High school 9	Daykundi	South West Afghanistan
2012	High school 1	Daykundi	South West Afghanistan
2012	High school 5	Daykundi	South West Afghanistan
2012	High school 6	Daykundi	South West Afghanistan
2012	High school 8	Daykundi	South West Afghanistan
2012	High school 2	Daykundi	South West Afghanistan
2014	High school 4	Daykundi	South West Afghanistan
2014	High school 3	Daykundi	South West Afghanistan
2012	High school 3	Samangan	North West Afghanistan
2012	High school 4	Ghor	West Afghanistan
2013	High school 5	Herat	West Afghanistan
2013	High school 2	Herat	West Afghanistan
2015	High school 1	Herat	West Afghanistan
2015	High school 3	Herat	West Afghanistan
2015	High school 4	Herat	West Afghanistan

Figure 4-27. A sample output of mapping high school locations/provinces into their respective regions.

4.1.12. Derive Multiple Columns from the Result Column for Analysis

A common task after importing data from an external data source is to either merge two or more columns into one or split one column into two or more columns for visualizations and other analysis activities. For instance, there are situations when a column that contains a *full name* is split into a *first name* and a *last name*. Or, a column that contains an *address* field may need to be split into *street*, *city*, *region*, and *postal code* columns. The reverse operation to merge multiple columns into a single column may also be true, such as, to combine the *first name* and *last name* columns into a single *full name* column.

Study of Kankor data indicates there is only one column (aka, attribute) named *result* which carries details very useful and significant for descriptive and predictive analyses such as, *field of study* and *educational institutions* candidates got admitted into. Furthermore, information such as *geographical location of educational institutions* and whether these institutions are public or private also can be derived from the *result* column.

There is no consistent pattern for data in the *result* column. Hence it is not easily possible either to produce accurate reports or to perform more analysis activities. Moreover, lack of a regular and fixed pattern makes it very difficult to split the *result* column into multiple other columns in order to derive the identified details e.g. *field of study*, *name and geographical location of educational institutions* and others.

This structure poses some serious challenges and has undesired side effects for data analysis activities and even for producing reports. For example, it makes it difficult to calculate the total number of candidates admitted into computer science faculty in a particular region/province or to find the total number of candidates for a particular educational institution and other relevant queries.

The author of this thesis proposes the MoHE Kankor Committee to split the *result* column into the following columns in order to improve the structure of the data and enable analysts to perform further analysis activities useful for decisions:

1. Field of study, such as, computer science, engineering, economics, medicine, and even teacher training, and other institutions;
2. Educational institution, such as, Herat University, Kabul University, Kabul Polytechnic University, and even semi-higher education institutions e.g. teacher training colleges, technical and vocational institutions and others;
3. Type of educational institution, such as, public university, private university, teacher training, technical and vocational institutions, Islamic centers;
4. Geographical location of the educational institution, such as, Herat, Kabul, Balkh and other provinces.

As previously mentioned, the result column does not have a fixed pattern, and this makes the process of splitting it difficult – But it does not mean it is impossible for the following reasons:

- **Geographical Location of Educational Institutions:** A close inspection of Kankor data shows that in many cases the province name is part of values in the Result column, and it is possible to determine the geographical location of educational institutions. Generally, the words at the end of the Result column represent geographical location of educational institutions. However, for some educational institutions, such as Shikh Zahid University, Alberoni University, and Pohantun-e Shaheed Professor Burhanuddin Rabbani the province name is not part of the name of these institutions at all – in such cases familiarity with the data and eyeballing can determine the geographical location of such institutions. Additionally, study of Kankor data reveals the following phrases represent that the candidates did not pass the Kankor examination: failed, result-less, fraud and cheating, left the exam, presently studying in one of the country's educational institutions, the candidate has no chance to attend the exam, and a few others.
- **Name of Educational Institutions:** Study of Kankor data shows that mainly the following words within the values of the Result column represent semi and higher educational institutions: university, higher educational institution, teacher training, institute, dar-ul-ulums, dar-ul-heffaz, qualified for semi higher education institution, qualified for private higher education institutions. These words can be located at the beginning or end of the values of the Result column or even in the middle. Like in the above, the following phrases represent that the candidates did not pass the Kankor examination: failed, result-less, fraud and cheating, left the exam, presently studying in one of the country's educational institutions, the candidate has no chance to attend the exam, and a few others.
- **Names of Fields of Study:** Examination of Kankor data also shows most of the values of the Result column contain the following words that represent the fields of study: faculty, teacher training, dar-ul-ulums, dar-ul-heffaz, agriculture institute, interdisciplinary institute, qualified for semi higher education institution, and qualified for private higher education institutions. These words can be at the

beginning or at the end of values of the Result column or even in the middle. Like in the above the following phrases represent that the candidates did not pass the Kankor examination: failed, result-less, fraud and cheating, left the exam, presently studying in one of the country's educational institutions, the candidate has no chance to attend the exam, and a few others.

Taking the above conditions into account, the author of this thesis used Power Query to extract names of educational institutions and fields of study, and their type and geographical locations as follows:

1. It is deemed efficient to rectify the Unicode as well as to remove all the spaces by apply the combination of basic transformation operations such as *trim*, *replace* and *substitute* on the *result* column prior to any other transformation operations described at the beginning of this Chapter.
2. For the sake of performance optimization, it is deemed efficient to exclude other columns except the *result* column. Moreover, it is deemed efficient to perform the *remove duplicates transformation* to address the duplicate instances and narrow down all the instances to their *distinct values*. As a result, out of 1.5 million records of Kankor data, the values of the result column are reduced to around 980 distinct records. Yet for the administrators/analysts, it is not feasible to enter the correct values for the identified columns (*field of study, educational institutions, type and geographical location of educational institutions*) for this number of distinct instances.
3. More advanced text manipulation operations, such as, the combination of *Text.Range*, *Text.PositionOf*, *Text.Start*, *Text.End*, *Text.Replace*, *Text.Length, Text.Contains* and *nested conditions* were used to split the Result column into additional columns such as *field of study, educational institutions, type and geographical location of educational institutions,* described above. Please refer to the following Power Query code for further clarification: For each of these new columns the possible values were extracted – However, the extracted values still need to be corrected.

```
let
    fxPreprocess = (txtInput as text) as text =>
let
    delimiter = " ",
    // Rectifies Unicode for txtInput
    fixUnicode = Text.Replace(txtInput, "ي", "ى"),
    /*
     * Removes leading and trailing spaces,
     * it also removes more than one connected spaces from middle.
    */
    splitText = Text.SplitAny(fixUnicode, delimiter),
    removeBlankItems = List.Select(splitText, each _ <> ""),
    result = Text.Combine(removeBlankItems, delimiter)
```

```
in
    result,
    // Load Kankor dataset for the year 2015
    Source = Kankor2015,
    // Keep the 'Result' column and exlude others
    #"Keep Result Column" = Table.RemoveColumns(Source,{"ID", "KankorCode",
"FirstName", "FamilyName", "FatherName", "GrandFatherName", "Gender", "School",
"GraduationYear", "Province", "Score"}),


    // Invoke 'fxPreprocess' method preprocessing 'Result' column
    #"Fix Unicode and Remove Spaces" = Table.TransformColumns(#"Keep Result
Column",{{"Result", each fxPreprocess(_) }}),


    // Remove Duplicates from 'Result' column to narrow down all instances to their
distinct values
    #"Remove Duplicates" = Table.Distinct(#"Fix Unicode and Remove Spaces"),


    /*
     * Possible patterns to extract possible 'Geographical Location' for Institutions
     *
     * Mostly, end of the 'Result' column represents 'Geographical Location'.
     * However, some instances in 'Result' do not contain province name as part of
their title.
     * Also, following words represent 'failed' status
     * 'Failed' => 'ناکام',
     * 'Result-less' => 'بی نتیجه',
     * 'Cheating' => 'نقل',
     * 'Invalid' => 'باطل',
     * 'Presently student' => 'محصل برحال',
     * 'No more chances' => 'چانس'
    */
    #"Possible Locations" = Table.AddColumn(#"Remove Duplicates", "Extracted
Locations", each
        if Text.Contains([Result], "شهیدربانی") or
           Text.Contains([Result], "شهید ربانی") or
           Text.Contains([Result], "شهیداستادربانی") or
           Text.Contains([Result], "غضنفر") or
           Text.Contains([Result], "پولی تخنیک") then
            "کابل"
        else if Text.Contains([Result], "شیخ زاهد") then
            "خوست"
        else if Text.Contains([Result], "البیرونی") then
            "کاپیسا"
        else if Text.Contains([Result], "ناکام") or
```

```
                Text.Contains([Result], "بی نتیجه") or
                Text.Contains([Result], "نقل") or
                Text.Contains([Result], "فرار") or
                Text.Contains([Result], "باطل") or
                Text.Contains([Result], "محصل برحال") or
                Text.Contains([Result], "چانس") then
            "ناکام"
        else
            Text.Range([Result],Text.PositionOf([Result], " ", Occurrence.Last))),


    /*
     * Possible patterns to extract possible 'Institutions'.
     *
     * Mainly, the following words represent 'Institutions':
     * 'University' => 'پوهنتون',
     * 'Higher Educational Institution' => 'موسسه تحصیلات عالی',
     * 'Teacher Training' => 'تربیه معلم',
     * 'Institute' => 'انستیتوت',
     * 'Dar-ul-Ulums' => 'دارالعلوم',
     * 'Dar-ul-Heffaz' => 'دارالحفاظ',
     * 'Qualified for Semi Higher Education Institution' => ' واجد شرایط به مؤسسات نیمه عالی',
     * 'Qualified for Private Higher Education Institutions' => ' واجد شرایط به مؤسسات تحصیلات عالی خصوصی'
     *
     * If the 'Result' contains one of the above words,
     * then extract it into the new 'Institution' column,
     * otherwise write 'other' into the new 'Institution' column.
    */
    #"Possible Institutions" = Table.AddColumn(#"Possible Locations", "Extracted Institutions", each
        if Text.Contains([Result], "پوهنتون") then
            Text.Range( [Result], Text.PositionOf([Result], "پوهنتون"))
        else if Text.Contains([Result], "موسسه تحصیلات عالی") then
            Text.Range( [Result], Text.PositionOf([Result], "موسسه تحصیلات عالی"))
        else if Text.Contains([Result], "تربیه معلم") then
            Text.Range( [Result], Text.PositionOf([Result], "تربیه معلم"))
        else if Text.Contains([Result], "انستیتوت") then
            Text.Range( [Result], Text.PositionOf([Result], "انستیتوت"))
        else if Text.Contains([Result], "دارالعلوم") then
            Text.Range( [Result], Text.PositionOf([Result], "دارالعلوم"))
        else if Text.Contains([Result], "دارالحفاظ") then
```

```
            Text.Range( [Result], Text.PositionOf([Result], "دارالحفاظ"))
        else if Text.Contains([Result], "واجد شرایط به مؤسسات نیمه عالی") then
            Text.Range( [Result], Text.PositionOf([Result], " واجد شرایط به مؤسسات
نیمه عالی"))
        else if Text.Contains([Result], "واجد شرایط به مؤسسات تحصیلات عالی خصوصی")
then
            Text.Range( [Result], Text.PositionOf([Result], " واجد شرایط به مؤسسات
تحصیلات عالی خصوصی"))
        else if Text.Contains([Result], "ناکام") or
                Text.Contains([Result], "بی نتیجه") or
                Text.Contains([Result], "نقل") or
                Text.Contains([Result], "فرار") or
                Text.Contains([Result], "باطل") or
                Text.Contains([Result], "محصل برحال") or
                Text.Contains([Result], "چانس") then
            "ناکام"
        else "Others"),


    /*
     * Possible patterns to extract possible 'Fields of Studies'
     *
     * Mainly and mostly, the following word represent 'Fields of Studies':
     * 'Faculty' => 'پوهنحی'
     * If the 'Result' contains the word 'Faculty' and the word 'University',
     * then extract the value starting from the 'Faculty' until the word 'University',
     * else if the 'Result' contains the word 'Faculty' and the word 'Higher
Educational Institution',
     * then extract the value starting from the 'Faculty' until the word 'Higher
Educational Institution',
     *
     * Additionally, else if 'Result' contains following words then use them 'Fields
of Studies':
     * 'Teacher Training' => 'تربیه معلم',
     * 'Institute of Management and Accounting' => 'انستیتوت اداره وحسابداری',
     * 'Dar-ul-Ulums' => 'دارالعلوم',
     * 'Dar-ul-Heffaz' => 'دارالحفاظ',
     * 'Institute of Agriculture' => 'انستیتوت زراعت',
     * 'Institute of Interdisciplinary' => 'انستیتوت کثیرالرشتوی',
     * 'Qualified for Semi Higher Education Institution' => ' واجد شرایط به مؤسسات
نیمه عالی',
     * 'Qualified for Private Higher Education Institutions' => ' واجد شرایط به
مؤسسات تحصیلات عالی خصوصی'
     *
     * otherwise, write 'other' into this new column.
```

```
    */
    #"Possible Fields of Studies" = Table.AddColumn(#"Possible Institutions",
"Extracted Fields", each
        if Text.Contains([Result], "پوهنحی") and Text.Contains([Result], "پوهنتون")
then
            Text.Range([Result], Text.PositionOf([Result], "پوهنحی"),
Text.PositionOf([Result], "پوهنتون") - Text.PositionOf([Result], "پوهنحی"))
        else if Text.Contains([Result], "پوهنحی") and Text.Contains([Result], " موسسه
تحصیلات عالی")  then
            Text.Range([Result], Text.PositionOf([Result], "پوهنحی"),
Text.PositionOf([Result], "موسسه تحصیلات عالی") - Text.PositionOf([Result],
"پوهنحی"))
        else if Text.Contains([Result], "تربیه معلم") then
            "تربیه معلم"
        else if Text.Contains([Result], "دارالعلوم") then
            "دارالعلوم"
        else if Text.Contains([Result], "دارالحفاظ") then
            "دارالحفاظ"
        else if Text.Contains([Result], "انستیتوت اداره وحسابداری") then
            "انستیتوت اداره وحسابداری"
        else if Text.Contains([Result], "انستیتوت علوم صحی") then
            "انستیتوت علوم صحی"
        else if Text.Contains([Result], "انستیتوت زراعت") then
            "انستیتوت زراعت"
        else if Text.Contains([Result], "کثیرالرشتوی") then
            "انستیتوت کثیرالرشتوی"
        else if Text.Contains([Result], "واجد شرایط به مؤسسات نیمه عالی") then
            "واجد شرایط به مؤسسات نیمه عالی"
        else if Text.Contains([Result], "واجد شرایط به مؤسسات تحصیلات عالی خصوصی")
then
            "واجد شرایط به مؤسسات تحصیلات عالی خصوصی"
        else if Text.Contains([Result], "ناکام") or
                Text.Contains([Result], "بی نتیجه") or
                Text.Contains([Result], "نقل") or
                Text.Contains([Result], "فرار") or
                Text.Contains([Result], "باطل") or
                Text.Contains([Result], "محصل برحال") or
                Text.Contains([Result], "چانس") then
            "ناکام"
        else
            "Others"),

    // Invoke 'fxPreprocess' method to clean values of extracted columns
```

```
    #"Invoke fxPreprocess" = Table.TransformColumns(#"Possible Fields of
Studies",{{"Extracted Locations", each fxPreprocess(_) }, {"Extracted Institutions",
each fxPreprocess(_) }, {"Extracted Fields", each fxPreprocess(_) }})
in
    #"Invoke fxPreprocess"
```

Table 4-14. Power Query code to extract possible fields of study, institutions, and geographical location of institutions from Result column.

4. As previously mentioned in *step 3* the extracted values for each of the identified columns need to be reviewed and corrected. Therefore, at this point, for each of the identified columns a *reference query* was created from the query in *step 3* and then outputted into Microsoft Excel spreadsheets for further cleansing and transformations.
5. For more clarification, let us examine the outcome of the possible values extracted for the *geographical location of institutions*. The author of this thesis created a lookup table containing almost all the province names written in different variations, as shown in the following figure:

Province Lookup Table

Possible Values	Rectified Values	Possible Values	Rectified Values	Possible Values	Rectified Values
ارزگان	Uruzgan	دایکندی	Daykundi	کندز	Kunduz
بادغیس	Badghis	زابل	Zabul	کنر	Kunar
بامیان	Bamiyan	سرپل	Sari Pul	لغمان	Laghman
بدخشان	Badakhshan	سمنگان	Samangan	لوگر	Logar
بغلان	Baghlan	غزنی	Ghazni	وردک	Maidan Wardak
بلخ	Balkh	غور	Ghor	میدان وردک	Maidan Wardak
مزار شریف	Balkh	فاریاب	Faryab	میدان وردگ	Maidan Wardak
پروان	Parwan	فراه	Farah	وردگ	Maidan Wardak
پکتیا	Paktia	قندهار	Kandahar	ننگرهار	Nangarhar
پکتیکا	Paktika	کندهار	Kandahar	نورستان	Nuristan
پنجشیر	Panjshir	کابل	Kabul	نیمروز	Nimruz
تخار	Takhar	کاپیسا	Kapisa	هرات	Herat
جوزجان	Jawzjan	کندوز	Kunduz	هلمند	Helmand
خوست	Khost	قندوز	Kunduz	ناکام	Failed

Figure 4-28. Province lookup table and for the sake of clarity, it is broken down horizontally.

Then the table containing the extracted values for *geographical location of institutions* was merged with the above lookup table. It shows in most cases (958 instances out of 980 instances in step 2) appeared to be correct – except for 22 instances, as shown in the following figure:

Result	Extracted Locations	Rectified Values
دیپارتمنت طب معالجوی(نظامی) 2 تن اناث 28 تن ذکور	ذکور	Not Matched
دیپارتمنت عمومی(لیلیه ندارد) پوهنحی تکنالوژی معلوماتی ومخابراتی وزارت مخابرات	مخابرات	Not Matched
انستیتوت اداره وحسابداری	وحسابداری	Not Matched
پوهنحی ستوماتولوژی پوهنتون کندز PCB دیپارتمنت	دیپارتمنت	Not Matched
دیپارتمنت فارمسی 1 تن اناث 4 تن ذکور	ذکور	Not Matched
انستیتوت میخانیکی	میخانیکی	Not Matched
انستیتوت زراعت و وترنری	وترنری	Not Matched
انستیتوت کثیرالرشتوی بگرامی	بگرامی	Not Matched
انستیتوت کثیرالرشتوی خاک جبار	جبار	Not Matched
انستیتوت کثیرالرشتوی ده سبز	سبز	Not Matched
انستیتوت زراعت ووترنری میدانکثیرالرشتوی چکاداره وحسابداری دشت توپکثیرالرشتوی جغتوتخنیکی میدان	میدان	Not Matched
تربیه معلم فیض آبادبدخشان	آبادبدخشان	Not Matched
تربیه معلم غوربندپروان	غوربندپروان	Not Matched
انستیتوت کثیرالرشتوی سروبی	سروبی	Not Matched
واجد شرایط به مؤسسات تحصیلات عالی خصوصی	خصوصی	Not Matched
ربودن کتابچه سوالات	سوالات	Not Matched
پاره نمودن ورق جوابات	جوابات	Not Matched
دست کاری در بار کود	کود	Not Matched
پارچه خودرا پاره نموده است.	است.	Not Matched
تغیر در ای دی	دی	Not Matched
جهت حل مشکل به ریاست امتحانات مراجعه نماید.	نماید.	Not Matched
تغییر در آی دی	دی	Not Matched

Figure 4-29. Out of 980 extracted locations only 22 instances did not match the province lookup table.

6. At this point a new column, named *Correct Formats*, needs to be added to the outputted table in which the administrators/analysts can enter the correct values for each instance with missing geographical location. It is worth mentioning that the values entered by administrators are not linked with their corresponding records in the original dataset. It means as soon as the original data or even the sort order of the original data changes the values lose their alignment with their corresponding records. To address this challenge, it is deemed efficient to use *self-reference join,* the same way as it was described to resolve *inconsistency of values for the province column* – and the rest of the steps and procedures are exactly the same. It is worth mentioning that instead of *self-reference join* other methods, such as, (Chris Webb 2014), (Bondarenko 2015), (Feldmann 2016), and (The Power User 2017) were investigated, edited, and experimented by the author of this thesis to resolve the identified challenge of splitting the result column efficiently but the output was not desirable and, in addition, performance optimization was an issue.

4.1.13. Extract Data from the Consolidated Master Table

Power Query is designed such that even less technical users can use it to cleans and reshape the data through its intuitive Graphical User Interface (GUI) features and built-in functions. It, however, can be programmed to give users seemingly unlimited potential to transform the data in just about any way possible using custom code and user-defined functions, as explained in the previous sections of this chapter.

As mentioned earlier in this Chapter it is deemed efficient to combine all the multiple Kankor datasets for the years 2003-2015 into one consolidated master table prior to some further transformation and other analysis activities. Sometimes it is deemed necessary to split this uniform master table into sub tables such as year-wise, location-wise, gender-wise sub-tables

and so on to perform certain analyses and experiments, as illustrated in the following diagram.

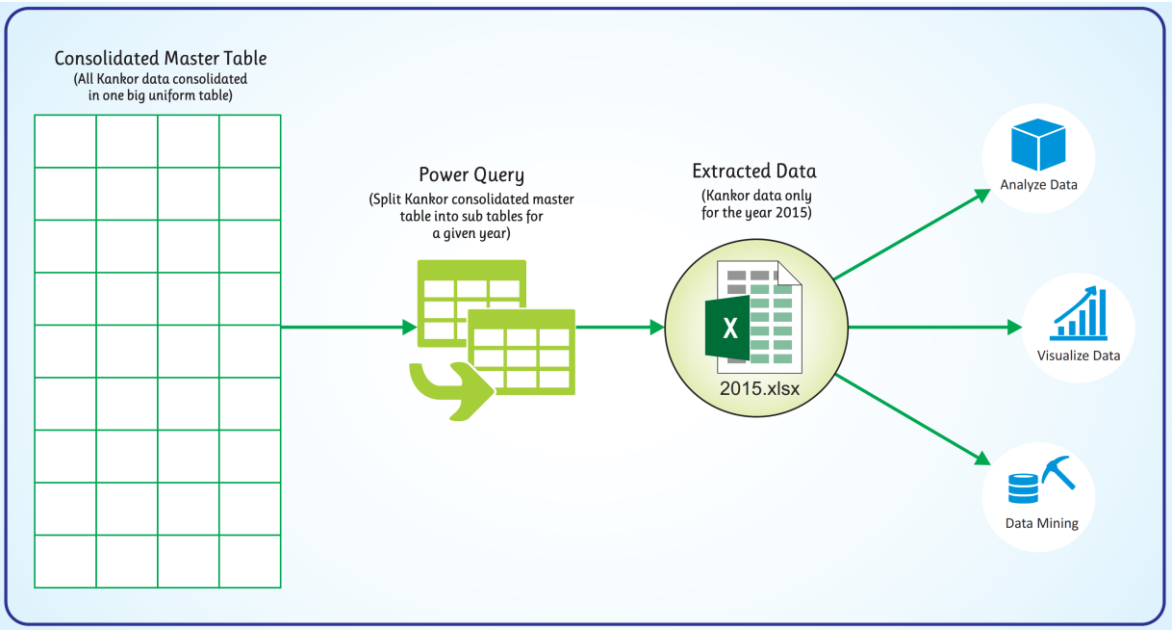

Figure 4-30. Extracts the required data from the consolidated master table (Visualization by the author).

In such cases custom code and user-defined functions are used to extract the required data from the master table. The custom functions are a great approach for sharing logic between queries, separating complex logic from simple ones, making the code more compact and reusable as well as making it easy to locate and isolate a faulty function for further investigations. To accomplish extracting sub data from the consolidated master table, the following Power Query parameterized custom function was created.

```
let
    Source = (#"Filter by Value" as number) => let
    Source = Excel.CurrentWorkbook(){[Name="MasterTable"]}[Content],
    #"Changed Data Types" = Table.TransformColumnTypes(Source,{{"Kankor Year",
Int64.Type}, {"Index", Int64.Type}, {"KankorCode", type text}, {"FirstName", type
text}, {"FamilyName", type text}, {"Gender", type text}, {"School", type text},
{"GraduationYear", Int64.Type}, {"Province", type text}, {"Score", type number},
{"Field of Study", type text}, {"Institution", type text}}),
    #"Filtered Rows by Parameter" = Table.SelectRows(#"Changed Data Types", each
[Kankor Year] = #"Filter by Value")
in
    #"Filtered Rows by Parameter"
in
    Source
```

Table 4-15. Power query parameterized custom function to extract the Kankor data from the master table for a given year as an input variable.

The above function requires a parameter e.g. the year, and accordingly extracts all the Kankor data for the given year as input variable/criteria. The above custom function filters Kankor data from master table for a given year as a criterion. This is deemed efficient to

create a small table for all the possible parameters, named *ParameterList*, as illustrated in the following diagram:

Kankor Year	Gender	Province	Field of Study	Institution
2015	Female		Computer Science	

Figure 4-31. The ParameterList table enabling users to enter criteria and filter Kankor data from master table accordingly.

This technique enables analysts to enter proper values for each parameter and the Power Query function extracts Kankor data from master table based on multiple conditions entered by an analyst for each parameter, as illustrated below:

```
let
    Source = () => let
    // Fetch all the parameters and the values from 'ParameterList' table
    Parameters = Excel.CurrentWorkbook(){[Name="ParameterList"]}[Content],
    P_KANKOR_YEAR = Parameters[Kankor Year]{0},
    P_GENDER = Parameters[Gender]{0},
    P_PROVINCE = Parameters[Province]{0},
    P_FIELD_OF_STUDY = Parameters[Field of Study]{0},
    P_INSTITUTION = Parameters[Institution]{0},
    // Apply filter on the 'MasterTable' based on the criteria set in 'ParameterList'
    Source = Excel.CurrentWorkbook(){[Name="MasterTable"]}[Content],
    #"Changed Data Types" = Table.TransformColumnTypes(Source,
        {{"Kankor Year", Int64.Type}, {"Index", Int64.Type},
         {"KankorCode", type text}, {"FirstName", type text},
         {"FamilyName", type text}, {"Gender", type text},
         {"School", type text}, {"GraduationYear", Int64.Type},
         {"Province", type text}, {"Score", type number},
         {"Field of Study", type text}, {"Institution", type text}}),
    // Conditional Code Branching based on the values of the parameters
    #"Filtered by Kankor Year" = if P_KANKOR_YEAR <> null
        then Table.SelectRows(#"Changed Data Types", each [Kankor Year] =
P_KANKOR_YEAR)
        else #"Changed Data Types",

    #"Filtered by Gender" = if P_GENDER <> null
        then Table.SelectRows(#"Filtered by Kankor Year", each [Gender] = P_GENDER)
        else #"Filtered by Kankor Year",

    #"Filtered by Province" = if P_PROVINCE <> null
        then Table.SelectRows(#"Filtered by Gender", each [Province] = P_PROVINCE)
        else #"Filtered by Gender",
```

```
    #"Filtered by Field of Study" = if P_FIELD_OF_STUDY <> null
        then Table.SelectRows(#"Filtered by Province", each [Field of Study] =
P_FIELD_OF_STUDY)
        else #"Filtered by Province",


    #"Filtered by Institution" = if P_INSTITUTION <> null
        then Table.SelectRows(#"Filtered by Field of Study", each [Institution] =
P_INSTITUTION)
        else #"Filtered by Field of Study"
in
    #"Filtered by Institution"
in
    Source
```

Table 4-16. This Power Query extracts Kankor data from master table based on multiple conditions entered by a user.

Typically, the *ParameterList* in Microsoft Excel is designed vertically instead of horizontally, as illustrated in the following diagram:

Parameter	Value
Kankor Year	2015
Gender	Female
Province	
Field of Study	Computer Science
Institution	

Figure 4-32. ParameterList is formatted vertically.

The following Power Query code, P_GENDER = Parameters[Value]{1}, is used in order to retrieve the value for the parameter Gender that is positioned in the second row – Power Query starts counting from 0, not 1. Essentially, over time this approach may cause problems for example if the position of the Gender parameter is changed from the second position to a different position, because in Power Query all parameters are referenced by their positions and after this change it is required to adjust the reference position for each parameter accordingly.

When the ParameterList is designed in horizontal formation, the following Power Query code, P_KANKOR_YEAR = Parameters[Gender]{0}, is used to retrieve the values for the same Gender parameter. This approach is more robust and ensures it always returns the value for the Gender parameter, no matter where it is positioned within the ParameterList (Feldmann 2015).

4.2. SUMMARY

The following section briefly recaps all the transformation and cleansing steps carried out on the Kankor data and high school marks data prior to using descriptive or predictive analysis activities on these datasets:

Kankor Results Data from 2003 - 2015

The author of this thesis initially created a general structure for the Kankor datasets with the following common attributes (*ID, Kankor Year, Kankor ID*, *First Name*, *Family Name*, *Father's Name*, *Grandfather's Name*, *Gender*, *High School Name*, *High School Graduation Year*, *Province*, *Kankor Score*, *Candidate's Result*).

Then, the data inconsistencies were identified and properly addressed: For example, province names appear in different formats such as (*Balkh*, *Mazar*, and *Mazar-e-Sharif*) using different naming conventions for the same province. Also, the gender column was identified with different values (e.g. *Male, M, 0,* and other Dari/Persian identifiers all as male gender identifiers). Furthermore, invalid data and typos were detected and corrected: For example, in the Kankor data for the year 2003, there were participants with school graduation year of 2007 and 2011 which is impossible.

Next, the missing values for the attributes useful for statistics i.e. *Gender* and *Province* of the candidates' schools were identified and filled with correct values. To address missing values for *gender* and *high school location*, first on the '*First Name*' and then on the '*High School Name*' were entirely sorted out respectively. For most candidates, *gender* and *high school location* were already identified and based on the already available information, the missing values were filled accordingly.

Afterward, to provide a better insight into the data, the 'Candidate's Result' attribute was divided into the following four attributes (*Field of Study*, *Educational Institution*, *Type of Educational Institution*, and *Location of Educational Institution*). *Type of Educational Institution* and *Location of Educational Institution* are only for analysis purposes and basic data mining experiments and are derived from *Educational Institution*. For example, '*Software Engineering at Computer Science faculty of Kabul University only for males*', was broken down into '*Computer Science*' as *Field of Study*, '*Kabul University*' as *Educational Institution*, '*University*' as *Type of Educational Institution*, and '*Kabul*' as *Location of Educational Institution*. It is worth mentioning that typos and different naming conventions were present and corrected during this division process: For example, *computer science*, *computer education*, *computer engineering*, *computer and networks*, *computer and IT*, *software engineering*, and other conventions were replaced by '*Computer Science*'. It is worth mentioning that the location and type of some of the educational institutions could not be identified even with advanced string manipulations techniques. On those special cases, the author's familiarity with the education structure of Afghanistan was useful. Thus, all the conventions were unified.

Next, attributes useful for generation of facts and figures were transformed into their equivalent English terminology. Finally, the private portion of the data including First Name,

Family Name, and other confidential and personal information was excluded to ensure candidate privacy.

School Performance Data 2011 - 2013

Around 6,000 school performance records of high school students from 2011 to 2013 were collected. The students' personal information and the scores for 10^{th}, 11^{th}, and 12^{th} grades were stored separately in different entities and based on the primary key, data was merged together. After this process, 1,500 duplicated data were found which was addressed through the merge rectification process. Then this dataset was compared with Kankor data based on the following attributes (*First Name*, *Father's Name*, *High School Name*, *High School Location*, and *High School Graduation Year*) to find out who had participated in the Kankor exam to evaluate the impact of their high school grades on their Kankor results. Of the 4,500 students, around 3,000 of them had participated in the Kankor. Finally, after validating the data, the data from high schools were merged together with Kankor data and a new dataset with around 3,000 records was formed.

This section provided cases and scenarios of data wrangling capabilities using Power Query – It can be concluded that Power Query makes it easy to *extract* data from almost any source including structured and semi-structured sources. It turned out that Power Query is essentially a macro recorder that keeps track of every step ranging from extraction to data manipulation and transformation. The applied steps will be repeated whenever the data change or when there are different datasets with similar structures – This saves a lot of time and eliminates repetitive cleansing and transformation processes. In the case of Kankor data, once one of the datasets for one year is preprocessed and cleansed, the applied steps are used for the other datasets as well as for future Kankor datasets.

Once the data is prepared and cleansed, a vast array of powerful descriptive and predictive analytical processes can be performed with ease. These analytical processes are described in the following Chapters 5 and 6.

Chapter 5

Chapter 5. Descriptive Analyses using Exploration and Visualization

Having a better picture of Kankor and high school data in Afghanistan through exploration and visualization techniques is essential before employing data mining techniques, recommender systems and other approaches to support both students and educational institutions better. Hence, the author of this thesis carried out exploration and visualization techniques on Kankor and high school data, in which he assessed and analyzed the data from the following angles: candidates' performance, the performance of public and private high schools, candidates' choice of field of study, the contribution of high school scores, etc.

This chapter provides productive contribution to the MoHE, and particularly its Kankor committee, and other educational institutions – by introducing the importance of data as a valuable asset and by providing a proper structure for recording and producing data effective for research studies, thus paving the way for in-depth descriptive and predictive analytics which in turn will improve the situation within the context of the existing structural models in Afghanistan.

The author's research target is to study Kankor using descriptive analytics supported by Kankor Results Data (KRD) of 1.5 million records from 2003-2015; Detailed Kankor Results Data which is a subset of KRD from 2004 – 2006 of 120,000 records; High school scores data from 2011-2013 of 6,000 records; data collected through conducting more than 2500 questionnaires among academicians as well as the author's personal observations, findings, and analyses as a student and faculty member at Herat University in the field of computer science.

5.1. Results of Descriptive Analytics

Data is a valuable asset to any organization. Graham Williams (2011b, 57) mentions in his book that *"Data is the starting point for all data mining – without it there is nothing to mine"*. Also, there is a quote from W. Edwards Deming *"Without data you are just another person with an opinion"*. Statistics as a scientific discipline provides methods to help organizations make sense of data and gain significant insights through exploring the data. These insights can deliver new discoveries that can offer benefits both to policymakers and in data mining projects. Through such insights and discoveries, organizations will increase their knowledge and understanding (Peck and Devore 2011; Williams 2011b).

Correct and clean data leads to accurate and transparent insights, and reports supported by facts and figures. The author of this thesis used several techniques for properly and efficiently cleansing and transforming the data, described in detail in Chapter 4.

More detailed data gives organizations the opportunity to try to discover much more interesting patterns and leads to more specific and significant analytics and insights. Hence,

the author of this thesis also added additional *calculated columns* and *calculated measures* to the data to perform further analyses. The *calculated columns* are evaluated for each row in a table and stored back into the table. The *calculated columns* are just like other columns in a table and can be used in any part of reports, but the *calculated measures* are evaluated and computed at query time on the fly.

An example of a *calculated column* is to classify all Kankor candidates into the following categories based on the type of institute they are admitted into: public university, private university, technical and vocational institution (including teacher training and Islamic education such as Dar-ul-Ulums) or Failed in order to create basic reports and insights. Another example of a *calculated column* is to group the candidates' Kankor scores into buckets such as (0 – 100], (100 – 150], (150 – 200], (200 – 250], (250 – 300], and (300 – 400] in order to analyze the distribution of candidates' Kankor scores in these buckets. Here is another example of a *calculated column* to group all the Kankor candidates into the following two categories: Pass or Fail.

An example of a *calculated measure* is the sum aggregation of all the candidates, which will be evaluated as soon as other *columns* or *calculated columns* are added to the reports or visualizations. Another example of a *measure* is the grand total of the candidates – the previous measure divided by this grand total measure will result in the percentage of the grand total of the participants.

In this section, the author of this thesis explains a few examples from the descriptive and exploration analytics perspectives. This enables organizations and policymakers to understand the 'lay of the land', and in turn, will drive the choice of the most appropriate and significantly applicable tools for preparing and transforming data for future educational data mining applications to improve the educational settings effectively and efficiently.

5.2. Accurate Facts and Figures

In Chapter 4 the author of this thesis properly preprocessed and cleansed the collected Kankor and high school data. For example, all the different identifiers which were used in the gender column to represent male and female were addressed and corrected, and the missing values for the gender column were properly filled out based on the existing ones as described in the previous Chapter. In addition, the *Result* column was divided into the following (*fields of study*, *educational institutions*, *type and location of educational institutions*) columns.

With using proper labels/tags/identifiers for the following attributes: gender, geographical location of high schools, fields of study, educational institutions, type and location of educational institutions, and other important attributes, accurate facts and figures could be easily generated.

5.2.1. Scenario I

Policy and decision makers may require a report to precisely calculate how many of the Kankor candidates got admitted into *public universities*, *private universities*, *technical and*

vocational institutions, or *failed*. With clean data, generation of such reports is easy and accurate as shown in **Figure 5-1**.

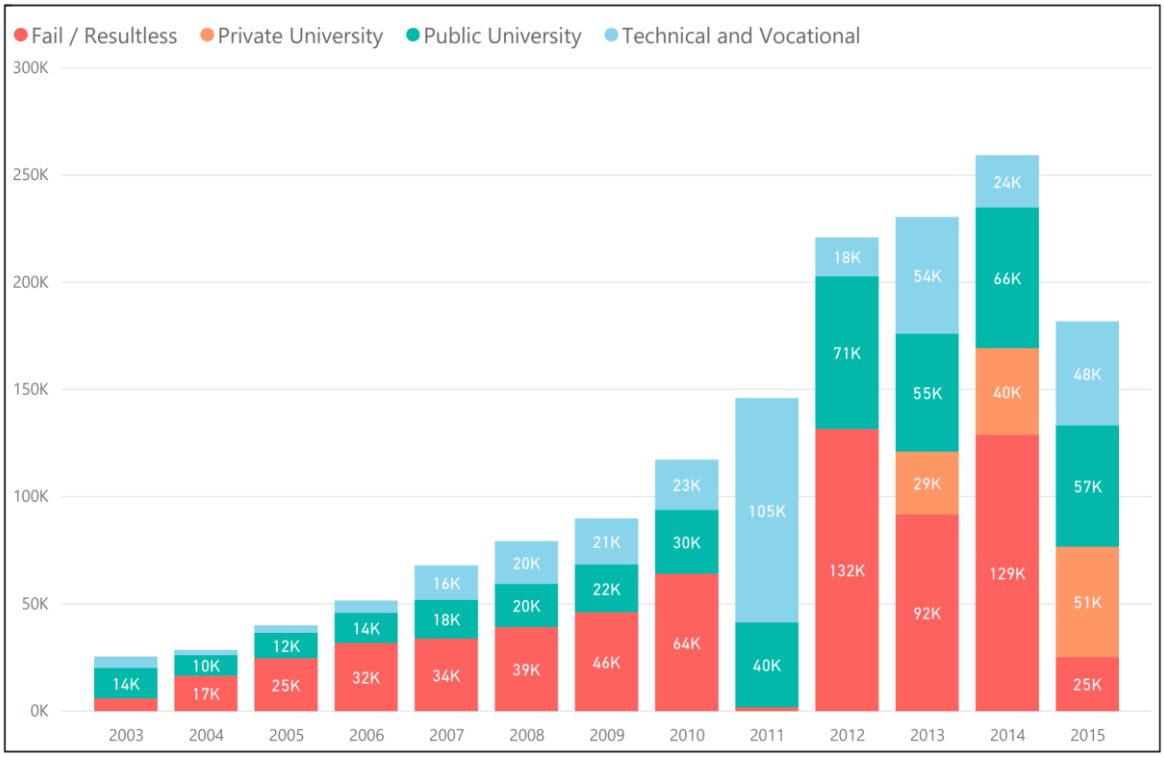

Figure 5-1. Year-wise admission of Kankor candidates to public universities, private universities, technical and vocational institutes or Fail.

It is worth mentioning that MoHE has decided to introduce some of the eligible candidates into *private universities* and *technical and vocational institutions* for several reasons:

- In order to absorb more candidates and decrease the failure rates in the Kankor,
- If too many candidates fail in the Kankor then MoHE and its Kankor Committee will be questioned by members of parliament, and they may even be at risk of losing their positions,
- Also, if a large percentage of candidates are not successful in the Kankor, then MoHE will be under pressure from candidates' families as well as from the media.

Therefore, more than 29,000 candidates in 2013, more than 40,000 candidates in 2014 and more than 51,000 candidates in 2015 were introduced into private universities, as presented in **Figure 5-1.**

Additionally, the data in 2011 shows that around 1,700 candidates failed or were announced as result-less, and the results for around 105,000 candidates were announced by the MoHE like this "All those candidates who failed to enter higher educational institutions can be sent into technical and vocational institutions though MoE". As mentioned in the second chapter of this monograph, the technical and vocational institutions are under the supervision of MoE. Hence, all the candidates who like to study at technical and vocational institutions must process their admission through MoE.

However, neither MoHE nor other organizations carried out research to find out how many of these candidates that were introduced into *private universities* and t*echnical and vocational institutions* are really attending and use these opportunities.

In some years the Kankor data shows that thousands of candidates were introduced into technical and vocational institutions, but it is not mentioned which fields in those institutions (such as agriculture, management, mechanic, etc.) they were to study. Similarly, thousands of candidates were introduced into private universities, but the names for the fields of study were not mentioned.

There may be candidates who are interested to study in private universities or technical and vocational institutions, but even though they performed well in the Kankor, they are not introduced to either. Thus, such negligence (not carrying out research) may make the entire situation unreasonable to them.

The Kankor data in 2012 shows around 81,000 candidates, out of 132,000 candidates, were marked as failed or result-less. Around 35,000 of the candidates who were announced result-less received a score greater than 190 in the Kankor. However out of 50,000 candidates who were marked as failed by the MoHE Kankor Committee only 7 of the candidates received a score greater than 190.

However, the Kankor data in 2013 shows out of 91,625 candidates not even one of them were announced as result-less, and none of them got a score greater than 176 in the Kankor. Only 1,058 of these candidates were marked failed for the following reasons: absence, cheating, impersonation, did not write their form number on the answer-sheets or their form number did not match with their question-sheets form number, and related reasons.

Like the year 2012, the Kankor data in 2014 also shows that out of almost 129,000 failed candidates more than 19,000 candidates were announced as result-less, and around 5,000 of these result-less candidates got a score greater than 190 in the Kankor. There is only 1 candidate who received a score greater than 190 out of around 109,000 candidates that were announced failed in the Kankor.

In the early years the term “result-less” and “failed” were used interchangeably to refer to candidates who failed in the Kankor. Later the term “failed” was used to represent those candidates who really failed in the Kankor, but the term “result-less” was used to represent those candidates who performed well in the Kankor but did not get admitted into their chosen fields of study. The candidates announced as “result-less” are given a chance to write a letter to the MoHE applying for higher education institutions or technical and vocational institutions. Upon successful completion of official processes, they may get a slot.

However, data from previous assessments show there are more candidates with a score less than 190 and even less than 160 but still marked as “result-less”. It is very important to define the role and concept of these two terms explicitly. Presently these terms are vague, and therefore, the author considered both as “failed” during the evaluation and analysis activities.

Finally, the Kankor data shows the number of Kankor participants was reduced from more than 259,000 in 2014 to 181,000 in 2015 – Although the MoHE expected the number of participants to be above 300,000. Due to crises and social problems because of the

presidential election dispute between Dr. Abdullah Abdullah and Dr Ashraf Ghani, most of the young generation left their home country and immigrated to European countries.

5.2.2. Scenario II

Policymakers may ask for a basic report with precise statistics on Kankor participants both gender-wise and year-wise as shown in **Figure 5-2**. This dataset is the same dataset used for visualization in the previous section, scenario I. In the following diagram the "Admission into Types of Higher Educational Institutions" attribute is replaced by "Gender" attribute. Therefore, all the explanation and discussion in the previous scenario are true and apply for this section, scenario II.

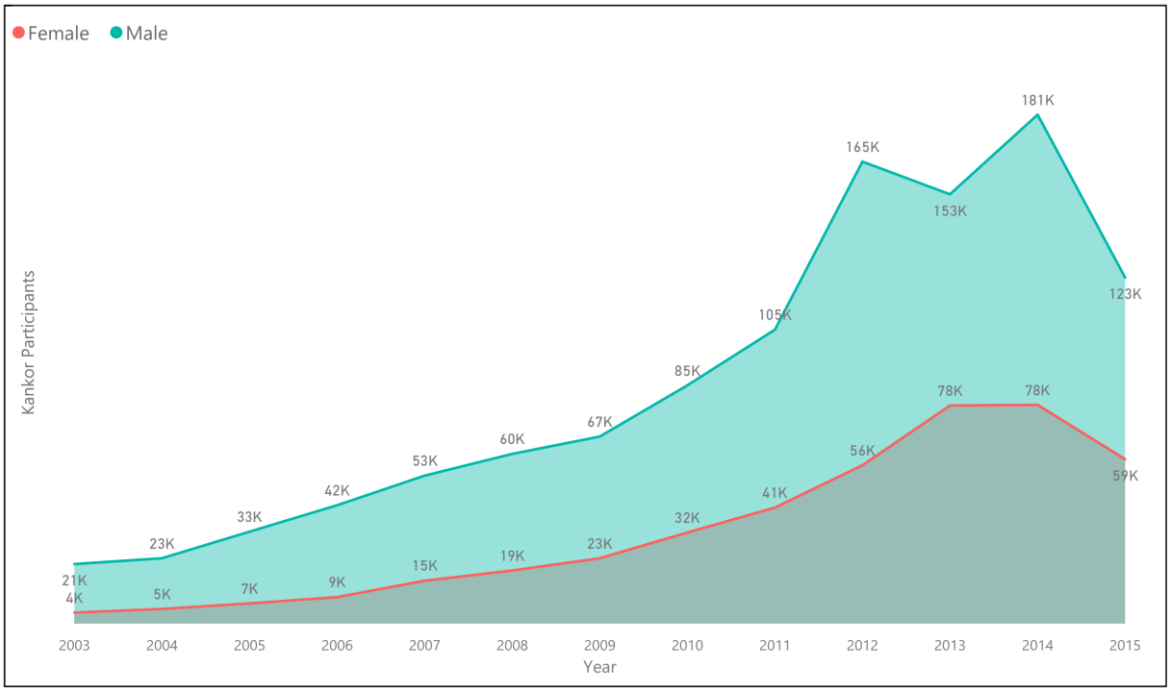

Figure 5-2. Kankor candidates gender-wise and year-wise.

Additionally, the above figure clearly shows the number of female participants is practically proportional to the number of male participants across the years. Except that the year 2013 shows an increase for the female participants and decrease for the male participants. Moreover, the year 2014 indicates an increase in the male participants while the number of the female participants is the same as the previous year. The author did not find any valid sources to show valid reasons for this – The reason could be foundation of new public university campuses e.g. the new campus of Balkh University was built on a 600-acre site in 2013, or the reason could be establishment of females' dormitories in major cities, and relevant reasons. This requires further investigation to find out the precise reasons.

The following figure illustrate that the rates of admission into higher education institutions and technical and vocational institutions are almost the same for both male and female participants. In 2011 the MoHE announced all those candidates who failed to get admission into higher educational institutions may get admission into technical and vocational institutions through MoE, as explained in the previous section, scenario I.

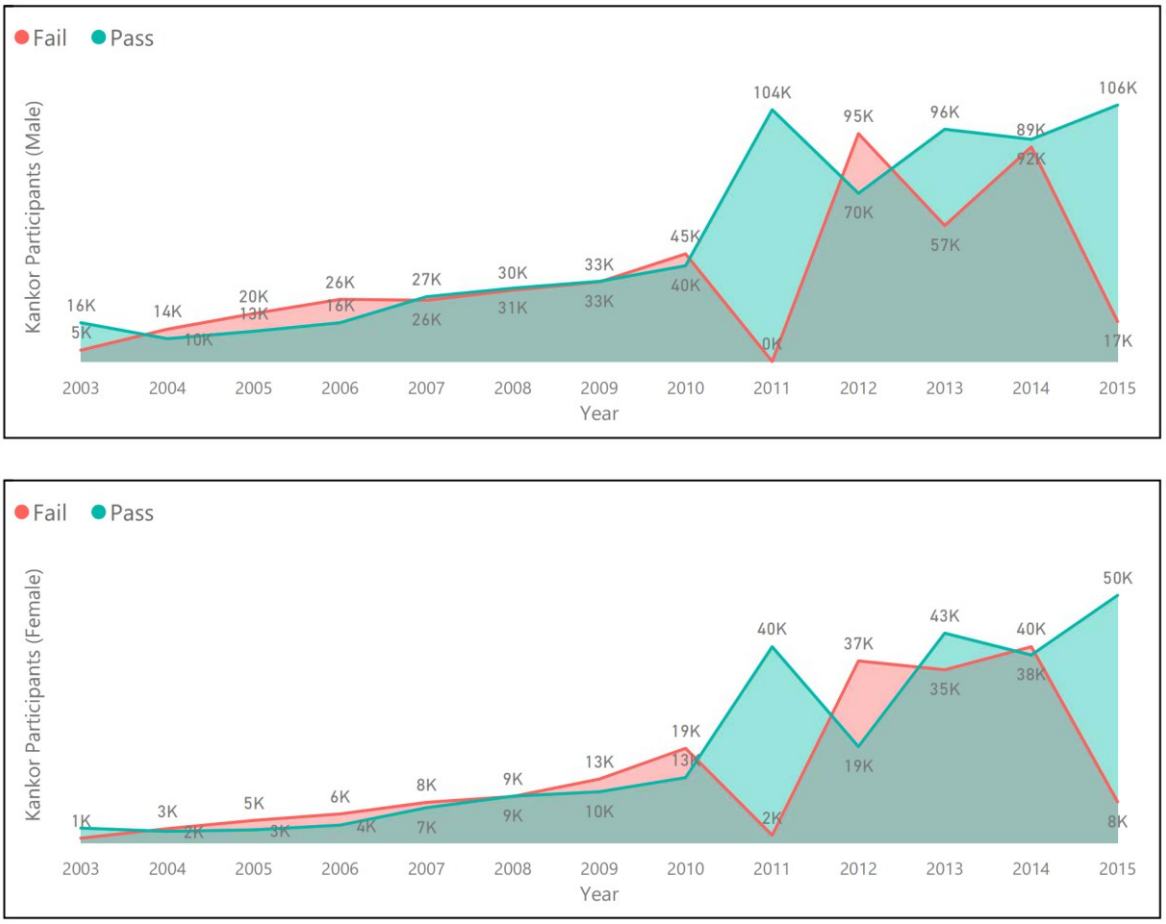

Figure 5-3. This chart illustrates the admission result for both male and female candidates since 2003.

5.2.3. Scenario III

The reports could be complex, and conditions could be grouped to produce further specific and detailed insights such as to generate the *top-n higher educational institutions* or *fields of study* with high enrollment trend location-wise, gender-wise or using other criteria for evaluation and assessment purposes as shown in **Figure 5-4**.

The author's analyses indicate that the enrollment rates are high for *Teacher Training College* and *Faculty of Education (Pedagogy)* across Afghanistan as shown in **Figure 5-4**. Despite this still primary, secondary and high schools suffer from lack of qualified teachers resulting in poor education, *please refer to Chapter 2 for more information and details*. Several factors could cause this:

- low quality in higher education institutions
- recruitment process for hiring of school teachers is not transparent,
- or there could exist another reason which might require further investigation.

Likewise, although *Technical and Vocational Institutions* are ranked the second top as shown in **Figure 5-4**, unfortunately, unemployment and lack of professionals are still noticeable and have increased across the country.

Moreover, statistics also show top enrollment in *Faculty of Agriculture* and *Technical Institute of Agriculture* as shown in **Figure 5-4**. Unfortunately, Afghanistan still suffers from not having experts in the field.

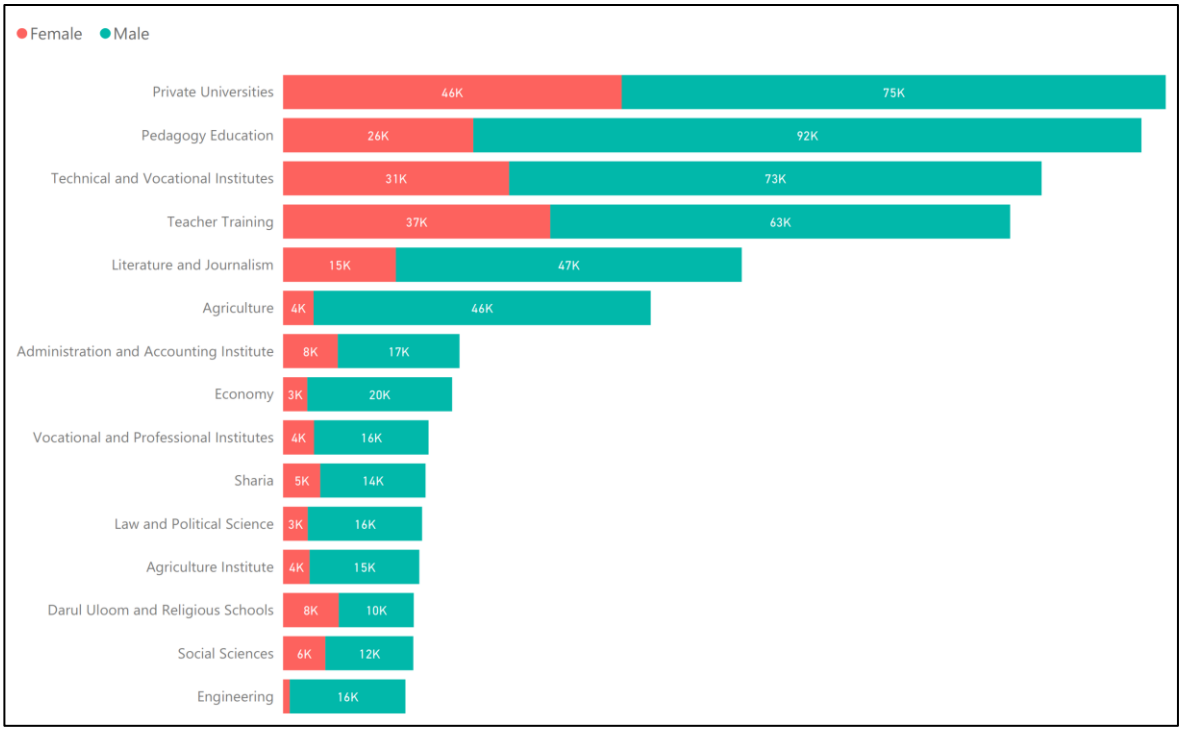

Figure 5-4. The top 15 fields of study in Afghanistan higher education institutions with high enrollment trend since 2003.

Additionally, although admission to private universities (MoHE does not specify name of fields of study for those candidates got admitted into private universities) started in 2003, the above figure reveals admission for both male and female candidates to Private Universities is very high in comparison with all other fields of studies.

Finally, the figure concludes the admission rate of female candidates for fields of study such as Engineering, Sharia, Agriculture and Economy is very low in comparison with male admission rate.

These reports and insights provide an opportunity for policy and decision makers to investigate the causes. Additionally, one can find the fields of study with more or less female enrollment for different Kankor years and or provinces. This will provide the opportunity for policymakers to increase female enrollment in those fields of study and provinces to boost women's role in the society in the future.

It is worth mentioning that these columns with proper labels and identifiers are useful in training and building a prediction model, *more details are provided in Chapter 6*.

5.3. Geographical Reports

The policy makers may require reports on all the participants across the country province-wise and year-wise for analyses as shown in **Figure 5-5**.

Province	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Total
Kabul	10,690	9,477	6,721	8,298	10,722	26,554	32,045	37,371	41,396	69,757	57,033	65,240	38,817	**414,121**
Balkh	1,945	982	1,704	2,001	3,004	4,420	6,215	7,653	10,355	18,658	22,101	23,534	15,168	**117,740**
Herat	1,356	1,644	2,110	2,955	4,111	5,628	7,321	8,592	10,986	12,892	15,506	16,903	12,039	**102,043**
Nangarhar	2,860	2,198	3,535	4,258	5,361	6,033	7,237	7,441	8,832	12,004	10,655	15,041	9,993	**95,448**
Badakhshan	1,150	1,140	2,458	3,176	4,129	4,507	4,801	6,314	8,331	10,452	11,082	11,965	9,837	**79,342**
Ghazni	757	1,301	2,445	3,306	4,446	3,676	3,274	5,643	6,954	8,787	10,677	10,927	9,160	**71,353**
Takhar	692	1,001	1,522	1,954	2,680	2,228	3,064	4,633	5,703	7,296	10,655	11,597	9,336	**62,361**
Baghlan	560	910	1,568	2,139	2,747	2,984	3,536	4,728	5,988	7,692	10,265	10,984	7,034	**61,135**
Parwan	807	1,591	2,684	3,298	4,355	2,134	2,043	3,332	4,808	7,258	9,246	10,190	6,518	**58,264**
Kunduz	388	666	1,062	1,515	2,230	2,161	3,292	3,956	5,656	9,519	9,785	10,787	6,060	**57,077**
Kapisa	537	1,099	1,831	2,341	3,101	2,545	1,489	2,588	3,674	6,317	5,223	6,214	4,162	**41,121**
Maidan Wardak	433	1,244	2,279	3,004	3,761	1,216	1,640	2,830	3,104	3,449	4,054	4,407	4,321	**35,742**
Bamiyan	259	310	631	973	1,326	1,081	1,389	2,315	3,311	5,216	4,858	6,640	5,438	**33,747**
Khost	567	107	841	1,144	1,452	1,118	1,247	2,347	3,051	4,786	4,811	5,510	4,583	**31,564**
Faryab	236	322	495	740	1,022	692	965	1,323	2,160	3,641	6,175	6,621	5,069	**29,461**
Jawzjan	550	441	386	485	676	1,051	1,532	1,728	2,585	3,443	6,023	5,978	4,410	**29,288**
Logar	356	1,024	1,656	2,129	2,561	1,107	1,170	1,879	2,138	2,920	3,381	3,625	3,577	**27,523**
Kunar		404	743	1,124	1,318	1,262	1,163	2,078	2,448	2,978	3,104	3,567	3,184	**23,373**
Laghman	397	806	1,551	1,956	2,366	1,378	685	1,636	2,230	2,615	2,435	2,625	1,744	**22,424**
Paktia	196	265	707	911	1,062	785	785	1,325	1,897	2,651	2,917	3,300	3,052	**19,853**
Daykundi		84	170	239	368	396	361	785	1,564	4,254	3,479	3,795	3,571	**19,066**
Kandahar	282	142	270	346	478	532	606	1,102	1,351	2,997	2,180	2,981	2,028	**15,295**
Samangan		174	299	343	585	312	607	724	1,116	1,737	3,240	3,181	2,042	**14,360**
Panjshir		407	864	1,165	1,499	1,499	414	642	789	1,117	1,607	2,322	1,968	**14,293**
Ghor		59	201	302	541	413	753	1,164	1,628	1,909	2,176	2,386	2,002	**13,534**
Sari Pul		141	246	296	395	145	345	556	833	1,384	2,209	2,430	1,519	**10,499**
Helmand	5	98	194	215	322	335	428	644	783	1,417	1,566	1,437	996	**8,440**
Farah	91	105	285	380	456	406	335	627	762	909	1,101	1,207	1,048	**7,712**
Paktika		133	112	130	235	342	142	381	468	742	779	942	889	**5,295**
Badghis	58	26	60	96	168	61	179	314	379	645	720	676	582	**3,964**
Uruzgan		56	162	163	200	90	51	126	234	565	631	844	644	**3,766**
N/A	252	49	1	19	63	1,885	520	2		40		179		**3,010**
Nimruz		22	42	55	78	136	105	227	177	382	349	491	352	**2,416**
Zabul		17	39	54	86	73	66	138	214	315	331	440	408	**2,181**
Nuristan		12	42	52	64	67	7	65	78	194	163	232	247	**1,223**
Total	**25,424**	**28,457**	**39,916**	**51,562**	**67,968**	**79,252**	**89,812**	**117,209**	**145,983**	**220,938**	**230,517**	**259,198**	**181,798**	**1,538,034**

Figure 5-5. This report shows all Kankor participants across the country province-wise and year-wise.

The report is sorted by the *Total* column in descending order and it confirms that Kabul had the majority of the Kankor participants. In the cases of *Balkh* and *Herat* provinces it can be seen that participants of *Herat* province were more than *Balkh* participants until 2011, but since 2012 the story changes and *Balkh* participants are significantly more than *Herat's*. One of the reasons for these changes could be establishment of the new campuses e.g. building a new campus for Balkh University. These changes require further analysis to find out the positive or negative factors that have caused them.

The report shown in **Figure 5-5** can be drilled down to reveal detailed data e.g. adding the gender column or other columns and then policymakers can analyze the data from different angles by focusing in on different queries.

Furthermore, policymakers may be interested to view the above report shown in **Figure 5-5** region-wise. Although the original Kankor data does not include such a column representing region but it is practical to derive region based on the province column by adding a calculated column. All the provinces in Afghanistan are categorized into the following *6 regions* as shown in **Figure 5-6**.

Region	Provinces
Central Afghanistan	Kabul, Bamiyan, Parwan, Kapisa, Maidan Wardak, Logar, Uruzgan
East Afghanistan	Nangarhar, Laghman
North East Afghanistan	Baghlan, Takhar, Badakhshan, Kunduz, Panjshir, Kunar, Nuristan
North West Afghanistan	Balkh, Jawzjan, Faryab, Samangan, Sari Pul
South East Afghanistan	Khost, Paktia, Kandahar, Ghazni, Paktika, Zabul
South West Afghanistan	Helmand, Daykundi , Nimruz
West Afghanistan	Herat, Badghis, Farah, Ghor

Figure 5-6. This diagram shows classification of provinces region-wise.

Taking the above data shown in **Figure 5-6** into account it is possible to create either a calculated column directly in the Kankor dataset or create a lookup table for the region and then merge it with the Kankor dataset.

Classifying the provinces into regions enables policymakers to create region-wise reports such as the following one shown in **Figure 5-7**.

Region	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Total
Central Afghanistan	**13,016**	**14,801**	**15,964**	**20,206**	**26,026**	**34,722**	**39,827**	**50,441**	**58,665**	**95,482**	**84,426**	**97,160**	**63,477**	**614,213**
Female	2,226	3,050	3,030	3,752	6,550	9,210	11,330	15,977	18,086	24,242	28,604	29,370	20,988	176,415
Male	10,790	11,751	12,934	16,454	19,476	25,512	28,497	34,464	40,579	71,240	55,822	67,790	42,489	437,798
North East Afghanistan	**2,789**	**4,540**	**8,259**	**11,125**	**14,667**	**14,707**	**16,277**	**22,416**	**28,993**	**39,248**	**46,661**	**51,452**	**37,666**	**298,800**
Female	446	685	1,470	2,100	3,022	2,869	3,723	5,475	7,863	10,523	14,873	15,108	11,550	79,707
Male	2,343	3,855	6,789	9,025	11,645	11,838	12,554	16,941	21,130	28,725	31,788	36,344	26,116	219,093
North West Afghanistan	**2,730**	**2,060**	**3,130**	**3,865**	**5,682**	**6,619**	**9,664**	**11,984**	**17,049**	**28,863**	**39,748**	**41,744**	**28,208**	**201,346**
Female	646	535	867	1,081	1,624	2,040	3,355	4,261	5,920	9,276	18,439	17,354	12,776	78,174
Male	2,084	1,525	2,263	2,784	4,058	4,579	6,309	7,723	11,129	19,587	21,309	24,390	15,432	123,172
South East Afghanistan	**1,802**	**1,965**	**4,414**	**5,891**	**7,759**	**6,525**	**6,120**	**10,936**	**13,935**	**20,278**	**21,695**	**24,100**	**20,120**	**145,540**
Female	92	244	546	834	1,314	1,093	867	1,542	2,172	3,405	4,520	4,361	4,009	24,999
Male	1,710	1,721	3,868	5,057	6,445	5,432	5,253	9,394	11,763	16,873	17,175	19,739	16,111	120,541
West Afghanistan	**1,504**	**1,834**	**2,656**	**3,733**	**5,276**	**6,506**	**8,588**	**10,697**	**13,755**	**16,355**	**19,503**	**21,172**	**15,671**	**127,250**
Female	360	482	701	1,081	1,874	2,525	3,438	4,411	6,275	6,355	8,787	8,895	7,049	52,233
Male	1,144	1,352	1,955	2,652	3,402	3,981	5,150	6,286	7,480	10,000	10,716	12,277	8,622	75,017
East Afghanistan	**3,257**	**3,004**	**5,086**	**6,214**	**7,727**	**7,409**	**7,922**	**9,077**	**11,062**	**14,619**	**13,090**	**17,666**	**11,737**	**117,870**
Female	130	156	456	434	738	577	378	529	593	934	1,077	1,365	803	8,170
Male	3,127	2,848	4,630	5,780	6,989	6,832	7,544	8,548	10,469	13,685	12,013	16,301	10,934	109,700
South West Afghanistan	**5**	**204**	**406**	**509**	**768**	**867**	**894**	**1,656**	**2,524**	**6,053**	**5,394**	**5,723**	**4,919**	**29,922**
Female	1	24	86	120	164	131	119	218	422	1,661	1,350	1,449	1,365	7,110
Male	4	180	320	389	604	736	775	1,438	2,102	4,392	4,044	4,274	3,554	22,812
N/A	**9**	**49**	**1**	**19**	**63**	**1,885**	**520**	**2**		**40**		**179**		**2,767**
Female	1	1		2	11	407	31	1		4		14		472
Male	8	48	1	17	52	1,478	489	1		36		165		2,295
Total	**25,112**	**28,457**	**39,916**	**51,562**	**67,968**	**79,240**	**89,812**	**117,209**	**145,983**	**220,938**	**230,517**	**259,196**	**181,798**	**1,537,708**

Figure 5-7. This figure represents Kankor participants across the country region-wise, gender-wise as well as year-wise.

The gender values for 326 participants are missing and even the method used for addressing and filling out missing values for the gender column was unable to fill these missing values. These records are excluded from the above reports for better readability and visibility.

Also, for the 2,767 of the participants either the values for the province column were missing or they completed their high schools abroad and those locations were recorded for the province column. The above report shown in **Figure 5-7** and similar reports provide opportunities for policymakers for instance to evaluate the participation of women in each region and take necessary actions to boost women's enrollment as required by MoHE policy plan. Also, a region may need skilled and educated manpower in education, another region may require talented manpower in agriculture and others in other domains – adding the

column field of study to the report will allow policymakers to evaluate the successful rate of enrollment of participants in fields of study which are most needed in various regions.

5.3.1. Limitations

It is worth mentioning that Kankor data lacks data to denote the original/permanent residence (province) of Kankor candidates. Hence the reports of **Figure 5-5** and **Figure 5-7** do not represent the Kankor participants' province or region of origin. This province represents only geographical location of high schools where Kankor participants completed their twelfth grade. There are participants who graduated from high schools located in *Herat* province but their province and even region of origin may be other than *Herat*. Therefore, with the current Kankor data it is not possible to answer questions such as the number of Kankor participants or percentage of successful participants province-wise and related questions. It is also not feasible to judge the number of Kankor participants and their admission rates considering the population of each province in order to evaluate and improve the quota system. The author of this thesis proposes to the MoHE Kankor committee to record the original residence of Kankor candidates to make further analyses possible and for other researchers in the future.

5.4. Candidates' Performance Assessment

In this section, the author of this thesis carried out analyses to evaluate the distribution of candidates' scores and results in the Kankor. Hence, a *calculated column* was added to the Kankor dataset in order to group the candidates' Kankor scores into the following bins/buckets: (0 – 100], (100 – 150], (150 – 200], (200 – 250], (250 – 300], and (300 – 400]. Statistics and analyses illustrate that most candidates from the year 2003 to 2012 score below 200 in the Kankor as demonstrated in the following **Figure 5-8**.

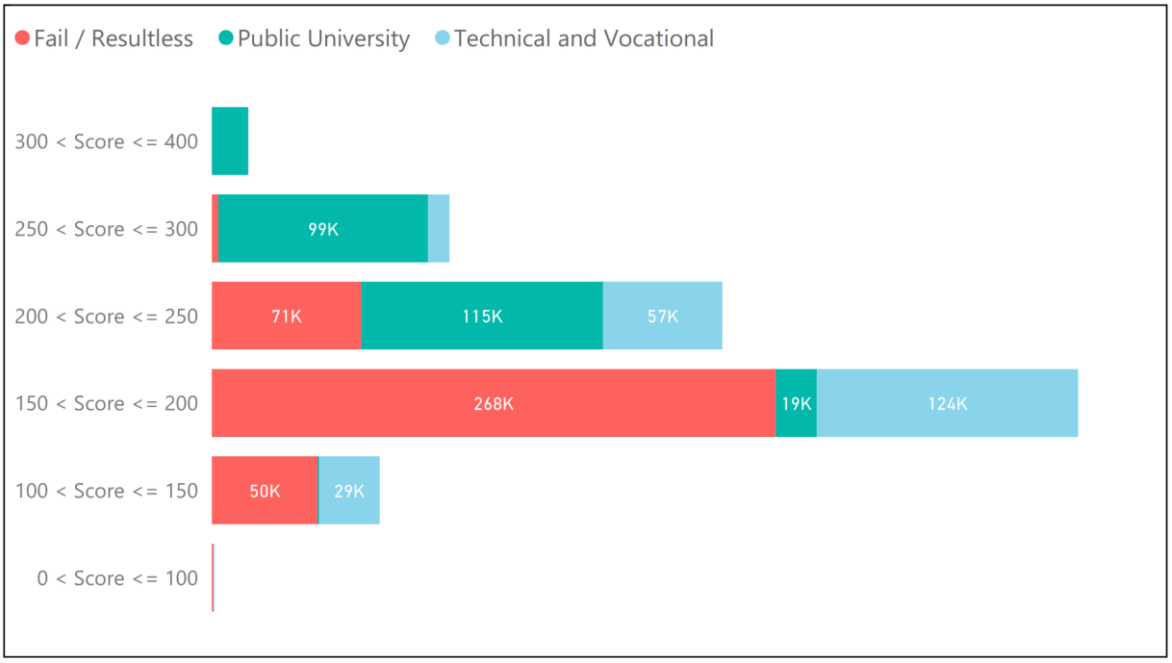

Figure 5-8. This diagram shows the distribution of Kankor candidates' scores as well as their results from 2003 - 2012.

The recent data also confirms that the majority of scores are less than 200 in Kankor, as shown by the following **Figure 5-9** which represents the candidates' Kankor scores and their results in the recent years 2013 – 2015.

It is important to notice that prior to the year 2013 almost 79,000 candidates received a score between 100 and 150, and around 29,000 of them were admitted into higher education and technical and vocational institutions. However, from 2013 onward, none of the candidates with a score in that range got admitted either in higher education institutions or in technical and vocational institutions.

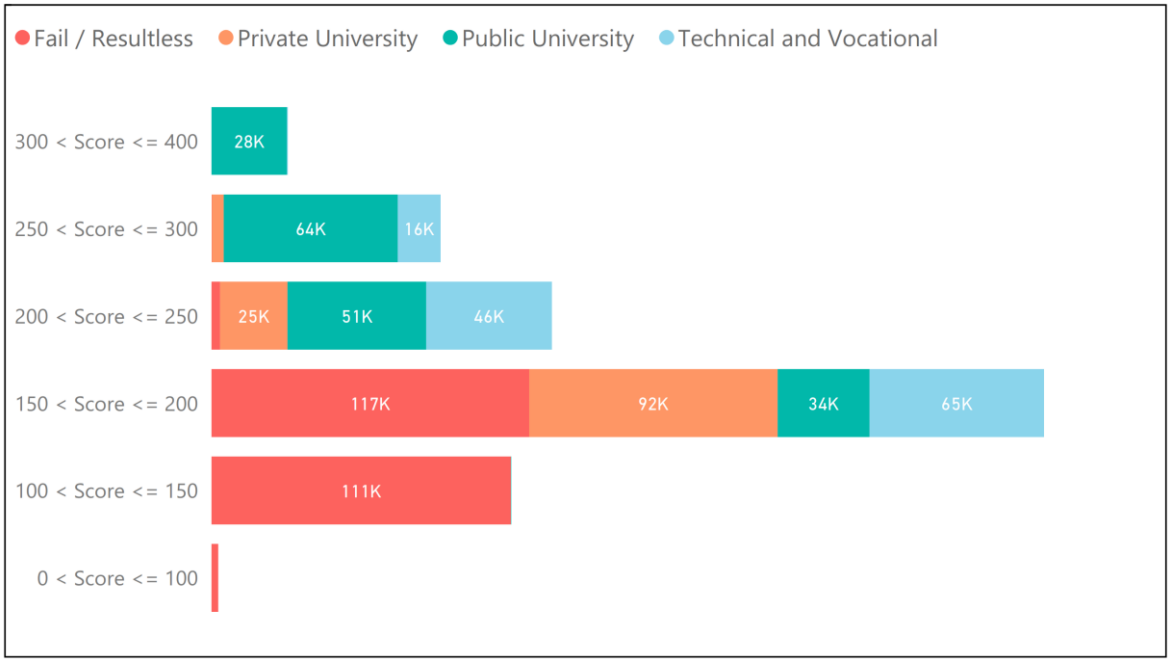

Figure 5-9. This diagram shows the distribution of Kankor candidates' scores as well as their results from 2013 - 2015.

Moreover, there are many candidates with good performance and high scores who got admitted into *private higher education institutions* rather than *public universities*, on the other hand, some candidates with lower performance and score did, as presented in **Figure 5-8** and **Figure 5-9**. It can be concluded that admission also depends on how many slots are available at the different educational institutions, thus, it is 'first come first served'. A candidate can perform well and get a high score yet be unlucky and still not get into his/her most desired field of study. Therefore, a candidate's uninformed/randomness choosing of field of study might lead to unsuccessful admission. For more information and detail about the Kankor and admission process please refer to *Chapter 2 section Education in Afghanistan*.

5.4.1. Limitations

The current structure of Kankor data neither includes the five choices the candidates selected in the Kankor nor the number of available slots for fields of study in higher education institutions. In addition to the descriptive analyses, these data might be significant in predictive analyses such as in building and training a prediction model. I contacted the head

of Kankor Committee of MoHE many times to request him to share with me if such data is available, but I did not receive a positive response. Also, I emailed twice to Prof. Abdul Tawab Balakarzai (Deputy Minister of Afghanistan Higher Education for Academic Affairs, MoHE) and my emails were forwarded to the head of the Kankor Committee and yet I did not get any reply.

5.5. Public and Private High School Performance Assessment

From the policymakers' perspective, it may be required to compare the performance of students from *public high schools* with those from *private high schools* or to perform further analysis to reveal interesting patterns and find answers for other significant queries.

As a matter of fact, the Kankor data lacks information to represent whether high schools are public or private. Thus, it is very challenging to create reports and insights that require knowledge of high school type.

However, in some cases the type of some high schools can be determined from their names. Taking only this as a factor into account does not make it possible to classify high schools into public and private. For example, the following names are written in several variations represent the same private high school:

- هیواد and هیواد (these two names written using different Unicode characters for the letter Yeh "ی". Moreover, these represent the high school in brief);
- خصوصي هیواد, خصوصي هیواد and خصوصیهیواد (the Unicode for the letter Yeh "ی" is different. Additionally, these names consist of information showing that it is a private high school);
- لیسه خصوصي هیواد (the name consists of information showing that it is a private school, and it is high school);
- لیسهخصوصیهیواد and خصوصیهیواد (more than one words are connected and in some those same words are unconnected);
- and there are more variations than the ones listed and elaborated.

Likewise, the following names are written in different formats represent the same public high school:

- انقلاب اسلامي;
- ذکورحضرت عمرفاروق;
- حضرت عمرفاروق;
- انقلاب اسلامی;
- نسوان حضرت عمرفاروق رح;
- نسوان حضرت عمرفاروق (رح);
- انقلاباسلامي;
- نسوانحضرتعمرفاروق (رح);
- and حضرتعمرفاروق.

The followings explain these irregularities and variations in detail:

- there are unnecessary spaces in some of the high school names;

- in others more than one words are connected and in some those same words are unconnected;
- the letter Yeh (ی) is written using different Unicode characters;
- in some cases, only the name high school is written;
- in some other cases the name consists of information showing that it is a high school;
- in some cases, the name contains information showing that it is a private high school;
- sometimes the name includes information that shows the high school is only for males or females;
- and even sometimes there are high schools whose names have been updated during the years.

These inconsistencies make the classification process very complex – sometimes it is even not practical without prior familiarity with the data and the environment.

Taking familiarity of the author of this thesis with high schools in Herat province into account, the Kankor data for the years 2013, 2014 and 2015 were selected for the evaluation purposes. The dataset for the three years consist of 671,256 records. These data were filtered only for Herat province and reduced to 44,371 records of Kankor applicants graduated from high schools located in Herat province. Proper preprocessing, cleansing and transformation steps were applied on the data. Finally, based on the author's knowledge, high schools were classified into public high schools with 37,946 Kankor applicants and private high schools with 6,425 Kankor applicants.

In this section, the author did an assessment to evaluate if private high schools' graduates perform better than public high schools' graduates in the Kankor, as represented in the following figure:

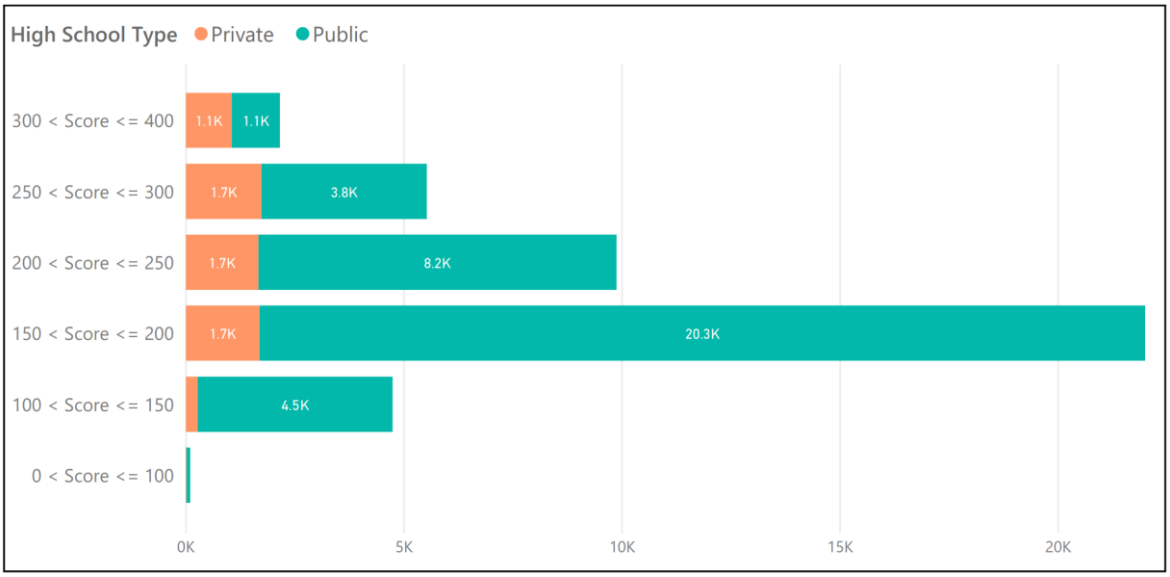

Figure 5-10. Public and private high school graduates' scores in Kankor.

The above figure shows that out of 6,425 Kankor participants from private high schools almost 1,100 received a score in the range of (300 – 400], but out of 37,946 of Kankor participants from public high schools around 1,100 received a score in the same range.

Likewise, it is true for the scores of the Kankor participants in the range of (250 – 300]. Even though the number of participants from public high schools was much larger than that from private high schools, an equal number of graduates from both received scores in the same range – it can be concluded that the Kankor participants from private high schools performed better than the Kankor participants from public high schools.

For further clarification purposes the author added another attribute that shows admission to types of higher educational and plotted the following two charts (figure 11 and 12). Figure 11 shows the Kankor participants scores/performance from public high schools as well as how many of them got admitted into public universities, private universities, technical and vocational institutions or were marked failed/result-less. Figure 12 shows the same information for the Kankor participants from private high schools.

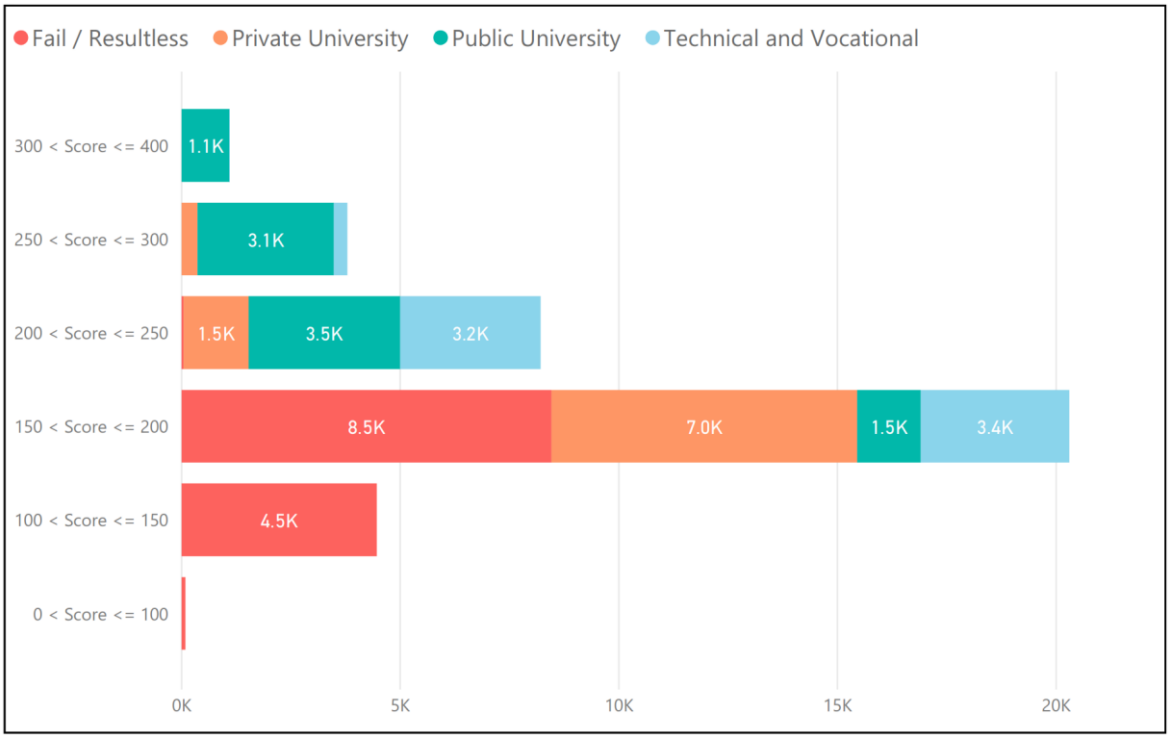

Figure 5-11. Performance and admission of candidates from public high schools.

Both figures 11 and 12 show similar information to that shown in figure 10 as well as the participants admission into types of higher education institutions. The admission of participants from public high schools to public universities is not proportional in comparison to the admission of participants from private high schools. These figures also confirm participants from private high schools were more successful than participants from public high schools.

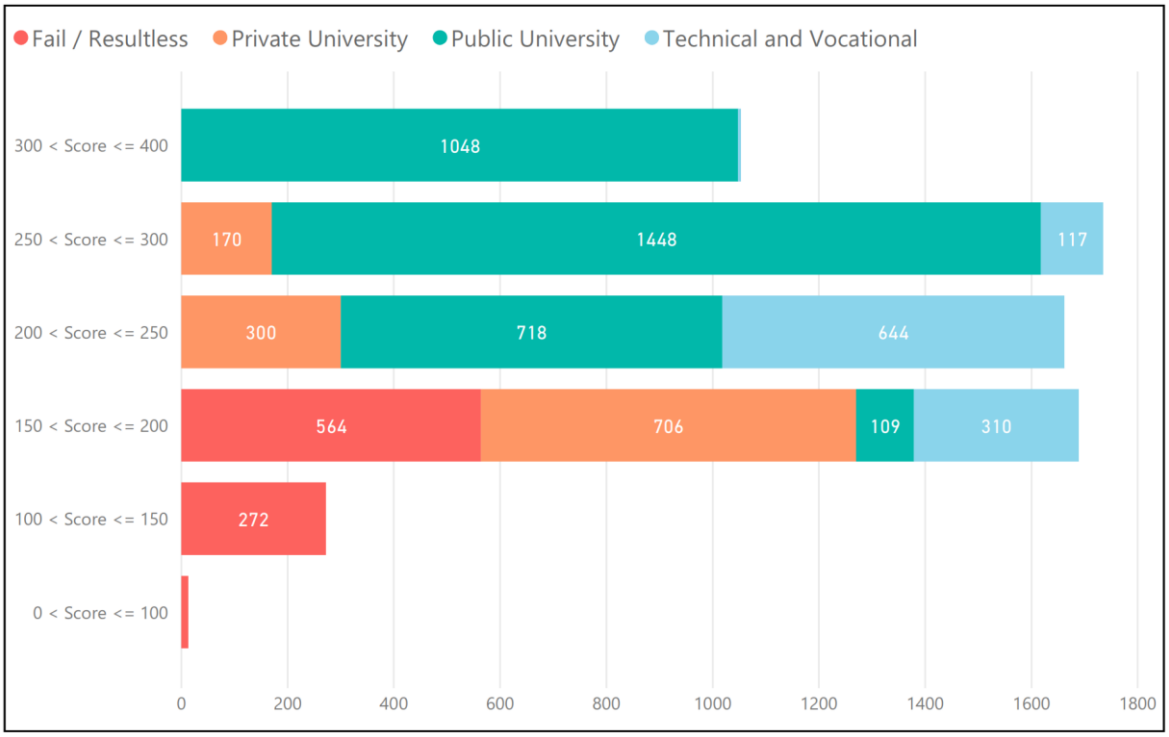

Figure 5-12. Performance and admission of candidates from private high schools.

It is worth mentioning that some private high schools have different strategies as they accept (only) the top and skilled students after they successfully pass different tests and competitions. Some private high schools provide their students more than the official mid-term and final exams for their better evaluation and preparation. Sometimes, some private high schools offer their students Kankor practice tests to make them familiar with the Kankor methods and procedures, Exam Cram, Timeliness, and others.

It can be argued that participants from all private high schools do not perform well in the Kankor. Based on the author's familiarity and observation the following three private high schools (Afghan Turk, Hiwad and Tawhid) are very well known in Herat province. Therefore, the author grouped these three private schools with 3,218 graduates who participated in the Kankor as one cluster, named *first cluster*. All the other more than 20 private high schools in Herat had 3,207 graduates who participated in the Kankor and they were grouped as another cluster, named *second cluster*.

The following figure confirms the argument that participants from *first cluster* perform well in the Kankor in comparison to the participants from the *second cluster* – Despite the fact that the total number of participants from both clusters are equal.

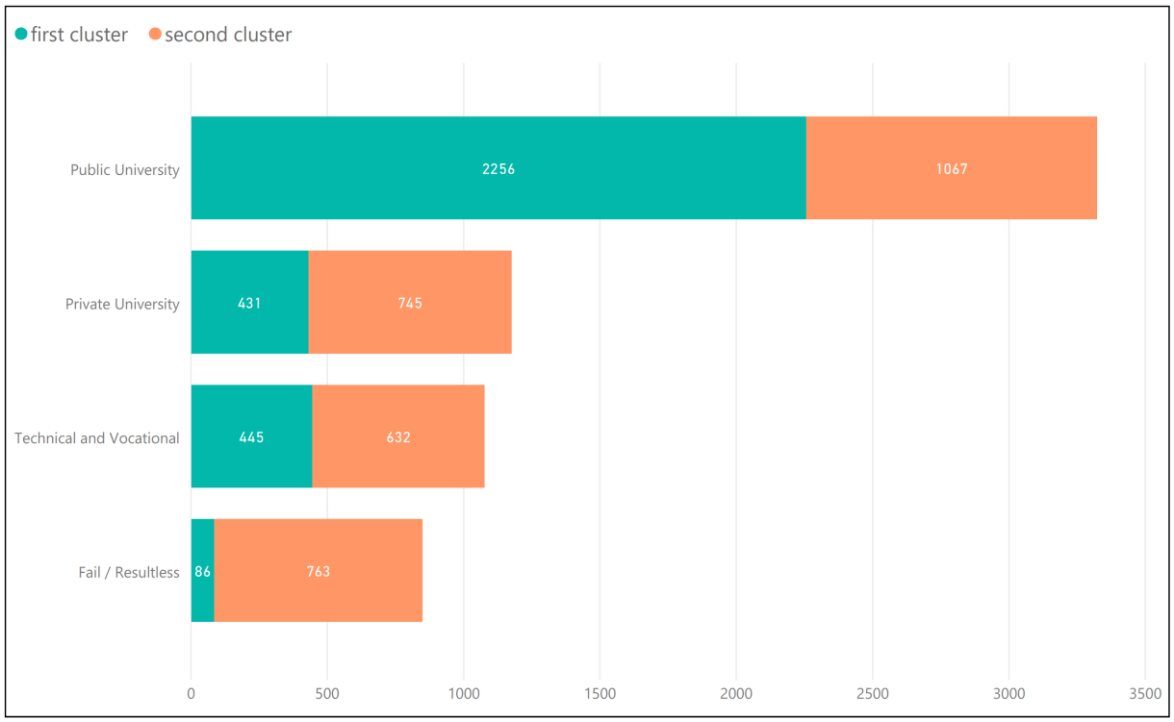

Figure 5-13. Classification of private high schools into two clusters with their admission rates in the Kankor.

The above figure shows the number of participants from *first cluster* admitted into public universities is more than twice than the number of participants from the *second cluster*. Furthermore, the following figure shows that the number of participants who scored in higher ranges such as (300 - 400] and (250 – 300] from *first cluster* is more than the number of participants from the *second cluster* who scored in the same range. The participants from *second cluster* who scored in the range of (150 - 200] are five-times more than the number of participants from the *first cluster* who scored in the same range.

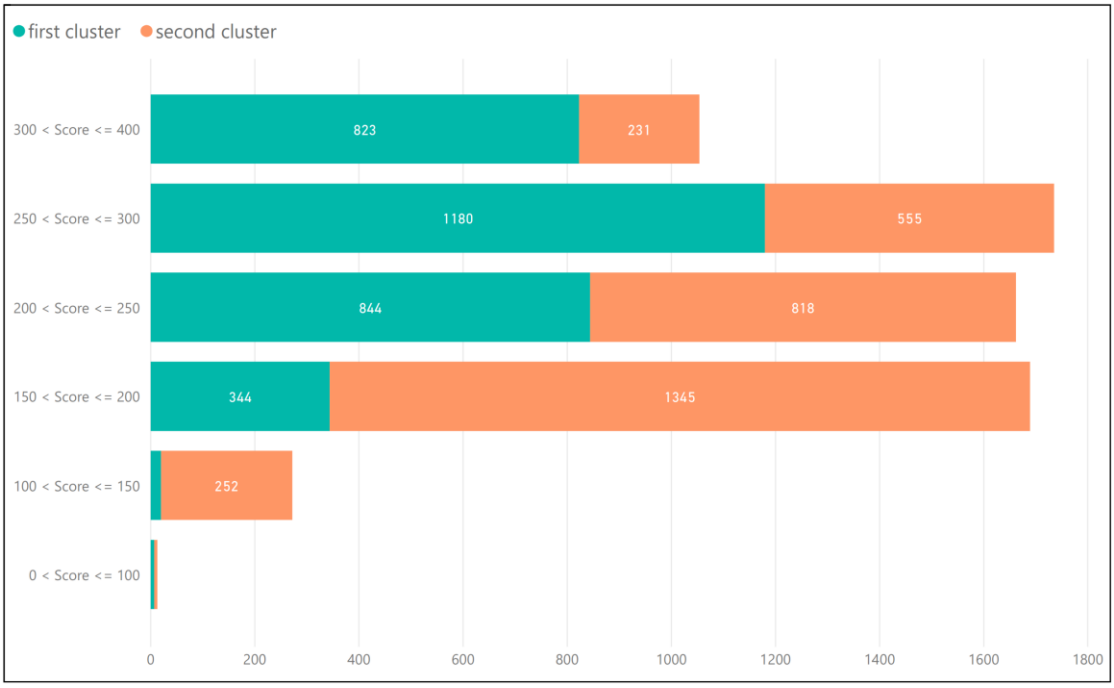

Figure 5-14. Classification of private high schools into two clusters with their scores in the Kankor.

Afghan Turk Private High School (ATPHS) is an International School established in 1995 in Afghanistan. Since 1995, ATPHS opened other branches in several provinces of Afghanistan such as Balkh, Herat, Kandahar, Nangarhar and other provinces. ATPHSs are very popular in Afghanistan. In January 2014, the former Minister of MoE, Ghulam Farooq Wardak said that he would like Turkey to increase the number of Afghan Turk schools in Afghanistan, opening a school in each province as an education role model. Graduates from ATPHSs perform well in Kankor and get high ranks, as represented in the following figure:

There are 1,915 graduates have attended Kankor for the years 2013, 2014 and 2015 from all ATPHSs in Afghanistan. More than 700 of their Kankor participants received a score in the range of (300 – 400] and got admitted into public universities. Also, almost 500 of them got a score in the range of (250 – 300] and got admitted into public universities too. These facts confirm that graduates from ATPHSs perform well in Kankor.

It is worth mentioning that ATPHSs have difficult procedures for enrollment and charge a large tuition fee in comparison to all other private high schools in Afghanistan.

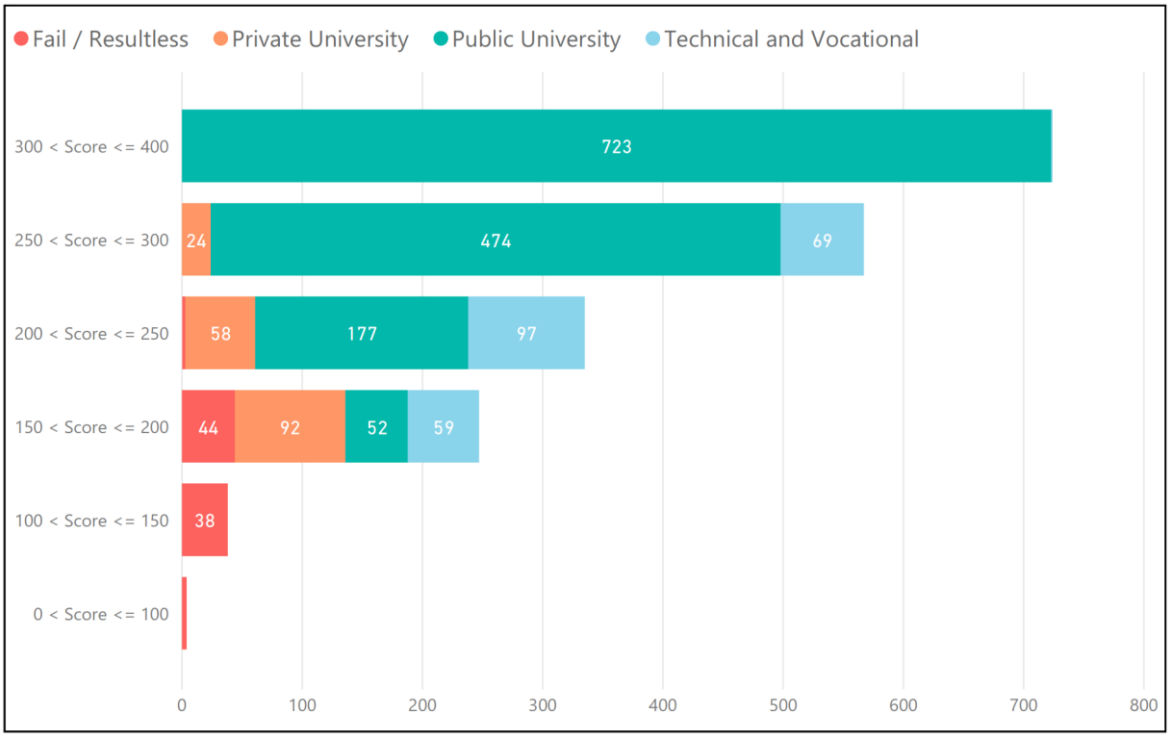

Figure 5-15. Graduates from Afghan Turk Private High Schools across the country.

5.5.1. Limitations

The Kankor data lacks information to represent whether high schools are public or private. The author proposes to MoHE Kankor Committee to record other details for high schools such as public or private, urban or rural along with the other Kankor data. Then the number of successful candidates from the two different high school types can be determined, and the question of whether private high schools prepare Kankor participants better than public high schools can be answered more precisely.

Likewise, it is recommended to record high school names properly. From the current structure of Kankor data it is very complicated to create a report showing the total number of Kankor participants for a particular high school – because different names and formats were used for that high school over the years, even in data which belongs to one year. If high school names are made consistent, in addition to basic reports becoming more accurate other interesting insights can be produced as well as it is practical to classify high schools into clusters prior to training and building a classification model such as a prediction model.

5.6. Assessment of Candidates' Choice of Field of Study

Candidates could select ten fields of study, and one location for each field of study. They also had three chances to participate in Kankor in their lifetime, until 2011. Since then, they can only select five fields of study and have only two chances in the lifetime. For further details please refer to *the section Education in Afghanistan in Chapter 2*.

The Kankor data collected by the author of this thesis for about 120,000 candidates from the years 2004 – 2006 shows the choices candidates got admitted into but does not show what fields of study they had selected. Additionally, these data contain additional data such as Kankor candidates' scores for Languages, Mathematics, and Natural and Social Sciences. Please refer to Chapter 4 for more details about the collected data.

Taking this data into account, the author of this thesis evaluated the candidates' acceptance based on the order of their choice of field of study in the Kankor. It can be concluded that some candidates with good performance and score did not get admitted into their first favorite choice of field of study. On the other hand, some candidates with lower performance and score did, as presented in **Figure 5-16**.

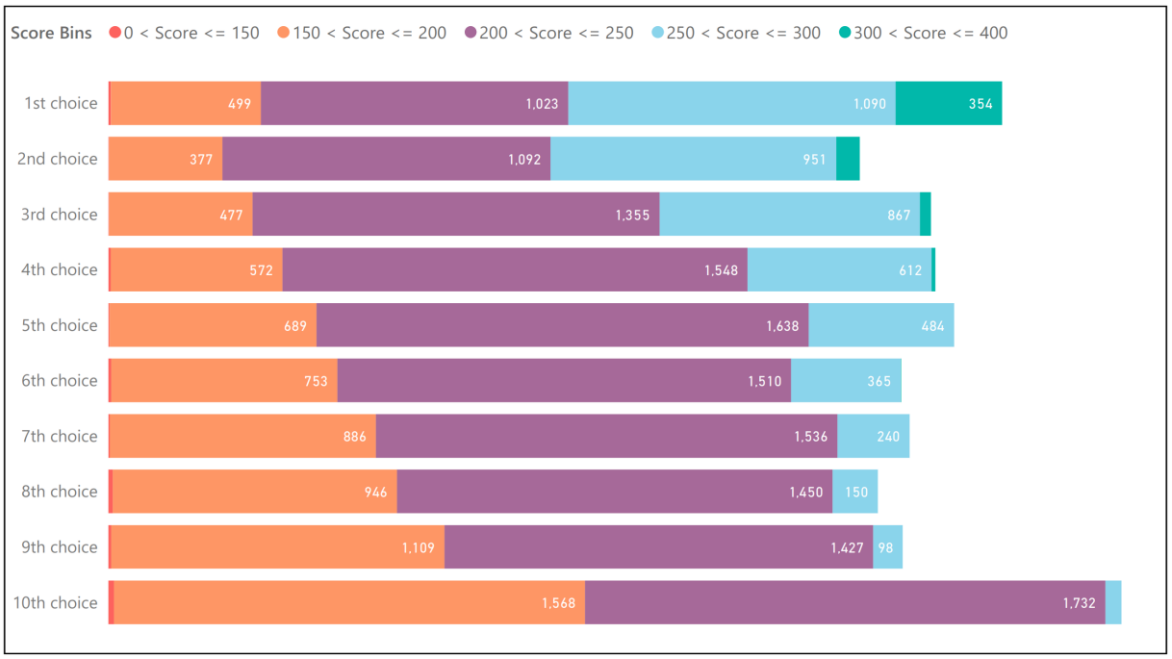

Figure 5-16. This visualization shows the candidates admission by their choices from 1 to 10.

One of the reasons that some candidates with lower scores in the Kankor got admitted into their first-choice field of study and some with higher scores did not is the order of the

selection of the choices. For example, **candidate x** had a score of 280 in the Kankor. S/he chose Medicine 1st, Stomatology 2nd, Computer Science 3rd, Engineering 4th, and Economics 5th. On the other hand, **candidate y** had a score of 260. S/he chose Computer Science as 1st choice. Assume that the minimum admission scores for Medicine and Stomatology are greater than 280, and for Computer Science it is less than 260. So, **candidate y** gets admission into Computer Science. All other candidates with eligible admission scores and Computer Science as their first-choice field of study will be admitted into Computer Science. Now if there are no free slots left for Computer Science **candidate x** and other candidates like **candidate x** do not get admission into Computer Science. Thus, it can be said that proper choice of field of study is very critical and it is recommended that candidates be offered advice by high school advisors to make sure their choices in Kankor better match their skills. But Afghan high schools are not equipped with systematic tools to evaluate Kankor candidates' skills and preparation for Kankor and fields of study. Nor is there a counseling mechanism in the structure of high schools to advise candidates concerning their academic career.

5.6.1. Limitations

The data about what five or ten fields of study the candidates select in Kankor can have significant uses in both descriptive and predictive analyses. To name a few:

- Policymakers may be interested to evaluate and determine the most popular fields of study selected by candidates, gender-wise, province-wise, year-wise and considering other conditions;
- Policymakers may be interested to know and compare the candidates' top five or ten choices. For example, they may want to compare the most selected field with the 2nd with 3rd most selected field and so on to find out if their choices are closely relevant or if choices are made randomly.
- Policymakers also may be interested to know the number of candidates who successfully got admitted into their 1st choice, 2nd choice, etc.
- Researchers may be interested in mining candidates' choices together with minimum and maximum admission scores for each choice and a combination of other useful attributes to organize the existing fields of study into main streams.
- This attribute may also play a crucial role to train and build a classification model such as a prediction model to recommend the right fields of study for candidates.

Presently, the Kankor data does not show the five choices of the candidates (such as Medicine, Engineering, Computer Science, etc.). The data does not even show whether candidates got admitted into their 1st, 2nd, 3rd, 4th, or 5th choices. Hence the current structure of Kankor data does not help policymakers and researchers to perform either descriptive or predictive analysis activities on the candidates' choices. But if such data is available then questions like "what fields of study are popular and trendy among the candidates?" could be answered. Also, it could be determined if candidates choose fields of study systematically, or at random. This will give opportunities for policymakers to introduce proper policies to help candidates choose fields of study in a more informed manner.

5.7. High School Performance and Grades Assessment

In some countries, high school data and scores/marks are considered a vital factor and are used as a basis for candidates' admission into higher education.

In Iran, for many years elimination of Kankor has been under discussion, and since 2012 it has been decided that candidates' high school GPA replace Kankor for determining candidates' admission into higher education. Also, in Afghanistan, in the past this subject was discussed several times as an alternative option to replace Kankor. However, these decisions have not been put to practice either in Afghanistan or in Iran, due to nepotism and selling of scores in high schools and other concerns and challenges.

In addition to the above scenario, it is very important to evaluate and demonstrate the importance of high school marks and their role in the success of the candidates in the Kankor. In Afghanistan, although MoE has planned to record high school marks for the senior three years in an electronic system, they are not yet available in electronic format – despite this the author of this thesis collected high school data for 6,000 students from Herat province consisting the students' personal information as well as their marks for grades 10, 11 and 12.

The 6,000 records were reduced to around 4,400 records after performing matching and removing duplicate transformations. Out of 4,400 records around 4,000 students had completed their high schools in Herat province and the other 400 had completed their high schools in other provinces.

To find out the relation between high school marks and the success rate of candidates in the Kankor exam, it is required to merge and compare the candidates' high school marks with the Kankor data. Since there is no common key between the high school's data and Kankor the only solution is to use the combination of following common attributes for comparison: First Name, Father's Name, Province, High School Name and Graduation Year.

Afghanistan consists of 34 provinces, hence, the values for the 'Province' attribute are limited and can be rectified with basic preprocessing and transformation steps. The values for the 'Graduation Year' are numeric and relatively straight-forward to be preprocessed and cleansed.

However, the values for the candidates' First Names, Father's Names and High School Names are tricky and cannot be rectified using basic preprocessing and transformation steps, as explained in Chapter 4. The author of this thesis proposed four different approaches to address the issue of comparing First Names, Father's Names, and High School Names, as elaborated in Chapter 4.

The combination of the mentioned common attributes was used to simulate the application of common key between the two datasets. High school data were for students who graduated from high schools located in Herat and for the years 2011, 2012 and 2013. Therefore, all the 1.5 million records of Kankor data were reduced to 37,196 records consisting of only the data for the candidates who attended Kankor in 2011, 2012, and 2013 and also completed their high schools in Herat province.

The high school data consisting of 4,000 records was merged with the Kankor data consisting of 37,196 records – the merge transformation using 'Exact String Matching' technique resulted in 1,972 records. However, the merge transformation after performing 'Strip All Spaces' technique on the First Name, Father's Name and High School Name attributes resulted in 2,966 records. This means out of 4,000 high school data; 2,966 students were found to have attended Kankor.

Then the author made an assessment in order to find out the level of correlation between the candidates' high school marks with their scores and results in Kankor. The outcome of the analysis depicts that high school marks are not strongly correlated to the success of candidates in Kankor since there are public and private high schools whose scoring systems are not trustworthy and realistic, as illustrated in **Figure 5-17**.

The following figure shows that almost 1/4 (one-fourth) Kankor candidates with great performance in high schools and with a GPA in the range 90% to 100% failed to pass the Kankor exam and did not get admission to any higher education institutions. Likewise, out of 570 Kankor candidates with a high school GPA of between 80% and 90%, 218 of the candidates failed Kankor exam and 42 of the candidates were introduced to private universities. In the same way, out of 788 Kankor candidates with a high school GPA between 70% and 80%, 429 of the candidates failed in the Kankor and 86 of them were introduced to private universities.

This can be concluded that with the current structure, high school performance and marks are misleading. Therefore, it is not reasonable to only consider the three years high school performance and marks as a vital factor for assessing candidates' admission rates into higher education institutions.

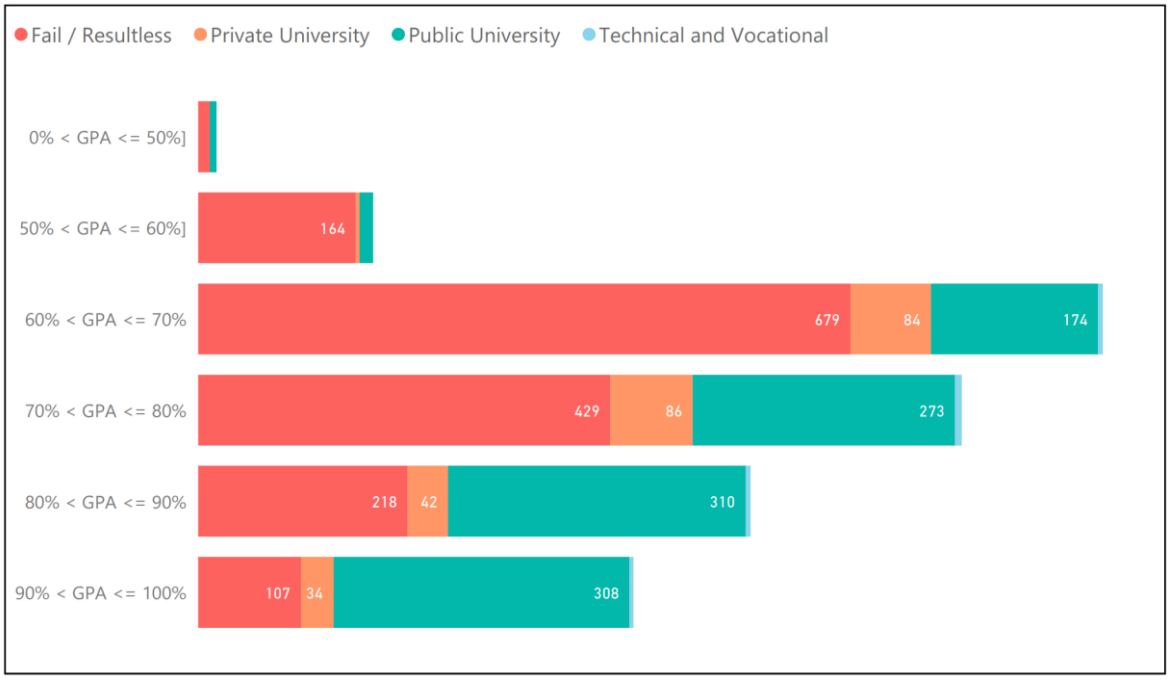

Figure 5-17. This visualization shows the relationship between high school GPA and success of candidates in Kankor.

In addition, from another assessment and evaluation done on the DKRD, it can be concluded that success in Kankor is not dependent on any one of the following categories of questions: Mathematics, Natural Sciences, Social Sciences, and Languages. One should get high scores in all the categories in order to get admitted into a prestigious field of study and university.

5.7.1. Limitations

If the five choices each candidate makes in Kankor and their score details are stored properly, the author proposes that further analyses can be performed to find out interesting patterns.

It is significant to compare and merge high school and Kankor data to perform further analysis activities using descriptive, predictive and recommender systems methods. Presently, high school data are not available in electronic format and maybe only basic data such as high school marks are stored. The author of this thesis recommends strong cooperation between MoE and MoHE in order to introduce a proper method to collect the required and useful data with the possibility to compare and merge them with Kankor data for further analysis activities.

5.8. KANKOR ADMISSION SCORE ASSESSMENT

The Kankor admission score overlaps for some fields of study. For example, a candidate with a Kankor score of 290 is eligible for Computer Science, Economics, Engineering, and maybe other fields of study. Moreover, the admission score varies for the same field of study from province to province and from one higher education institute to another. For example, the minimum admission score for Computer Science at Kabul or Herat Universities is 260 while for Nangarhar, Kunduz, and Badakhshan Universities the minimum admission score for it is around 180.

The minimum and maximum of the Kankor admission scores alone may not represent the whole range of scores and cannot be good metrics to rely on and classify the fields of studies at a university or the same field of study that is offered at several educational institutions. It is deemed efficient to consider other indicators and consider aggregations such as mean, median, variance, and standard deviation to classify the fields of study at a university, for example to classify all the 15 fields of study at Herat University into N clusters. Or to classify a field of study such as Computer Science that is offered by 12 public universities across the country by province or by university. These classification techniques reduce the number of options and are useful in classification models e.g. prior to training and building a prediction model. Please refer to the next chapter for more information and case studies.

Hence the author of this thesis carried out experiments on Kankor data for the years 2013, 2014, and 2015 to evaluate the Kankor admission scores for the fields of study offered at Herat University using the mentioned aggregations, as demonstrated in the following **Figure 5-18**. The author chose the data for the years 2013, 2014, and 2015 for the following reasons:

1. Prior to year 2013 a few of these fields of study were not offered at Herat University. But since 2013 all these 15 fields of study are offered at Herat University.
2. The outcome widely varies for the data only for the years 2013, 2014 or 2015.

Field of Study	Count	Min	Q1	Q2	Q3	Max	Mean	Median	Mode (Freq)	StdDev
Medicine	460	298	321	332	337	349	328	332	332 (33)	11.81
Stomatology	310	292	307	322	326	345	318	322	324 (26)	12.31
Engineering	451	199	289	315	326	347	306	315	321 (16)	25.04
Law and Political Science	482	272	288	309	316	342	304	309	316 (26)	16.9
Economy	360	267	291	308	316	347	304	308	317 (19)	16.74
Computer Science	463	262	289	302	311	339	300	302	297 (21)	16.01
Public Administration	455	256	274	287	301	325	287	287	283 (22)	15.96
Literature and Journalism	1709	188	246	270	290	343	264	270	291 (55)	31.33
Sharia	1000	187	249	266	281	338	257	266	262 (27)	36.36
Science	1529	184	235	262	281	350	257	262	259 (31)	31.85
Agriculture	833	179	221	253	272	330	248	253	254 (21)	31.21
Fine Arts	600	194	230	247	273	334	251	247	298 (18)	29.21
Pedagogy Education	2743	176	214	241	257	318	238	241	240 (69)	30.09
Veterinary	240	209	231	241	256	311	244	241	231 (11)	20.6
Social Sciences	1079	178	220	231	246	322	234	231	229 & 222 (32)	22.79

Figure 5-18. This matrix shows Kankor admission scores for all the 15 fields of study at Herat University.

From the above matrix shown in **Figure 5-18** it can be concluded that the admission scores for Medicine and Stomatology at Herat University are relatively similar and these fields can be classified into one cluster. Similarly, the admission scores for Engineering, Law and Political Science, Economy and Computer Science are very close, and these fields can be classified into another cluster, etc. For further clarification, the author of this thesis also plotted side by side box plots of the Kankor admission scores vs. the fields of study at Herat University, as illustrated in **Figure 5-19**.

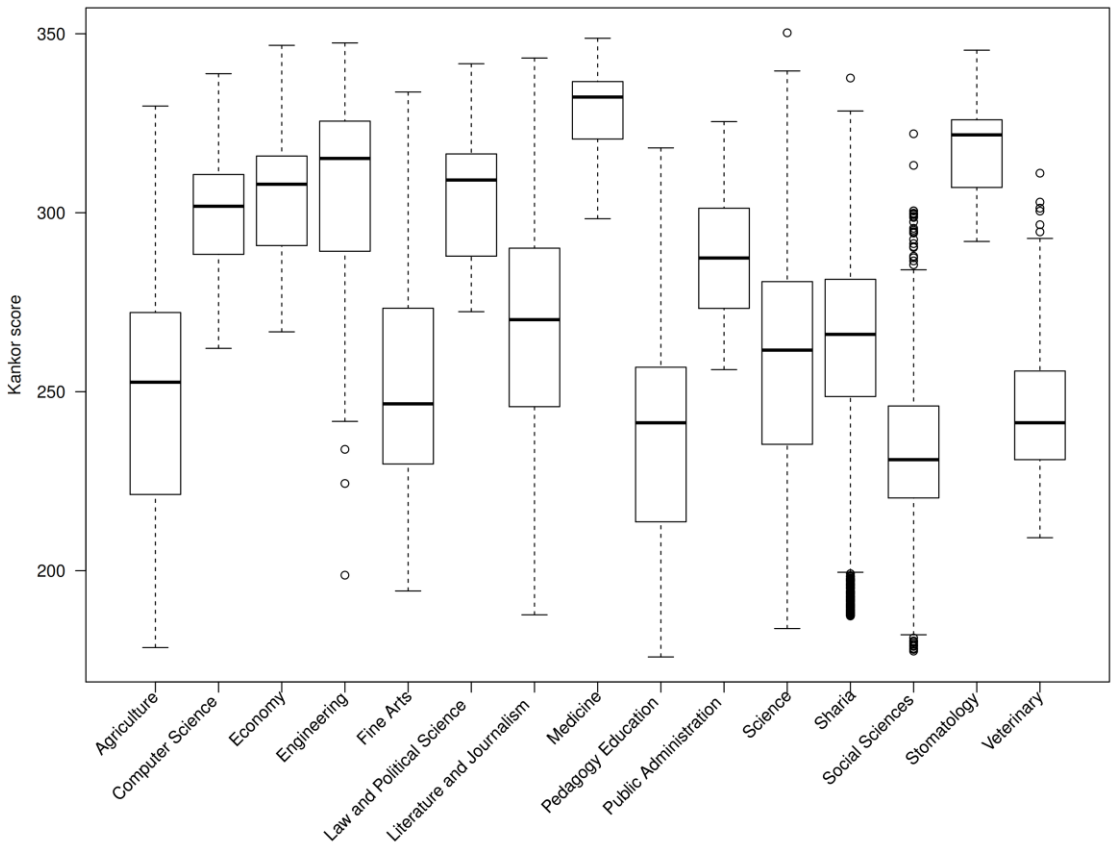

Figure 5-19. This boxplot shows Kankor admission scores for all the 15 fields of study at Herat University.

The Box Plot (also known as Whisker Plot) enables one to have a closer look at the data. It is a convenient way of graphically representing groups of numerical data through their quartiles. It shows the outliers present in the data, the cluster of data points and the volume of data between two extremes. In general, it represents the following basic statistical information:

- the 1st, 2nd, 3rd, and 4th quartile,
- the mean,
- the median (it is good when there are extreme values in either direction i.e. outliers up or down),
- and range of values.

Similar analyses were carried out to analyze the admission scores for Computer Science for 2014 and 2015, as demonstrated in **Figure 5-20**. The data for the years 2014 and 2015 were chosen for the following reasons:

1. Prior to the year 2014 a few of these public universities across Afghanistan did not offer Computer Science. But since 2014 all the 12 higher educational institutions offer Computer Science.
2. The outcome widely varies for the data only for the year 2014 or 2015.

University	Count	Min	Q1	Q2	Q3	Max	Mean	Median	Mode (Freq)	StdDev
Kabul University	419	260	293	307	324	345	306	307	298 (17)	18.62
Herat University	310	262	283	304	311	339	297	304	307 (14)	18.16
Polytechnic University	360	254	279	295	309	333	295	295	309 (15)	18.32
Kabul Education University	240	255	265	288	299	322	283	288	300 (13)	17.61
Sheikh Zayed University	290	169	211	278	308	324	259	278	304 (23)	52.92
Kandahar University	300	191	254	273	283	323	267	273	273 (16)	26.08
Kunar University	301	214	223	272	284	322	257	272	222 & 217 (12)	31.82
Nangarhar University	359	174	241	269	303	324	267	269	304 (12)	40.32
Balkh University	190	208	223	255	339	347	275	255	343 (12)	54.36
Parwan University	150	213	225	247	310	341	262	247	309 (9)	40.73
Badakhshan University	140	187	206	239	296	329	250	239	292 (9)	46.18
Kunduz University	249	182	198	223	305	330	247	223	197 & 193 (8)	51.52

Figure 5-20. This matrix shows Kankor admission scores for Computer Science across the country.

The above matrix, **Figure 5-20,** reveals that the admission scores for Computer Science at Kabul, Herat, Polytechnic, and Kabul Education universities are close and these universities can be considered as one cluster. Similarly, the admission scores for Shaikh Zayed, Kandahar, Kunar, and Nangarhar universities are relatively in the same range and these universities can be classified into another cluster. For further clarification, these data were visualized using box plots of the Kankor admission scores vs. the public universities that offer Computer Science across the country, as illustrated in **Figure 5-21**.

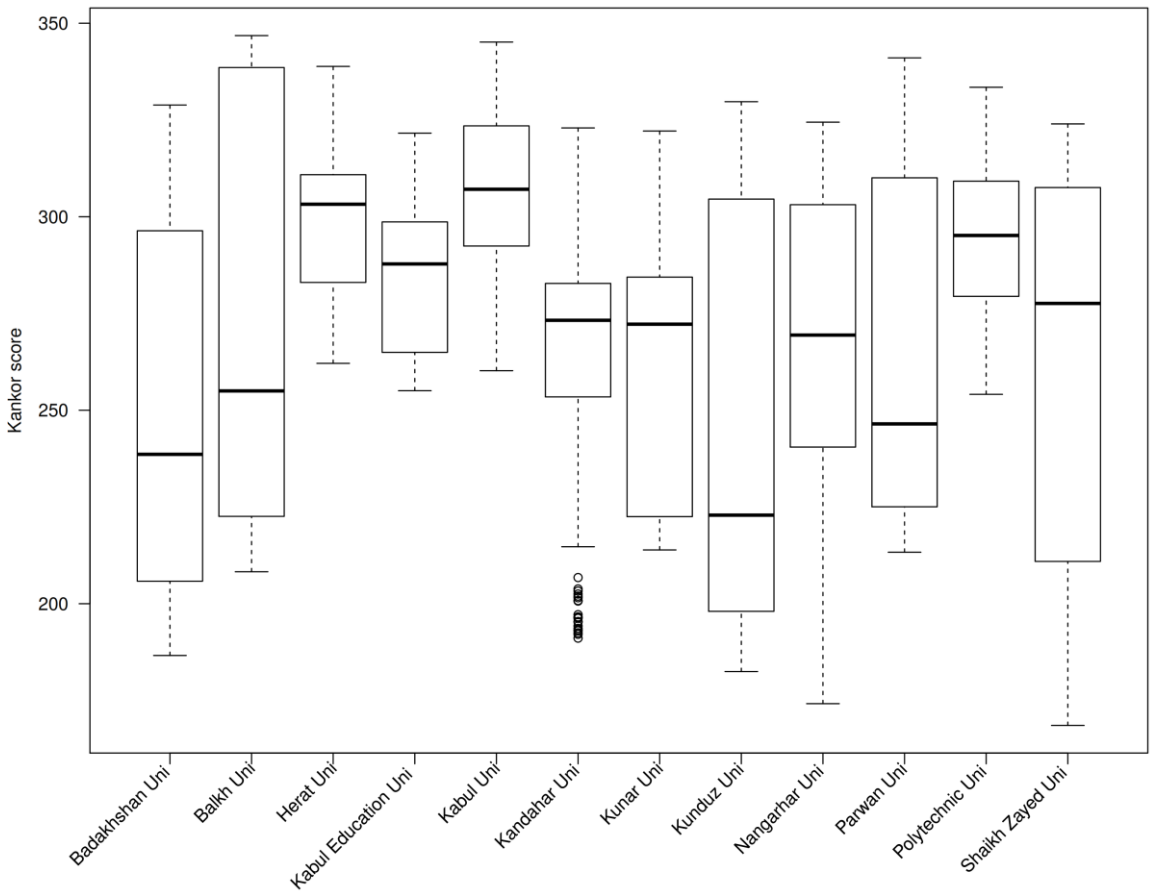

Figure 5-21. This boxplot shows Kankor admission score for Computer Science for all public universities.

5.9. Discussions and Recommendations

With the current ineffectual method of recording and producing data, accurate statistics and detailed insights are not produced. The data will not be valuable for effective research studies unless they are immensely pre-processed and cleansed. The author of this thesis proposes the following recommendations for storing and producing effective data as well as improving Kankor settings:

1. Attributes with predefined domain values should have proper tags/labels/identifiers and should be divided into their own separate attributes. This increases the opportunities to easily roll up and drill down data to generate accurate and detailed reports and insights: For example, the attribute 'Candidate's Result' should be divided into the following attributes: field of study, higher educational institution, type of educational institution, geographical location of educational institution, etc.
2. A better approach would be to categorize the more than hundred fields of study into the following main streams: Natural Sciences, Social Sciences, Health Sciences, Literature and Humanities, Islamic Education, Technical and Vocational Education, etc. Also, ensure that the same system and the same terminology are used across all provinces. Currently, in some provinces, the following fields of study: Languages, Mathematics, Physics, Biology, Chemistry, History, Geography, Psychology, and Philosophy fall under 'Education'. In some other provinces, they fall under Sciences and Social Sciences. While in some other provinces some of the mentioned fields of

study are taught under faculties of Humanities and Sciences. With such a mechanism, unified and transparent reports could not be generated easily.

3. Storing and producing detailed data is essential: For example, if the candidates' choices of fields of study and the order of choices are recorded and produced, then the popular fields of study among the candidates could be discovered precisely. Also, assessment could be made if the candidates' choices are random. As an example, Theology, Fine Arts, Medicine and Engineering are not systematic since they are not relevant and do not fall under one stream or similar streams. Additionally, it is recommended to store and produce score details i.e. scores for mathematics, natural sciences, social sciences, and languages; the number of available seats for each field of study in every higher education institutions, and other valuable data.
4. The following structure with the following attributes is proposed by the author of this thesis:
    - Kankor year,
    - candidate's Kankor ID,
    - candidate's personal information (first name, last name, father name),
    - gender,
    - cell phone number,
    - candidate's current province,
    - candidate's permanent province,
    - candidate's high school,
    - high school location (province),
    - high school ownership (public or private),
    - high school geographic location (urban or rural),
    - fields of study (1-5) the candidate has chosen,
    - department,
    - field of study the candidate has been admitted into,
    - department,
    - main streams (Natural Sciences, Social Sciences, Health Sciences, Literature and Humanities, and so on),
    - educational institution,
    - educational institution type (public university, private university, teacher training, Islamic education, technical and vocational institutions),
    - educational institution's geographical location,
    - educational institutions' available seats,
    - score for each category of questions,
    - total score,
    - result (Pass, Fail, Fraud, Result-less, and other identifiers),
    - and other useful and valuable data.
5. The author assumes Kankor Committee of MoHE, through its Kankor Management System, is capable of storing and producing the data as mentioned. If so, unfortunately, the data is not recorded properly and never shared in detail for research studies – Of course, the candidate's confidential data should be excluded before

sharing with other parties. If the Kankor Management System lacks the feature to store and produce data in detail and in a proper format, then it is highly recommended that they should seriously consider including that in the next version.

Presently Kankor is not able to systematically and methodologically identify skills and abilities of candidates. This chapter in this context concludes the following recommendations:

1. It is reasonable to systematically revise the Kankor scoring system based on candidates' choice of field of study: For example, fields of study relevant to Sciences require strong understanding and knowledge of mathematics and science subjects. While Theology and similar fields of study do not. This will be a pioneering step leading toward offering specialized Kankor for the main streams.
2. Choosing 5 favorite fields of study out of more than hundred fields of study is very difficult while participants do not have sufficient knowledge of them. A better approach would be to categorize them into the following general main streams: Natural Sciences, Social Sciences, Health Sciences, Literature and Humanities, Islamic Education, Technical and Vocational Education, etc.
3. When Kankor results are announced, the MoHE website becomes inaccessible due to traffic loads. Kankor candidates and their families are impatient to find out about the results. They call friends that have access to the Internet, or they go to internet clubs for hours in order to find out their results. This is a time-and money-consuming process. With the latest statistics announced, more than 90% of Afghans have access to GSM services via mobile phones. If MoHE introduces a policy to store cell phone numbers of Kankor candidates, which they already do, which ensures privacy, then Kankor result notification could be sent to candidates' cell phone numbers. This also ensures privacy for Kankor participants as compared to sharing Kankor results in excel sheets, which include all the demographic and other personal information of candidates, on social media and putting participants at risk.

5.10. SUMMARY

Descriptive analyses reveal that available data is not sufficient to generate accurate and in-depth insights unless they are immensely pre-processed. The author of this thesis proposed a proper structure to record and produce detailed data of Kankor participants as a solution, which greatly improves the existing structural models.

The current Kankor structure is found not to be the proper medium to identify the level and extent of candidates' competence and skills. For example, according to the author's observation and surveys, since Kankor lacks questions to identify Computer and English literacy, most of the candidates that get admitted into Computer Science do not have enough knowledge of basics of Computer and English language which the cornerstones for the field of Computer Science are. More details can be found in *Chapter 2 under section Education System in Afghanistan*.

Also, analyses illustrate that admission is strongly dependent on the number of available slots at educational institutions. A candidate can perform well in Kankor and get a high score

but still not get into his/her favorite field of study because of the order in which s/he has chosen fields of study. Therefore, a candidate's uninformed choice of field of study might lead to no admission.

Furthermore, statistics reveal high schools lack a standard and realistic scoring system; and, based on the author's conclusion and observation the social and political conditions are not favorable. Hence, high school marks do not play a fundamental role in the success of candidates in Kankor. As a result, direct admission into higher education institutions based on candidates' high school GPA is not recommended. On the other hand, elimination of Kankor may be extremely challenging in the short-term. To transition gradually, Specialized Kankor may serve as an intermediate bridge to enhance the existing admission processes. A Specialized Kankor may allow allocation of human capital into appropriate fields of study, leading to a more specialized labor force in the long-term. Results of the questionnaires for university and high school administrators also show that everyone believes that Specialized Kankor is a better means to identify the skills and competence of candidates. With the current situation, the author sees that offering specialized Kankor for every one of the existing fields of study may be difficult, in which case fields of study could be categorized into main streams proposed in this chapter. Specialized Kankor can then be held for each of the main streams.

Descriptive analytics are useful for data exploration, visualization, and producing insights while predictive analytics are useful to guess/predict the future with the help of data mining methods and techniques (Larose and Larose 2015). In the next chapter, the author, besides testing and validating the predictive models using data mining approaches, also uses descriptive analytics to confirm the outcomes and accuracy of predictive analytics.

Chapter 6

Chapter 6. Proposed Tools, Processes and Approaches in Modelling Decision Making

As previously mentioned in chapter I, in Afghanistan policymakers and high school advisors do not have the necessary tools such as data mining techniques, recommender system methods, and other proposed tools to carry out assessments and provide academic career guidance methodologically, using the advantages of science and technology. Additionally, the education system needs to be overhauled to accommodate structural reforms related to institutionalization of academic counseling, which is a part of what this study addresses.

The following different techniques are suggested to support policymakers in shaping the education system and to allow high school advisors to be able to provide guidance to those candidates seeking career advice.

- Narrowing down choices.
- Kankor Practice Test (e-Kankor).
- Prediction model using data mining approaches.
- Prediction model using recommender systems approaches.

The followings are the main reasons the author of this thesis proposed different techniques rather than proposing and relying on only one:

- In the chapter 2 of this dissertation, it was explained that the education system in most countries allows high school graduates to change fields of study several times. However, in Afghanistan high school graduates are given only two chances to attend the Kankor exam in a life time. Therefore, guidance on academic career, mainly, choosing a field of study is very critical and requires more precision. It is therefore considered that help from a third party would be vital to better support a more informed decision making.
- Since high school studies in Afghanistan are not specialized as it is in Iran, Turkey and other countries, and there are no pre-college courses prior to entering higher education, it is deemed efficient to evaluate the skills of applicants in the 12th grade through a specific set-of Kankor practice tests. The surveys and interviews conducted by the author of this thesis also confirm that a suitable option is to offer several Kankor practice tests to high school students, specially grade 12 students. This allows students to get familiar with the real Kankor test through Exam Cram. The purpose of e-Kankor in the context of data mining is to estimate the candidates' Kankor score through the set of Kankor practice tests, then to use the candidates' data and the estimated scores as unseen data (test data) for the data mining predictive model to provide suggestions to the candidates prior to their actual Kankor examination.
- Furthermore, the collected Kankor data lacks some important features that might be useful in building the prediction models with higher and better accuracy. The Kankor

data does not include features such as the five choices the candidates chose in the Kankor as well as the order of their choices, the number of available slots for educational institutions, and detailed Kankor score for each category of questions i.e. scores for mathematics, sciences, social sciences and languages. Likewise, the Kankor data does not show types of high schools and whether they are public or private, or if the high schools are geographically located in urban or rural areas. Therefore, in addition to the prediction model, it is deemed efficient to diversify a modality of alternatives.

- Moreover, a closer analysis of the Kankor data reveals that there are more than a hundred fields of study, and these fields of study are not classified into streams such as sciences and social sciences. Building a prediction model with more than a hundred fields of study often does not lead to a good model. Therefore, the author of this thesis proposed approaches to narrow down these choices properly prior to building the prediction model supported by a scientifically structured framework.
- Finally, as mentioned in chapter 2 of this dissertation, in Afghanistan the former Minister of MoHE and law makers have discussed and suggested to eliminate Kankor as an entrance medium to higher education and consider high school grades/marks in lieu. In some countries high school grades are considered as an entrance medium to higher education. Therefore, the high school marks have been taken into consideration as part of the data modelling for the proposed solution. It is, however, worth mentioning that analyses carried out by the author of this thesis in the previous chapter validates that high school marks alone are not trustworthy as a standalone metric to rely on.

The proposal stands unique in its nature in the sense that in Afghanistan no scholars have investigated this issue at the time of research and the literature review shows a lack of similar research attempts to combine the techniques federated in this study.

It is worth mentioning that the author's research differs from similar research conducted in other countries in three important ways:

1. The studies mentioned in chapter 3 of this dissertation dealt with simple cases and scenarios i.e. two to three class values while in Afghanistan we have a complicated system of over 100 choices for fields of study.
2. The school students in those countries are offered specialized studies at secondary or high school level.
3. Academic advice is available to students either in secondary and high schools or through other offices in the country, which guide the students concerning their academic career.

But in the context of Afghanistan none of the above is true. Therefore, it requires more attention and needs more research.

6.1. Narrowing Down Choices

There are more than a hundred fields of study programs, which are not streamlined distinctively, and candidates are required to choose a maximum of five and a minimum of

one field of study, usually creating a dilemma of choices for those candidates unsure of their decision and their ability to succeed.

Typically, families and the candidates are interested in the most common fields of study such as medical science and engineering because these professions are very popular among the community, and the families and applicants are familiar with the names of these professions.

This research introduces two approaches to substantiate the efficacy and optimization of streamlining fields of study and narrowing down the choices for the candidate to the maximum possible extent through a clearly pre-defined set of rules. The first instance silos irrelevant subject matters for the sole purpose of bringing the candidate closest to his/her choice and favorite institutions and provinces, whereas the second instance narrows down the choices such that it would make the study programs manageable and consistent with their *knowledge set* and their tendency.

The followings are examples of corresponding scenarios for the first instance:

- Most Kankor candidates know in advance in which university or province they would like to continue their higher education studies. Hence, taking this preference of the candidates into account, the available fields of study can be reduced to the existing of fields of study in that province or university. More information on this in the following section of this chapter.
- Also, almost all the girl candidates prefer either to study in their home province or in a province with a proper girls' dormitory. This is another method, which can be used to reduce the choices of fields of study province-wise. More information in the following section of this chapter.
- There are those candidates for whom the province or the university is not a factor affecting their decisions, but instead they prefer to know which institution offers the most relevant field of study for them. That an institution is located somewhere other than in their home province does not matter for this category of applicants. Thus, their choices will only be limited to those institutions which offer the study program they want. For instance, *candidate x* is interested in *field of study y*. All the fields of study are narrowed down to the *field of study y* for the *candidate* x independent of the geographical location of the institution. More information on this in the following section of this chapter.

Moreover, the *knowledge set and true tendency* of the Kankor candidates also play a decisive role to suggest them relevant fields of study. Identifying the *knowledge set* of the candidates allows to narrow down the choices by eliminating unnecessary fields of study. Presently, families and high school teachers are not equipped with such an assessment method to take an interest inventory as described in the following section.

A special test comprising non-technical 'yes/no' questions will be designed to identify *knowledge set and the true tendency* of the applicant for a desired field of study. The questionnaire includes 150 questions in order to find out and streamline choices for an assumed number of 10 fields of study, with 15 questions per field. Since there are more than a hundred fields of study, the design proposes classifying the fields of study into categories

to reduce the analysis workload. In some countries similar methods are used in some counseling centers as well as in some education and higher education institutions to provide academic and career advice for students. Based on the interview with officials of the Afghan Turk high school in Herat, this high school also uses a similar technique to provide guidance to their students concerning their academic career. It, however, includes only a few disciplines and so is not robust and rich and therefore expert feedback and contribution at the national level is needed to adjust it in accordance with the Afghanistan education and higher education systems.

Below is an example of the table of questions (10 out of 15) to identify those candidates whose interest falls under *Computer Science* field of study:

No	Questions
1	Are you interested in learning about technological innovations?
2	Do you develop any type of apps or programs that would require coding?
3	Do you easily consider yourself an individual to solve numerical problems?
4	Do you perform any activities related to architectural drawing, such as map of buildings, apartments, houses, maps, and directions using computer programs?
5	Do you think technology has the potential to provide a solution to all types of humankind's problems?
6	Do you always plan and document for your activities before implementing?
7	Do you carry out any accounting and or data analysis tasks in a company?
8	Do you have access to technical and scientific magazines and journals related to your career interests?
9	Do you easily get fascinated with discovery and innovations?
10	Do you examine how electronic devices and circuits work?

Table 6-1. Example of questions used to determine the tendency of applicants towards Computer Science.

It is worth mentioning that these questions are only at drafting stage and may be replaced with standardized questions in consultation with professional test developers. Likewise, the total number of questions and scoring mechanism for each stream or each field of study will be determined by these experts. The questions need to be carefully designed in order to find out the baseline of knowledge, creativity and interest of the test takers in relation to the fields of study. For instance, in the above table, questions #2 and #4 are intended to realize whether or not the test takers have basic knowledge in computer studies. Likewise, questions #3 and #7 are purposed to understand whether or not the test takers have basic knowledge of math. Similarly, questions #1 and #8 find out if test takers are interested in IT and Computer Science and finally questions #6, #9 and #10 are bundled to determine creativity of the test takers.

6.2. KANKOR PRACTICE TEST (E-KANKOR)

Let's assume the prediction model is developed and tested based on previous Kankor datasets and previous high school marks. For the model to make predictions for the Kankor candidates, it requires the candidate's biodata, high school marks, and Kankor score as input.

Then the prediction model evaluates the given data and accordingly provides suggestions to the candidate which are closely relevant to her/his profile.

The high school marks for the Kankor candidates are available prior to registering for the Kankor exam. However, the Kankor performance and score for the Kankor candidates are not available prior to taking the actual Kankor exam. But the main purpose is to provide suggestions to the Kankor candidates before they take the actual *Kankor* exam and to measure the candidate's strengths and weaknesses to assist them in gauging their odds in the actual examination.

The approach is to design a system to simulate the real Kankor exam considering the Kankor Committee of MoHE standards and regulations. The following conditions show why a Kankor simulation is needed:

- A few of private high schools and other educational institutions offer only paper-based *Kankor* practice tests to increase the candidates' awareness and to make them familiar with actual *Kankor* exams through exam cram methods.
- Also, there are a few private high schools where fragmented efforts are applied to streamline students' abilities based on *Kankor* practice tests in order to provide students with best possible recommendations.
- The surveys and interviews conducted by the author of this thesis also confirm that a suitable option is to offer several Kankor practice tests to high school students during the preparing academic year.

To this end, an integrated platform will be developed to enable the candidates to enter their biographical data and high school marks. The system also will simulate the real Kankor exam (will not replace it) to enable the candidates to take the Kankor test.

The purpose of *Kankor* practice test in the context of data mining is to estimate the candidates Kankor score through a standardized set of Kankor practice tests, followed by using candidates' data and the estimated Kankor practice test scores as unseen data in the data mining predictive model in order to provide recommendations to the candidates. Of course, the recommendations should be elaborated and discussed with the tested candidate by high school advisors and experts.

Additionally, it also offers greater accessibility and features to prepare the Kankor candidates and increase their chances of admission into their favorite field of study when they take the actual Kankor. The preliminary steps suggested for the e-Kankor can be operated in both online and offline modes, require minimum computer literacy and are user-friendly. The following features are integrated into the e-Kankor system:

- Providing comprehensive guidance based on the latest rules and regulations of the Kankor Committee of MoHE,
- Identifying the strengths and weakness areas of the candidate (i.e. science vs. social science),
- Providing best estimates about the candidates' improvements with every test taken, by keeping a record of candidates' performance,

- Allowing the candidates to learn the importance of timeliness, methods and techniques through exam cram,
- Allowing the candidate to choose the area of study confidently,
- A histogram performance bar showing the positive or negative progress in real time during candidate's interaction with the platform,
- Finally, the purpose of e-Kankor in the context of data mining is to estimate the candidates Kankor score through the set of Kankor practice tests, then to use the candidates' data and the estimated scores as unseen data to the data mining predictive model to provide recommendations to the candidates.

This will be a new initiative in its kind, which will allow high school officials, higher education officials and local and national policy makers to better prepare students for the Kankor, while assessing all the choices a candidate has. Therefore, the e-Kankor will be an all-encompassing medium to self-evaluate, and to improve and minimize the identified gaps, as illustrated in the following figure.

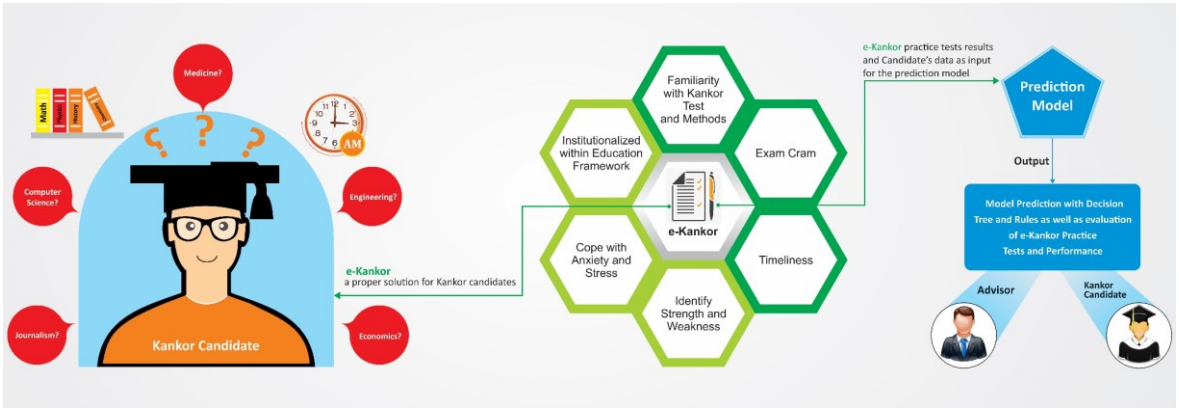

Figure 6-1. The e-Kankor for the Kankor candidates and data mining applications.

6.3. Prediction Model using Data Mining Approaches

Descriptive analytics are useful for data exploration, visualization, and producing insights, explained and detailed in the previous chapter while predictive analytics are useful to guest/predict the future with the help of data mining methods and techniques (Larose and Larose 2015).

The research author's aim is to assess the condition systematically using the advantages of science and technology. The approach is to build a prediction model and to generate representative rules supported data mining algorithms. The previous KRD datasets are used to build the prediction model. The candidates' data and e-Kankor practice tests results are used as input for the prediction model. The predictive model and the produced rules support high school advisors to suggest appropriate fields of study for the candidates by considering their performance and their choice of fields of study in e-Kankor, as is illustrated in the following figure.

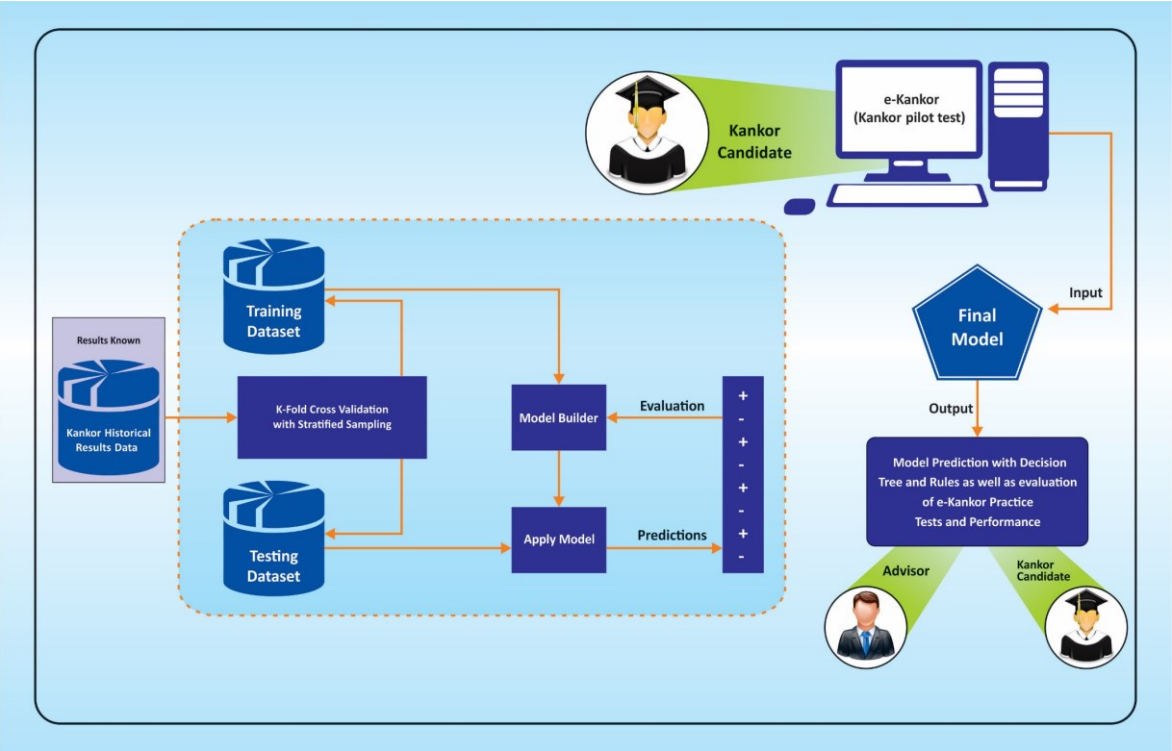

Figure 6-2. The prediction model to support advisors to recommend proper disciplines for Kankor candidates.

In this approach, the previous Kankor datasets collected, preprocessed and cleansed by the author of this thesis as explained in the previous chapters is considered to train and test the prediction model. To develop the model the Kankor datasets is required to be further preprocessed and divided into two sets, a training set is required to train the model, and a testing set where all class values are hidden from the trained-model to evaluate and test the accuracy of the model. If the model classifies most cases in the test set correctly, it is assumed that it will also work accurately on future unseen data. Otherwise, it is assumed as a wrong model.

Typically, designing a classification prediction model is an iterative process. To improve accuracy of the model, one will also need to take into consideration other algorithms with appropriate parameter settings as well as methods to divide the datasets into training and testing sets properly such as cross-validation with stratified sampling.

The author picked the following classification methods to build the prediction model, the motive for picking these methods will be briefly explained. However, this does not stand that the Kankor datasets are not experimented using other methods.

- **Decision Tree:** Decision Trees are one of the most widely used and practical methods in statistics, machine learning, and data mining. There are numerous distinct advantages to using decision trees in many classification applications: Decision trees can work on both categorical and continuous input and output variables. They are also capable of performing multi-class classification on a dataset. Above all, the output of the Decision trees is very easy to understand, read and interpret even for the layman. Its graphical representation is very intuitive, and the users can easily relate their hypothesis – that makes it possible to account for the reliability of the

model (Dahan et al. 2014; Kotu and Deshpande 2014a). Hence, in this thesis, the decision tree was picked to develop a model for prediction.

- **K-Nearest Neighbor:** The K-Nearest Neighbors (KNN) is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure such as, Euclidean Distance and Cosine Similarity (Larose and Larose 2015; Sayad n.d.; Kotu and Deshpande 2014a). In case of Kankor, the admission score overlaps for some fields of study, for instance, a candidate with a score of 290 in Kankor is eligible for Computer Science, Economics, Engineering, and possibly another field simultaneously. Hence, the KNN was selected as a methodology to develop the model with a proper value of K to compute the K nearest neighbor fields of study for the new applicant with an estimated score of 290, and in turn, assign a proper field of study by a majority vote of its neighbors.
- **Ensemble/Collective techniques:** The goal of ensemble techniques in data mining is to improve accuracy of the model through combining predictions from single algorithm such as, Bagging, Random Forest or combining predictions from several learning algorithms to obtain better predictive output than could be obtained from any of the constituent learning algorithms alone. Ensemble methods have been called the most influential development in data mining in the past decade. Fast algorithms such as decision trees, Random Forest, are commonly used in ensemble methods. Hence, the author also picked Random Forest to build the prediction model.
- **Random Forests:** Decision trees are prone to overfitting, which can give poor results when applied to future unseen data, causing a significant problem. Although there are several approaches and reasonable strategies handling overfitting in building decision trees such as, pre-pruning, and post-pruning; small changes in the data can drastically affect the structure of trees. This last point has been exploited to improve the performance of the trees through Random Forests. A Random Forests algorithm takes the decision tree concept further by producing many decision trees. It takes a random sample of the data and identifies a key set of features to grow and build each decision tree. Random Forests is a type of ensemble learning method, where a group of weak models combine to form a powerful model. The result from forest of trees is usually better than the result from one of the individual trees (Hartshorn 2016; Kotu and Deshpande 2014a). Hence, Random Forests was chosen to handle overfitting issues while building a model for prediction.

There are many ways to validate the trained-model such as Split Validation, Cross Validation, Leave One Out Cross Validation and Out of Time Validation. The following section briefly explains each one of them:

- **Split Validation:** Split validation also called Holdout validation the simplest validation method. It splits up the whole dataset into a training dataset and test dataset and evaluates the model on the test dataset (unseen dataset). The size of the training and testing datasets can be adjusted through parameter, but the common proportions are 70%/30% training/validation. This method works well with large datasets and it is not efficient for small datasets.

- **Cross Validation:** In Cross Validation the whole dataset is divided into k folds (a.k.a. subsets) of equal size. Out of the k folds, a fold is reserved as the test dataset, and the remaining k - 1 folds are used as training dataset. The Cross-Validation process is then iterated k times, with each of the k folds used exactly once as the test data. In the end, the k results (validation accuracy) from the k iterations are averaged to produce a final estimation, as it is shown in the below figure for further clarification. It is worth mentioning that the value k can be set through parameter. This is great and efficient for small datasets. However, it is computationally quite expensive on large datasets.

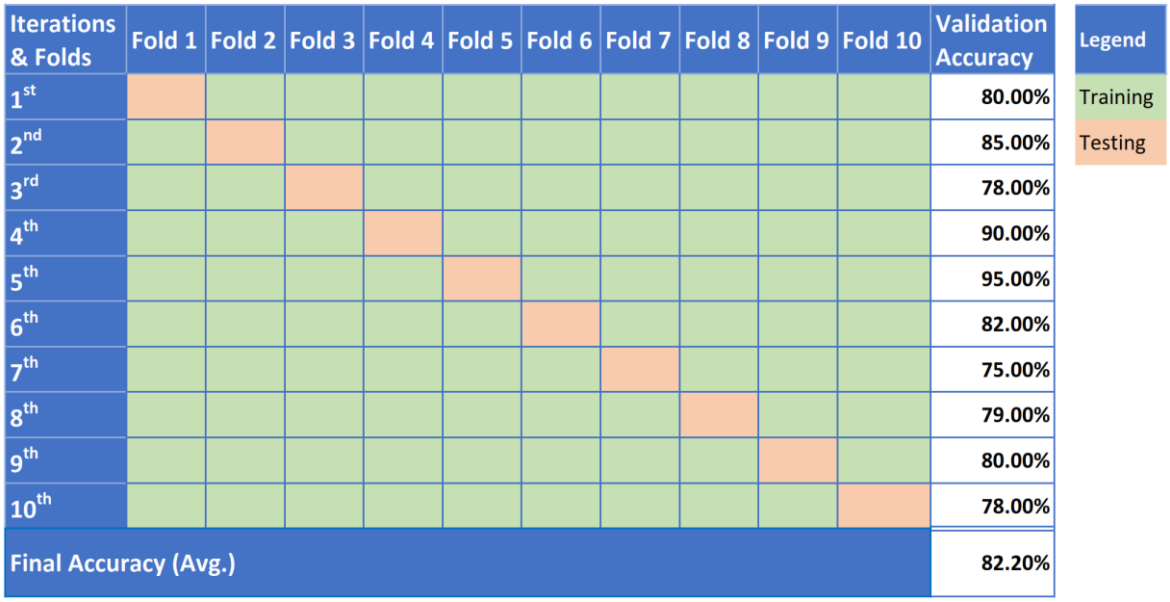

Figure 6-3. Cross Validation Workflows (Visualization by the author).

- **Leave One Out Cross Validation:** It is also called as "All but One Cross Validation". It is a special type of Cross Validation and works the same as the Cross Validation. However, it considers only one sample from the whole dataset for testing, and the remaining samples are used as the training data. This process is iterated such that each sample in the whole dataset is used once as the test data. In this form of validation k is equal to n where n is the total number of samples in the whole dataset. Therefore, it is iterated n times. It is computationally very expensive, as the training and testing subprocesses are repeated as many times as the number of available samples in the whole dataset. This is great and efficient for very small datasets.

Cross Validation and Split Validation with stratified sampling were used to split the dataset into the training and testing sets. Where the stratified sampling ensures that the class distribution in the subsets is the same as in the whole dataset so that it addresses and reduces bias in the subsets (Larose and Larose 2015; Kotu and Deshpande 2014a).

The author collected the Kankor data since the year 2003 to 2015 including of around 1.5 million records. All the collected data were preprocessed, transformed and cleansed as well as were used for exploration and visualization purposes in the previous chapters. In this section, the author of this thesis chose Kankor data for the years 2013, 2014 and 2015 altogether (henceforth, KRD1315) with 671,513 samples (or examples, records, instances)

of Kankor participants for building the prediction models. Furthermore, the Kankor data only for the year 2015 (henceforth, KRD15) with 181,798 samples of Kankor participants were used to train the same prediction models for comparison purpose. It is because across the years the minimum admission score for semi and higher educational institutions changes.

The techniques used are classification with three algorithms explained above in this section, namely, Decision Tree, Random Forests, and K-Nearest Neighbors (K-NN). To overcome overfitting and reduce noise Ensemble techniques (Bagging, Boosting, and Stacking) were also used to train and build the models. For each model a confusion matrix was obtained to calculate sensitivity, specificity, and accuracy.

It is worth mentioning that RapidMiner Studio[6], a leading open-source system for knowledge discovery and data mining, picked by the author for experiments to build the models and to evaluate the models' accuracy.

There are more than a hundred fields of study, and these fields are not classified into major streams such as sciences, social sciences, and other relevant streams. It is very complex to develop a classification model with more than a hundred values/fields of study in the target attribute. However, clustering techniques from data mining and techniques from descriptive analytics can be used to classify the fields of study to reduce the number of possible values for the target attribute prior to designing the classification prediction model. The following sections explains several approaches the author of this thesis carried out to build prediction models for difference cases and scenarios to help policy makers in shaping the education system and allow high school advisors help the Kankor candidates on their academic career of choosing a field of study in the Kankor examination.

6.3.1. Binary and Multiclass Classification

In this section, the following predictor (a.k.a. independent) variables (gender, province, and Kankor score) are selected as input, and to provide feedback and prediction in the high level the (candidate's result) is selected as class (a.k.a. target or dependent variable). After preprocessing and transformation steps carried out, the candidate's result contains the following four values: Public University (Pub. Uni), Private University (Pri. Uni), Technical and Vocational Institute (TVI), and Fail, as illustrated in the following sample dataset.

#	Gender	Province	Kankor Score	Kankor Result
1	Male	Herat	290	Public University
2	Female	Kabul	190	Private University
3	Male	Herat	150	Fail
4	Female	Herat	250	Technical and Vocational Institute

Table 6-2: A sample of Kankor data for Binary and Multiclass classification.

[6] RapidMiner Studio is a powerful visual programming environment for rapidly building complete predictive analytic workflows.

6.3.1.1. Binary Classification

Binary classification is the task of classifying samples in the dataset into two groups such as 'pass or fail', 'true or false', 'positive or negative'. Therefore, all the samples of the KRD1315 dataset are classified into 'Pass or Fail' class values.

Then, the data mining workflow process was created in RapidMiner Studio to develop a prediction model using the following classifiers: Decision Tree, Random Forests and KNN, explained above. Cross Validation method with 10 number of folds and stratified sampling type was used to estimate how accurately the model (learned by the mentioned learning classifiers) would perform in practice, as shown in the following figure.

The accuracy, precision, and recall of Decision Tree, Random Forests, and KNN (K=15) were nearly 100% for the KRD15 dataset. The tree generated by the Decision Tree is very simple with the following two straightforward rules:

1. The candidates with a Kankor score greater than 150 were considered as Pass,
2. else the candidates were considered as Fail.

However, the model's accuracy for the KRD1315 decreased to 87.75%, 85.11% and 82,41% for KNN (K=9), Decision Tree and Random Forests classifiers respectively. The Decision Tree model infers the following rules:

- Almost all the candidates with a Kankor score greater than 177 were considered as Pass,
- More than half of the candidates with a Kankor score less than or equal to 177 and greater than 150 were considered as Fail,
- All the candidates with a Kankor score less than or equal 150 were considered as Fail.

The following confusion matrix elaborates the above rules that were produced by the Decision Tree as well as estimates how accurately the model (learned by the Decision Tree classifiers) performs in practice on the test sets produced by the Cross Validation.

Accuracy: 85.75% +/- 0.10% (mikro: 85.75%)			
	true Pass	**true Fail**	**class precision**
pred. Pass	340,452	10,293	97.07%
pred. Fail	85,392	235,376	73.38%
class recall	**79.95%**	**95.81%**	

Table 6-3. The confusion of matrix of the Decision Tree classifier for the KRD1315 with two class values.

From the above confusion matrix, it can be concluded that out of 425,844 Kankor participants who successfully passed the Kankor exam 85,392 of them were predicted Fail by the classifier model. While on the other hand, the classifier model predicted 10,293 of candidates as Pass out of 245,669 candidates who were failed in the Kankor. Comparison of the mode trained on the KRD1315 with the model trained on the KRD15 concluded that one of the main reasons is the minimum admission Kankor score to higher education in the year

2015. Additionally, it can be concluded that building a prediction model on an average of the recent three years of Kankor dataset is efficient than building a model on only the recent year dataset as the admission score to higher education changes in time periods.

6.3.1.2. Multiclass Classification

Multiclass classification is the problem of classifying all the samples in the dataset into more than two classes, for example, to classify the Kankor participants' admission into Pub. Uni, Pri. Uni, TVI, and Fail. Multiclass classification assumes that each sample is assigned to one and only one class, for example, a Kankor participant admission is either to Pub. Uni or Pri. Uni but not both at the same time. According to the above definitions and principles, all the samples of the KRD1315 dataset were classified into the following three class values: Pub. Uni, Pri. Uni & TVI and Fail.

Next the workflow process was created in RapidMiner Studio with same settings as the Binary classification approach explained above to develop a prediction model. The model's accuracy for the KRD15 dataset turns out to be 81.67% for Decision Tree, 81.21% for Random Forest, and 87.07% for KNN (K=15) respectively.

The tree generated by Decision Tree infers the following rules:

- If the Kankor score > 178 then the candidates most likely the candidates got admitted to public universities (Pub. Uni),
- else if the Kankor score > 150 and the Kankor score ≤ 178 then most likely the candidate got admitted to either private universities or technical and vocational institutions (Pri. Uni & TVI),
- finally, if the Kankor score >= 150 then the candidates were considered Fail.

The following confusion matrix further explains the above rules as well as estimates how accurately the built model performs in practice on the test sets formed using the Cross Validation.

Accuracy: 81.67%+/- 0.13% (mikro: 81.67%)				
	true Pub. Uni	**true Pri. Uni & TVI**	**true Fail**	**class precision**
pred. Pub. Uni	46,610	23,199	0	66.77%
pred. Pri. Uni & TVI	10,118	76,731	0	88.35%
pred. Fail	0	0	25,140	100.00%
class recall	**82.16%**	**76.78%**	**100.00%**	

Table 6-4. The confusion of matrix of the Decision Tree classifier for the KRD15 with three class values.

From the above confusion matrix, it can be concluded that out of 56,728 participants in Pub. Uni almost 1/4 were classified to Pri. Uni & TVI by the classifier model. While on the other hand, 23,199 participants from Pri. Uni & TVI were classified to Pub. Uni by the classifier model. Uni. Finally, all the 25,140 candidates who failed in the Kankor exam were considered failed by the model.

The following table is the confusion matrix for the KNN classifier. It seems that KNN's *precision* for the class Pub. Uni better than the Decision Tree's. while for the same class the Decision Tree's *recall* is better in comparison to the KNN's.

Accuracy: 81.67%+/- 0.13% (mikro: 81.67%)				
	true Pub. Uni	**true Pri. Uni & TVI**	**true Fail**	**class precision**
pred. Pub. Uni	44,002	10,707	0	80.43%
pred. Pri. Uni & TVI	12,726	89,191	45	87.47%
pred. Fail	0	32	25,095	99.87%
class recall	**77.57%**	**89.25%**	**99.82%**	

Table 6-5. The confusion of matrix of the KNN classifier for the KRD15 with three class values.

Another similar experiment was conducted in which all the samples of the KRD15 dataset were classified into the following four values (Pub. Uni, Pri. Uni, TVI, and Fail). The model's accuracy decreased to 77.32% and 72.48% for Decision Tree and Random Forests with Information Gain Ratio (IGR) as criterion respectively. Finally, for KNN (K=15) an accuracy 77.35% was achieved. The Decision Tree model infers the following main rules:

- if the Kankor > 198 then the candidates mainly got admitted to Pub. Uni,
- else if the Kankor score ≤ 198 and the Kankor score > 169 then the candidates most probably got admitted to Pub. Uni,
- else if the Kankor score ≤ 198 and the Kankor score ≤ 169 then the number candidates got admitted to Pub. Uni were more than the candidates got admitted to Pri. Uni or TVI,
- and finally, if the Kankor score ≤ 150 then the candidates were confidently considered Fail.

The descriptive analytics in the chapter of this dissertation infer allocation to higher education institutions is limited to the number of available seats, and the participants' first choice of field of study. It is 'first come, first served'. Hence, it cannot be said that the model's prediction accuracy is low.

Furthermore, the same classifiers with the same settings as above were used to develop a model for the KRD1315 and 72.03%, 72.03% and 76.02% accuracy were achieved for the Decision Tree using Gini Index (GI) as the criterion on which attributes will be selected for splitting, Random Forest and KNN (K=9) classifiers respectively. The Decision Tree model infers the following basic rules:

- Nearly all the candidates with a Kankor score greater than 240 were classified to Pub. Uni,
- Most of the candidates with a Kankor score less than or equal to 240 and greater than 176 were classified to Pri. Uni & TVI, some of the candidates with a Kankor score in this range were classified to Pub. Uni and a few of the candidates were considered failed,

- Most of the candidates with a Kankor score less than or equal to 176 were considered failed.

This can be concluded that the candidates with a Kankor score greater than 240 certainly get admission to public universities and the candidates with a Kankor score in the range 176 and 240 get admission either to private universities or technical and vocational institutions and finally the candidates with a Kankor score less than 176 cannot get admission to either of ones.

In this section, the candidate's results were categorized with basic and possible limited values such as "Pass or Fail", "Pub. Uni, Pri. Uni & TVI and Fail" or "Pub. Uni, Pri. Uni, TVI and Fail". The developed models generated straightforward rules, which are useful for high school advisors to provide feedback to Kankor candidates. But these basic rules are not enough. In next section, the author's attempt is to generate other interesting and representative patterns and rules to support the high school advisors to help the candidates better.

IG and IGR have been applied on the KRD1315 dataset to measure the association between independent variables (gender, province, high school, Kankor score, fields of study, higher education institutions, and location of higher education institutions) and the dependent variable the (candidate's result) as class attribute. The result reveals the following variables (or attributes, features, dimensions): fields of study, higher education institutions, and location of higher education institutions also plays a strong role. The author also applied other feature selection techniques such as Optimization Selection to discover more about the contribution and importance of attributes. This also shows a similar result to IG and IGR.

The big challenge is that there are more than a hundred fields of study, many higher education institutions, and 34 provinces. Moreover, the admission score varies for some fields of study from province to province and from one higher education institutions to another and overlaps for some others. Including these features with all the possible values would lead to a complex tree comprising many rules what are not reliable, and (of course with a low accuracy rate of the developed model). To address this challenge and reduce the complexity, the author of this thesis inspects different scenarios and cases which are both practical and realistic, explained in the following sections.

6.3.2. Classification of Similar Fields of Study in Public Universities across the Country

As noted at the beginning of this chapter, there are more than a hundred fields of study which were not classified into streams such as sciences and social sciences and so on. The interviews and surveys conducted by the author of this thesis validates that in such a structure it is very challenging for the candidates to choose a maximum of five fields of study from among more than a hundred existing disciplines. The candidates have partial understanding about a few disciplines that are very common but are unfamiliar with most disciplines.

Furthermore, using more than a hundred fields of study as target values to develop a prediction model that can suggest a suitable field of study to the applicants is challenging and perhaps in certain cases even impossible due the following reasons:

- The Kankor data currently does not include variables such as the five disciplines selected by the applicants and lacks the information on the order they were chosen. There is no information on whether the applicant is a graduate of a public or a private high school and relevant data on the high school geography is not recorded by MoHE. Additionally, the data does not include a variable to show the breakdown of Kankor scores for various subjects such as mathematics, sciences, social sciences and languages as well as a variable to determine the seating capacity of higher education institutions. Availability of such key variables could make constructing the prediction model intuitive and illustrative.
- There are multiple fields of study whose admission scores are the same. For example, Herat University is comprised of 15 faculties and some of its faculties have several departments. Some candidates with a score of for instance 280 or 290 can get admitted to Engineering Faculty, while candidates with the same score can also get admitted to the Computer Science, Economics or into Law and Politics faculties. Therefore, it becomes a challenge for the prediction model to classify the candidates since the same score can allow candidates to qualify for more than one field of study.
- Some universities across the country offer the same fields of study. For instance, disciplines such as Engineering, Economics, Agriculture, Law and Politics, Computer Science, Fine Arts, Medicine and others are offered by several universities in Afghanistan. Analysis of the Kankor data shows that the admission score for the same fields of study in different universities are the same. This makes the design and development of the prediction model more challenging.

One of the solutions proposed by the author of this thesis at the beginning of this chapter is to systematically narrow down the existing number of fields of study.

As previously mentioned for some Kankor candidates the geographical location of the university does not impact their decisions. For example, some candidates decide to study Medicine either at Herat University, Kabul University or any other universities that offer Medicine. In such a scenario, the key deciding factor is the availability of the field of study regardless of the geography. For such type of candidates, the proposed solution is to identify the universities that offer fields of study they are interested in. Let's assume Computer Science is one of the disciplines this category of candidates is interested to pursue. In Afghanistan, there are twelve universities that offer Computer Science. Considering the field of study from a geographic pool reduces the total number of fields to twelve from among more than a hundred existing disciplines.

As stated above, most candidates for instance with a score of 280 or 290 were admitted to Computer Science faculties of Kabul, Herat and Polytechnic Universities. In such a case, building a prediction model with the current independent variables and the twelve identified universities as target variable to classify the candidates into Computer Science discipline is not realistic. It would be ideal to first classify these twelve universities into clusters using a combination of statistical and empirical techniques and then use these clusters as target variable to build the prediction model.

In order to validate the above argument, the KRD1315 dataset was filtered to include only Computer Science as field of study for the year 2015. Computer Science was picked since the author has a Computer Science background and is familiar with this field of study across the country. The KRD1315 dataset was reduced to data on 1,799 participants who were admitted into Computer Science at 12 public universities across the country. The following features (gender, province, and Kankor score) were selected as input attributes and (public universities) with 12 class values as target attribute.

The Decision Tree with IG, IGR and GI as criterion, was developed and validated to classify the data and had an overall accuracy of 40%, 52%, and 47% respectively. Likewise, Random Forest and KNN (K=5) classifiers were used to build and validate models to classify the data which achieved an accuracy of 50% and 50% respectively. Similar results were achieved when the KRD1315 was filtered to include only Computer Science for the years 2014 and 2015 with 3,308 instances of Kankor participants.

In both cases, neither the accuracy of the models is good enough, nor the generated tree is representative. It is risky to provide suggestions to the candidates based on the generated rules. A better option is to classify the same fields of study across the country using either the descriptive statistics approach or the data mining clustering approach. The author of this thesis proposes the following two methods for classification of same fields of study that are offered by more than one universities across Afghanistan.

6.3.2.1. Method#1 - Descriptive Statistics Approach:

A close analysis of the Kankor data reveals that the minimum and the maximum admission scores are not reliable metrics for classifying existing universities into clusters, as explained in the previous chapter. That is because these two metrics alone do not represent the entire range of admission scores for any one field of study.

A close inspection of the Kankor data shows a few candidates may have received high scores in Kankor and been admitted into a field of study, while many other candidates with lower scores were also admitted into the same field of study at the same university. For example, the Kankor data for the years 2014 and 2015 reveals that, among twelve universities, Balkh University has the highest admission score of 347 for Computer Science. Moreover, out of 190 admitted candidates to Computer Science Faulty of Balkh University, fifteen had received a score of 340 that is greater than the maximum scores which were admitted into Computer Science at nearly any of the other eleven universities. Taking only the minimum and maximum admission scores into consideration would classify Balkh University into the first cluster which represents the top and prestigious universities. Likewise, the maximum and the ‘mode’ metrics would classify Computer Science Faculty of Parwan University as part of the first cluster. This scenario is illustrated in the following figure.

University	Count	Min	Max	Mode (Freq)
Balkh University	190	208	347	343 (12)
Kabul University	419	260	345	298 (17)
Parwan University	150	213	341	309 (9)
Herat University	310	262	339	307 (14)
Polytechnic University	360	254	333	309 (15)
Kunduz University	249	182	330	197 & 193 (8)
Badakhshan University	140	187	329	292 (9)
Sheikh Zayed University	290	169	324	304 (23)
Nangarhar University	359	174	324	304 (12)
Kandahar University	300	191	323	273 (16)
Kabul Education University	240	255	322	300 (13)
Kunar University	301	214	322	222 & 217 (12)

Figure 6-4. The min, max, count, and mode of Kankor admission score for computer science discipline across the country.

It, however, turns out that Computer Science faculties of both Balkh and Parwan Universities become part of either the second or third cluster when we use other techniques of statistical data analysis such as mean, median, first quartile, second quartile and third quartile for the classification. Furthermore, the standard deviation of Computer Science candidates' scores in both these universities is very high in comparison to the standard deviation of Kabul, Herat, Polytechnic and Kabul Education Universities. It is therefore assumed that the minimum and the maximum admission scores alone do not represent an accurate picture of scores for clustering purposes. The following figure better shows the above-mentioned metrics to better depict this scenario. The data is sorted by the Median column.

University	Count	Min	Q1	Q2	Q3	Max	Mean	Median	Mode (Freq)	StdDev
Kabul University	419	260	293	307	324	345	306	307	298 (17)	18.62
Herat University	310	262	283	304	311	339	297	304	307 (14)	18.16
Polytechnic University	360	254	279	295	309	333	295	295	309 (15)	18.32
Kabul Education University	240	255	265	288	299	322	283	288	300 (13)	17.61
Sheikh Zayed University	290	169	211	278	308	324	259	278	304 (23)	52.92
Kandahar University	300	191	254	273	283	323	267	273	273 (16)	26.08
Kunar University	301	214	223	272	284	322	257	272	222 & 217 (12)	31.82
Nangarhar University	359	174	241	269	303	324	267	269	304 (12)	40.32
Balkh University	190	208	223	255	339	347	275	255	343 (12)	54.36
Parwan University	150	213	225	247	310	341	262	247	309 (9)	40.73
Badakhshan University	140	187	206	239	296	329	250	239	292 (9)	46.18
Kunduz University	249	182	198	223	305	330	247	223	197 & 193 (8)	51.52

Figure 6-5. More descriptive statistics of the Kankor admission score for computer science disciplines across the country.

Taking the above reasons into account, the author of this thesis, after some testing, considered other metrics in addition to the minimum and the maximum scores such as standard deviation, mean, mode, median, first, second and third quartiles to classify the universities into clusters. Furthermore, the following factors are also important, and it is deemed efficient to consider them while classifying universities for the same field of study:

- **Capacity of the University:** The capacity of universities varies greatly. Some are equipped with sufficient infrastructure to accommodate larger classes and labs, and have stable electricity, internet facilities and reasonable ratio of professors to students. Some universities offer specialty programs.
- **Life Standards of the Province:** Some provinces such as Kabul, Herat and Balkh have better life-standards in comparison to other provinces. The standards can range from job opportunities to public transportation, entertainment, and safety and security. This factor affects demand by students.
- **Student Dormitory:** Some universities in some provinces offer dormitory for their male and female students who live farther than 35 Km from the university campus.
- **MoHE Ranking:** Ministry of Higher Education also have some ranking standards and based on that all the public universities in Kabul (Kabul University, Kabul Education University, Kabul Polytechnic University and Kabul Medical University) are considered as "parent universities". Next are the "major provincial universities" possessing large capacity such as Herat, Balkh, Kandahar and Nangarhar Universities and the remaining universities are considered as "provincial universities".

Taking the above indicators and metrics into account and based on the Kankor dataset for the year 2015 with 1,799 instances of candidates admitted into Computer Science discipline, all the twelve universities were classified into three clusters, as illustrated in the following table.

University	Count	Min	Q1	Q2	Q3	Max	Mean	Median	Mode	StdDev	Label
Kabul University	239	260	285	295	302	333	293	295	298 (17)	13.01	Cluster 1
Herat University	160	262	273	283	289	327	282	283	287 & 283 (10)	11.88	Cluster 1
Polytechnic University	179	254	272	279	289	309	280	279	277 (10)	12.30	Cluster 1
Kabul Education University	120	255	261	265	273	287	267	265	263 (11)	8.49	Cluster 1
Kandahar University	149	191	244	253	263	317	250	253	263 & 261 (8)	26.27	Cluster 2
Nangarhar University	209	174	220	244	260	318	239	244	257 (7)	30.92	Cluster 2
Parwan University	100	213	221	232	246	313	236	232	234 (5)	19.85	Cluster 2
Balkh University	115	208	217	230	245	288	233	230	217 (6)	19.51	Cluster 2
Kunar University	149	214	218	222	231	268	227	222	222 & 217 (12)	11.81	Cluster 3
Sheikh Zayed University	150	169	186	214	226	310	212	214	214 & 174 (6)	28.05	Cluster 3
Badakhshan University	80	187	199	208	223	304	213	208	189 (5)	20.90	Cluster 3
Kunduz University	149	182	194	201	216	299	206	201	197 & 193 (8)	18.93	Cluster 3

Table 6-6. Classification of public universities offering computer science discipline across the country using descriptive statistics approach.

6.3.2.2. Method#2 - Data Mining Clustering Approach:

Clustering is another approach in which each public university is assigned to one of a set of clusters. To validate and confirm the accuracy of the descriptive statistics method explained above, the author of this thesis applied k-means clustering algorithm (K = 3) with different maximum run times to classify all the twelve public universities that offer Computer Science discipline into three clusters. The output of clustering method is similar to the output of descriptive statistics method, as echoed in the following cluster centroid table.

University	Cluster 1	cluster 2	Cluster 3	Label
	287.8955	251.6245	207.8120	<= Score
Kabul University	0.3666	0.0194	0.0000	Cluster 1
Polytechnic University	0.2251	0.0689	0.0000	Cluster 1
Herat University	0.2090	0.0530	0.0000	Cluster 1
Kabul Education University	0.0723	0.1325	0.0000	Cluster 1
Kandahar University	0.0466	0.1661	0.0426	Cluster 2
Nangarhar University	0.0434	0.2244	0.0900	Cluster 2
Parwan University	0.0113	0.0795	0.0786	Cluster 2
Balkh University	0.0096	0.0919	0.0933	Cluster 2
Sheikh Zayed University	0.0113	0.0442	0.1931	Cluster 3
Kunduz University	0.0032	0.0159	0.2259	Cluster 3
Badakhshan University	0.0016	0.0283	0.1031	Cluster 3
Kunar University	0.0000	0.0760	0.1735	Cluster 3

Table 6-7. Classification of public universities offering computer science discipline across the country using clustering approach

The author of this thesis included Kabul Education University in the first cluster for the following two reasons:

1. Kabul Education University is geographically located in the capital of Afghanistan and based on the MoHE's ranking it is considered as one of the "parent universities". Moreover, it is equipped with adequate infrastructure.
2. The other reason is that inspection of the cluster centroid table confirms it.

Likewise, Balkh University was chosen to be part of the second cluster rather than the third cluster.

The Decision Tree classifier with IG as the criterion, Random Forests, and KNN (K=5) were trained and evaluated to classify the data with clustered values as target variable, and achieved an overall accuracy of 77%, 77%, and 79% respectively. The generated tree by the Decision Tree classifier is easy to interpret. It is also realistic for providing feedback and guidance to Kankor candidates based upon the generated rules, as shown in the following figure.

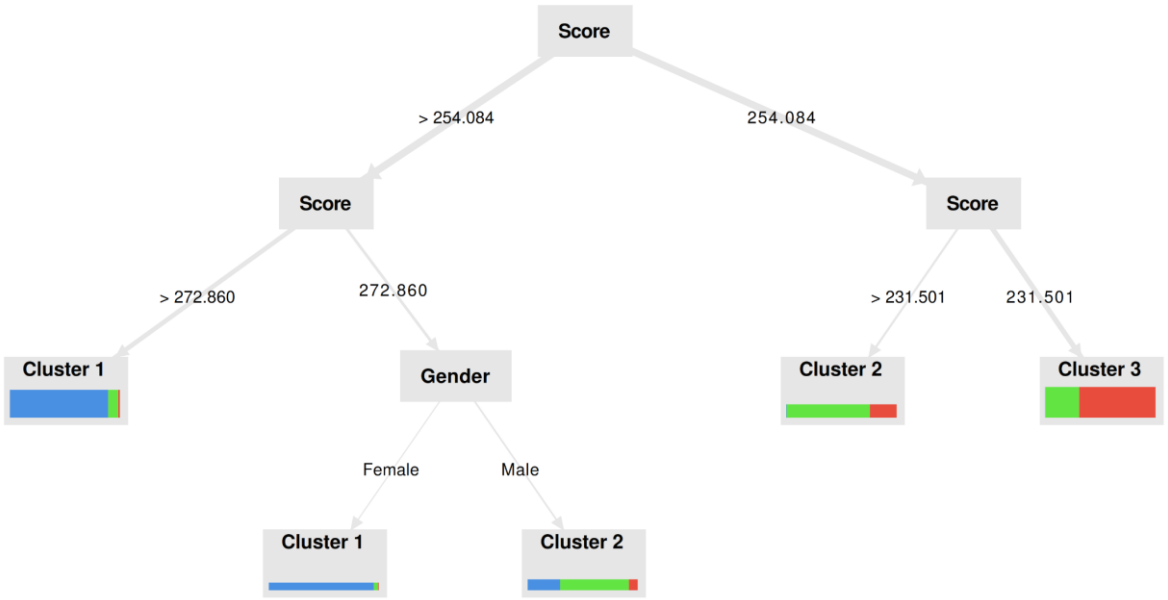

Figure 6-6. The output of the Decision Tree after Public Universities that offer Computer Science discipline across the country were classified in clusters.

The RapidMiner's 'Tree to Rule' operator, a meta learner, was used to create rule model from the above tree learner. The following rules are produced:

1. if Score > 254.084 and Score > 272.860 then Cluster 1 (510 / 52 / 9)
2. if Score > 254.084 and Score ≤ 272.860 and Gender = Female then Cluster 1 (128 / 5 / 1)
3. if Score > 254.084 and Score ≤ 272.860 and Gender = Male then Cluster 2 (60 / 127 / 16)
4. if Score ≤ 254.084 and Score > 231.501 then Cluster 2 (0 / 194 / 62)
5. if Score ≤ 254.084 and Score ≤ 231.501 then Cluster 3 (0 / 195 / 440)

Furthermore, Naïve Bayes classifier on the same dataset with the same settings as above was used to develop the classifier model and an accuracy of 80.32% was achieved. The confusion matrix of the Naïve Bayes classifier illustrated in the following table is nearly the same with the confusion matrix of the Decision Tree classifier.

Accuracy: 80.32% +/- 2.24% (mikro: 80.32%)				
	true Cluster 1	**true Cluster 2**	**true Cluster 3**	**class precision**
pred. Cluster 1	678	92	9	87.03%
pred. Cluster 2	15	377	129	72.36%
pred. Cluster 3	5	104	390	78.16%
class recall	**97.13%**	**65.79%**	**73.86%**	

Table 6-8. The confusion of matrix of the Naïve Bayes classifier for Computer Science discipline.

From the above confusion matrix, it can be concluded that out of 698 Kankor participants from first cluster 15 and 5 of them were predicted in the second and third clusters by the classifier model respectively. Out of 573 candidates who were part of the second cluster, the classifier model predicted 92 and 104 candidates to be part of first and third clusters. Finally, 9 and 129 candidates were predicted to be part of the first and second clusters out of 528 candidates who were part of the third cluster. The reason the candidates from one cluster were predicted by the classifier to belong to a different cluster is because a few candidates from third clusters had received high scores in the Kankor and vice-versa.

It is worth mentioning that the suggested methodology in the identification of universities based on a clustering system has not only been tested empirically but vast consultations have also been conducted with the epistemic community to ensure that the outcome of the clustering achieves the most realistic results.

6.3.3. Classification of Available Fields of Study for a Specific Public University

As mentioned in the beginning of this chapter most candidates know in advance which fields of study in which province or public university they want to opt for. These candidates are divided into the following three categories:

1. First, most candidates prefer their home-province universities.
2. Second, candidates who prefer to study at universities which are equipped with better facilities, including dormitory (especially for female candidates), regardless of which province the university is in.
3. Third, candidates who prefer provinces where there is better life-standards with greater job opportunities, public transportation, entertainment, safety and security.

In addition to the scenario presented in the previous section, another practical approach is to consider the above preferences into account. This method narrows down all the existing fields of study to those fields that are offered at the preferred university or province of the candidate.

In order to test and validate the above argument, the KRD1315 dataset was filtered for the year 2015 and only Herat public university was chosen as a sample for piloting the concept. The dataset was reduced to 4,139 instances of participants in 15 fields of study offered by Herat University.

As discussed in the previous section, some candidates with a score of for instance 280 or 290 were admitted into Engineering Faculty, while at the same time candidates with the same score were admitted into Computer Science, Economics or into Law and Political Science faculties. Therefore, it is a challenge for the prediction model to classify the candidates with 15 possible fields of study in the target variable since the same score can allow candidates to qualify for more than one fields of study.

The author of this thesis therefore proposes to classify the existing fields of study offered by a university or province into clusters prior to building the prediction model using the following two approaches.

6.3.3.1. Method#1 Descriptive Statistics Approach:

Like in the previous section, in addition to the 'minimum' and 'maximum', other descriptive statistics for the Kankor admission scores for each field of study were calculated. In addition, the high demand for some disciplines have also been considered in this calculation. Fields of study were thus classified and reduced to the following three clusters, as reflected in the following pivot.

Field of Study	Count	Min	Q1	Q2	Q3	Max	Mean	Median	Mode (Freq)	StdDev	Label
Medicine	160	298	308	314	323	338	315	314	313 (12)	9.91	Cluster 1
Stomatology	110	292	297	304	309	327	304	304	304 (9)	7.65	Cluster 1
Engineering	190	199	269	284	296	331	283	284	289 & 278 & 274 (7)	20.03	Cluster 2
Law and Political Science	180	272	278	284	289	325	285	284	288 & 284 (13)	10.35	Cluster 2
Economy	120	267	274	284	291	314	284	284	284 (9)	10.76	Cluster 2
Computer Science	160	262	273	283	289	327	282	283	287 & 283 (10)	11.88	Cluster 2
Public Administration	160	256	264	268	275	315	270	268	266 (12)	10.24	Cluster 2
Sharia	160	248	253	259	268	298	261	259	255 & 252 (10)	9.78	Cluster 2
Literature and Journalism	570	200	226	248	258	307	243	248	246 (21)	21.61	Cluster 3
Science	529	201	227	237	249	295	239	237	229 (23)	17.71	Cluster 3
Fine Arts	180	194	213	226	238	288	227	226	229 (8)	18.98	Cluster 3
Veterinary	80	209	216	226	241	295	230	226	228 (7)	17.51	Cluster 3
Agriculture	320	179	207	221	243	296	225	221	216 (13)	22.72	Cluster 3
Pedagogy Education	920	176	205	219	231	291	219	219	219 (27)	18.58	Cluster 3
Social Sciences	300	199	207	218	239	280	224	218	205 (18)	18.51	Cluster 3

Table 6-9. Classification of fields of study at Herat University using descriptive statistics with other metrics.

The following independent variables (gender, province, and Kankor score) were selected as input, and (output of clustering) was selected as target variable. Then, the data mining workflow process was created in RapidMiner Studio to develop prediction models using the following classifiers: Decision Tree, Random Forests, KNN and Naïve Bayes. In order to estimate how accurately the models learned by the mentioned learning classifiers and would perform in practice, cross-validation method with 10 folds and stratified sampling type was used.

The overall accuracy of Decision Tree, Random Forests, KNN (K=5) and Naïve Bayes was 89%, 85%, 87% and 88% respectively. From the result, using Decision Tree and Naïve Bayes and KNN give the best percentage of correctly classified instances of 89%, 88% and 87%. Random Forest gives a lower percentage of correctly classified instances of 85%. The tree generated by the Decision Tree classifier is reflected in the following figure.

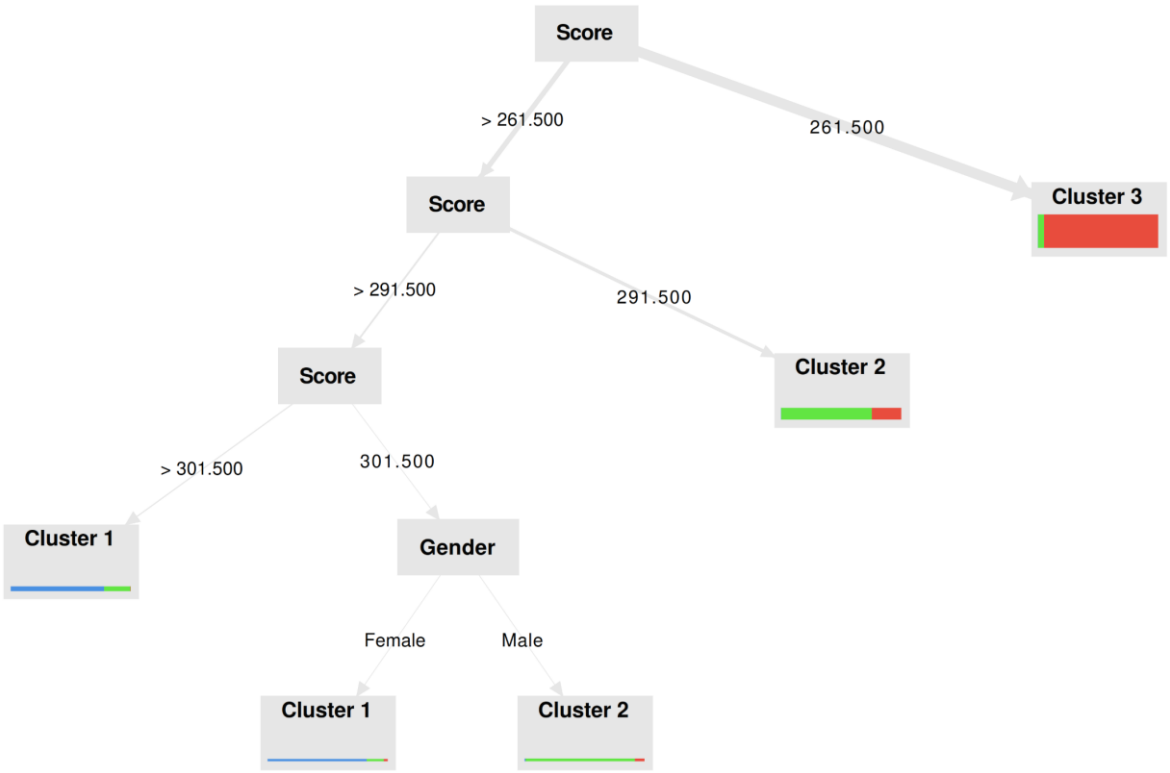

Figure 6-7. Decision Tree output after fields of study at Herat University were classified into three clusters.

The RapidMiner's 'Tree to Rule' operator was used to create rule model from the Decision Tree classifier and it converted the above tree into the following text-based rules and shows that 3,699 out of 4,139 training examples were correctly classified.

1. if Score > 261.500 and Score > 291.500 and Score > 301.500 then Cluster 1 (217 / 61 / 1)
2. if Score > 261.500 and Score > 291.500 and Score ≤ 301.500 and Gender = Female then Cluster 1 (52 / 9 / 2)
3. if Score > 261.500 and Score > 291.500 and Score ≤ 301.500 and Gender = Male then Cluster 2 (1 / 91 / 8)
4. if Score > 261.500 and Score ≤ 291.500 then Cluster 2 (0 / 668 / 217)
5. if Score ≤ 261.500 then Cluster 3 (0 / 141 / 2671)

To verify the validity of the approach, an additional experiment was added where fields of study were classified into four clusters, adding one more cluster to the sample of testing. These include the following clusters.

1. Medicine and Stomatology (Cluster 1);
2. Engineering, Law and Political Sciences, Economics and Computer Science (Cluster 2);
3. Public Administration, Sharia, Literature and Journalism and Science (Cluster 3);
4. Fine Arts, Veterinary, Agriculture, Pedagogy Education and Social Sciences (Cluster 4).

The same classifiers with the same settings were used to construct the models and an accuracy of 72.59%, 66.85%, 70.84% and 69.27% was achieved for Decision Tree, Random

Forests, KNN (K=5) and Naïve Bayes respectively. From the result, using Decision Tree gives the best percentage of correctly classified instances with an overall accuracy of nearly 73%. The Decision Tree classifier produces the tree illustrated in the following figure.

Figure 6-8. Decision Tree output after fields of study at Herat University were classified into four clusters.

The RapidMiner's 'Tree to Rule' operator produces to the following text-based rules with detailed information using the above tree.

1. if Score > 262.500 and Score > 291.500 and Score > 301.500 then Cluster 1 (217 / 59 / 3 / 0)
2. if Score > 262.500 and Score > 291.500 and Score ≤ 301.500 and Gender = Female then Cluster 1 (52 / 8 / 2 / 1)
3. if Score > 262.500 and Score > 291.500 and Score ≤ 301.500 and Gender = Male then Cluster 2 (1 / 86 / 10 / 3)
4. if Score > 262.500 and Score ≤ 291.500 and Score > 272.500 then Cluster 2 (0 / 378 / 110 / 19)
5. if Score > 262.500 and Score ≤ 291.500 and Score ≤ 272.500 then Cluster 3 (0 / 89 / 221 / 30)
6. if Score ≤ 262.500 and Score > 244.500 then Cluster 3 (0 / 24 / 478 / 187)
7. if Score ≤ 262.500 and Score ≤ 244.500 then Cluster 4 (0 / 6 / 595 / 1560)

6.3.3.2. Method#2 Data Mining Clustering Approach:

To ensure the most scientifically correct approach is chosen for accurate clustering purposes, additional techniques are also used. Therefore, the author of this thesis is also applying k-means clustering algorithms (K=3, number of clusters specified by the analyst) to classify similar fields of study at Herat public university together. The dataset comprises the following features (gender, Kankor score, and fields of study at Herat University). The output of clustering is slightly different compared to that of the descriptive statistics technique, as reflected in the cluster centroid table.

Faculty	Cluster 1	Cluster 2	Cluster 3	Label
	294.0520	252.4846	214.4179	<= Score
Law and Political Science	0.1863	0.0053	0.0000	Cluster 1
Medicine	0.1733	0.0000	0.0000	Cluster 1
Engineering	0.1430	0.0374	0.0012	Cluster 1
Computer Science	0.1278	0.0280	0.0000	Cluster 1
Stomatology	0.1192	0.0000	0.0000	Cluster 1
Economy	0.0986	0.0194	0.0000	Cluster 1
Public Administration	0.0498	0.0761	0.0000	Cluster 2
Literature and Journalism	0.0336	0.2397	0.1048	Cluster 2
Science	0.0238	0.1889	0.1304	Cluster 2
Sharia	0.0217	0.0935	0.0000	Cluster 2
Pedagogy Education	0.0087	0.1262	0.4208	Cluster 3
Agriculture	0.0054	0.0688	0.1234	Cluster 3
Fine Arts	0.0043	0.0367	0.0704	Cluster 3
Social Sciences	0.0022	0.0648	0.1170	Cluster 3
Veterinary	0.0022	0.0154	0.0320	Cluster 3

Figure 6-9. Classification of fields of study at Herat University using clustering approach (K=3).

Another similar experiment was conducted with k-means (K=4) and it also found almost the same result, thus no considerable variance was observed which could be of concern. A close investigation of the pivot in descriptive statistics section also shows the number of admitted candidates into the second and third quartiles for the Law and Political Science, Engineering, Economics and Computer Science disciplines are similar to the number of candidates who were admitted into Medicine and Stomatology disciplines. Hence, these disciplines were classified into ‘cluster 1’ in both cases (K=3 and K=4).

The following features (gender, province, and Kankor score) were selected as input, and (the result of clustering of fields of study) was selected as target attribute. Then the Decision Tree, Random Forests, KNN (K=15) and Naïve Bayes classifiers were used to build and evaluate models to classify the data with the three possible clustered values which achieved an overall accuracy of 72.38%, 71.51%, 74% and 72.46% respectively. From the result it can be concluded that all the classifiers give similar percentage of correctly classified instances. Most trees produced by the Random Forest classifier also lead to nearly similar rules to the rules that are produced by Decision Tree classifier. The Decision Tree classifier outputs the tree illustrated in the following figure.

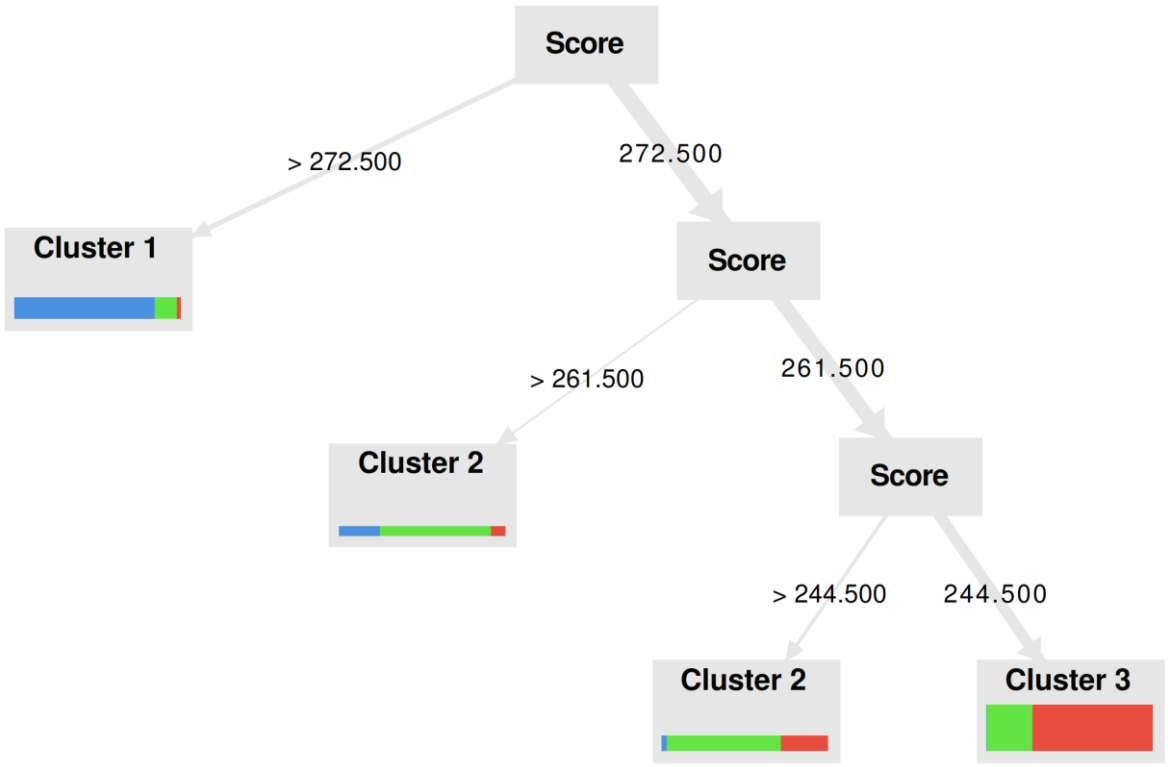

Figure 6-10. Decision Tree rules after fields of study at Herat University were classified into three clusters.

As conclusion, it is revealed that classification of the existing fields of study at a university or classification of similar fields of study offered by more than one universities into clusters can vastly increase the accuracy of the model. Moreover, the outcome of classifiers helps high school advisors to make data-informed recommendations on the fields of study to the candidates by 'discipline', 'university' and 'province'. Such recommendations would serve as realistic and vital guidance to both candidates and education officials.

Of course, if the process of classification of fields of study by 'province', 'university', and 'discipline' is carried out with more factors then the settings would be improved. As a result, more useful and realistic patterns and rules would be generated.

As discussed above in this chapter, the Kankor data currently lacks some key features whose presence while constructing the models might potentially increase the accuracy. The overall accuracy of the models developed in this section and the previous section is relatively low because of the following reasons:

- Generally, candidates with higher scores who were admitted into fields of study that were classified to a lower cluster can switch to fields of study that were classified to a higher cluster. However, scientific research has not been carried out by scholars to show the accurate facts and figures who switched disciplines from fields of study that are part of a lower cluster to a higher cluster and therefore the dataset lacks such information. Thus, the constructed models predict and classify these candidates to a higher cluster because they were built and trained without such information. This is one of the main reasons that has reduced the accuracy.
- Mainly, all the fields of study have limited seating capacities. The priority is given to candidates who pick these disciplines as their first choices out of the five choices

they have. Once a candidate picks a field as her/his first choice, if his/her Kankor score matches the admission score of that discipline they are admitted into that field of study. Otherwise, their second, third, fourth and fifth choices will be taken into consideration in the same way. This is one of the main reasons some candidates with a high score are either result-less or were admitted into disciplines that are classified into a lower cluster. It is 'first come, first served' and has been explained in the second chapter. Presently, the Kankor dataset lacks information on seating capacities of the disciplines. Moreover, it lacks the candidates' choices and the order they chose their five disciplines. Hence, the constructed models classify some candidates with a higher score in Kankor to a higher cluster while in actuality they are in a low cluster, and vice-versa, which thus is a major factor for the accuracy to drop.

Additionally, the Kankor data currently collected from the relevant government bodies lack key features such as the breakdown of scores for the following subjects: Mathematics, Sciences, Social sciences and Languages.

Furthermore, the information on the classification of high schools into public and private as well as relevant information to show whether they are geographically located in rural or urban areas is entirely lacking from the data architecture. A dataset with such key features could be very substantial in constructing and training the models that would certainly lead to better and higher accuracy of the models and serve as important tools for future decision making at the national level. Similarly, more useful and realistic patterns and rules would be generated that can help high school advisors to guide applicants in choosing their fields of study through informative tools. The availability of the identified key missing features will have national consequences that would lead educationalists to bring changes in the system constructively.

6.4. Prediction Model using Recommender Systems Approaches

Introducing a proper course of study requires careful and precise attention and assessment, particularly in Afghanistan, where candidates have (only) two chances in their lifetime to take the Kankor exam. Hence besides experimenting with Kankor results data through data mining techniques in the previous section, the author of this thesis also proposes consideration of high school students' marks from grades 10, 11, and 12.

Analysis carried out in the previous chapter revealed that high school marks do not have strong correlation to the success of candidates in the Kankor. Some high schools in major cities and most high schools located in rural areas, villages and in provinces with poor infrastructure and life-standards are very generous in giving scores to their students. The reason for such a biased and disorganized system of scoring is the inadequate monitoring and control by the Ministry of Education over its schools.

Furthermore, multiple analyses were carried out in the previous chapter to find out how graduates from public high schools performed in the Kankor in comparison to graduates from private high schools. The findings show that there are certain types of private high schools with unique criteria, rules and policies whose candidates really performed great in the Kankor. These unique conditions include providing more opportunities such as, the

provision of hostels so students can have more focused time to practice school work without external distractions, conducting more focused and regular testing requirements to sharpen students' skills to perform better in standardized testing. These schools also give a preference in their enrollment process to students with more advanced learning skills than those less committed. However, performance of graduates from the remaining public and private high schools were the same in the Kankor.

The analyses carried out validate that high school marks alone are not reliable metrics and with the current structure it is not deemed efficient to consider them as a basis for admission into higher education and eliminate the Kankor in Afghanistan. However, the role of high school marks in the calculation will gain further vitality due to the following implications:

- High school marks can play a fundamental role in the success of candidates to the condition that a standard and realistic scoring system is used in high schools. As already discussed in chapter 2, many policy advocates have proposed and are trying to replace Kankor with high school performance instead. However, due to strong opposition, the current system of Kankor remains a central policy for admissions. In a similar policy change in Iran, a portion of the admission of students in universities was awarded based on high school marks. Then, in 2010 it was announced that high school performance will entirely replace Kankor by 2012 which has not been implemented yet.
- As already mentioned in chapter 2, Kankor is a common measure in several countries, however at the same time, most education systems evaluate their students by means of different evaluation tools such as mid-term and final tests, assignments and exercises which result in a mark that reflects not only the students' knowledge but also their competence and choices of majors.

Hence, assessment and integration of high school marks alongside the Kankor results data will make the model more effective and will also generalize its usability and adoptability in other countries that lack the Kankor exam as a criterion for admission to higher education.

The curriculum for high schools in Afghanistan includes math, physics, geometry, trigonometry, chemistry, biology, geology, national languages (Pashtu and Dari), foreign languages (English, Arabic, even other languages in certain schools), computer, history, geography, Islamic studies (Islamic Beliefs and Interpretation of holy Quran), civil studies, Ethics, sport and vocational studies. In Afghanistan's educational system one grading scale (0 to 100) is used for both general and higher education. In general education, a grade below 40 and in higher education a grade below 55 for a specific subject is deemed unsatisfactory and a fail.

Since some high school subjects play a strong role in one field of study and are less significant in another, to streamline the strengths and weaknesses of applicants in the relevant fields of study, it is deemed efficient to identify the weighting of high school subjects in relation to the fields of study. For instance, Engineering requires in depth knowledge of subjects such as, math, trigonometry, geometry and other relevant high school subjects, while this is hardly the case in Social Sciences. Experts in the fields can classify the more than one hundred higher education fields of study into main streams such as

sciences and social sciences and then determine the contribution of high school subjects for each stream.

It is worth mentioning that experts (school teachers, university professors, policy makers, and other relevant branches) will need to know the weighting of every subject in every field of study precisely, so the system can infer a number for each student in each field of study based on the marks obtained in those subjects and the weightings given by experts. For instance, experts from mathematics may agree that if a student has a high mark in math it means that he/she will be better prepared for Engineering, Computer Science and Economics disciplines.

To enrich this insight with supplementary data, as part of the research, a survey was conducted using multiple methodologies including questionnaires, interviews and focus groups among 53 university professors, in order to determine the weighting of high school subjects for higher education fields of study. The following fields of study were chosen as a sample for the pilot concept: Agriculture (10 respondents), Economics (8 respondents), Computer Science (7 respondents), Engineering (4 respondents), Law and Political Sciences (5 respondents), Literature (9 respondents), Medicine and Stomatology (6 respondents), Journalism (2 respondents), and Theology (2 respondents). It is important to note that, since the number of respondents is rather small, the output of these questionnaires and interviews are not all encompassing and are statistically limited. However, even this limited survey allows us to test the technique and gain insight into the weighting of high school subjects for university fields of study. Moreover, the approach applied in this analytical framework can in the future be extended in a national level survey to include a wider range of respondents across many disciplines. Such a research would lead to more statistically accurate weighting for high school subjects.

The professors from each discipline were asked to determine the importance of each subject through a weighting system and to provide their argumentation as to why one subject should be weighted differently than another. In this section, the research shows a consolidated view of this weighting system for some of the disciplines.

Economics:

Subject	Rationale
Arabic Language	The contribution of this subject in Economics is to be 0%, 5% and 10%. Those who gave weights of 5% and 10% to the subject expressed the following reasons: Understanding the traditional rules of trade which provides a deeper foundation to the study of economics particularly in the Arab world whose trade history has influenced the world's major schools of thoughts in business, trade and investment. Thus, Arabic helps students to learn the Arab civilization through its own language. On average it is weighted at 5%.
Biology	All respondents believed that this subject has nothing to do with Economics and therefore was weighted at 0%.
Chemistry	It is weighted at 0%.
Civil Studies	On average it is weighted at 30%. Respondents gave civil studies weights ranging from 20% to 50% with the reasoning that civil studies enable students to learn the dynamic

	functions of a society regulated by norms and codes of conduct. Afghanistan's current civil code of conduct follows universal norms but is mainly contextualized from the French civil rights code.
Computer	This is weighted at >=50%. All professors believed that a changing world is more than ever reliant on the role of technology and how it transforms every function particularly the role it plays in helping students analyze data using statistical tools, conducting research and understanding economic trends, learning new research methodologies by applying them on computer applications as well as knowledge management and overall academic management.
English Language	This is weighted at >=50%. All respondent believed this subject is significant and empowers students to use a wider range of economic literatures that is mostly written in English and that English as an international language is useful for communication as well as establishing the grounds for business relationships with international organizations. Additionally, the MoHE plans to have the curriculum delivered only in English. It is considered a necessity in today's global world to be able to communicate with the rest of the world through an international language.
Geography	This is weighted at 20%. All respondents believed it is required to understand the spatial dimensions of regional economic theories.
Geology	This is weighted at 0% because it has no direct impact and relationship.
Geometry	This is weighted at 5%.
History	On average, this is weighted at 20%. Only through the study of the past, the solutions for the future can be speculated and designed.
Islamic Studies	The opinions on this were widely varied ranging from being weighted at 0% to 30%. It should be noted however that this subject is taught across all disciplines nationally. The average respondent's weight of this subject is determined at 10%.
National Languages	This is weighted at 30%. Dari and Pashto are the two national languages and are spoken distinctively in different regions. However, both languages are key if graduates plan to work in governmental positions.
Math	On average, this is weighted at 50%. As there are several subjects in Economics that require fluency in Math to understand the data, this subject is considered a key component in the study of this discipline.
Physics	On average, this is weighted at 15%. The relevance of the subject gains significance when the study of products, transportation, and marketing comes into play.
Vocational Studies	This is weighted at 5%. This subject is considered to motivate students in adapting creativity to their day-day activities.
Trigonometry	On average this is weighted at 30%. The skill is needed in some branches of Economics to understand data.

Computer Science:

Subject name	Rationale
Arabic Language	On average this is weighted at 10%. Those advocated for its teaching believed knowing an additional language always contributes to the competency of the candidate in a fast growing and interconnected world. Those who were not supportive did not see any professional value in the teaching of an extra language other than English.
Biology	On average this is weighted at 10%. The reason for its relevance would be if graduates decide to continue their higher education, this might provide a better foundation for those pursuing careers or fields of study concerning the study of living things. Artificial

	Intelligence (which uses neural networks are partly inspired by biology) has led to a massive technology revolution Since 2012.
Chemistry	Like biology, this is weighted at 10% on average.
Civil Studies	The weighting of respondents was highly skewed, ranging from 0% to 70%, averaging to 15%. Those against its role argued it has no direct impact but those who argued for it considered it vital in how students function in a society.
Computer	For evident reasons, this is weighted at 100%.
English Language	For apparent reasons, this is weighted at 100%. Additionally, plans are being prepared for the curriculum to be delivered in English.
Geography	On average, this is weighted at 25%. Those who advocated for it, argued that this skill plays a key role when students are engaged in using geographical applications.
Geology	On average, this is weighted at 10%. When and if graduates follow a professional career in the field, this field is considered to have value in various aspects.
Geometry	On average this is weighted at 50% although the respondents' answers were highly skewed. The argument was that certain tasks such as to design 3D objects, robotics, animation and gaming require knowledge of geometrical computation.
History	This is weighted at 0%.
Islamic Studies	This is weighted at 20% on average although the answers were widely varied. Those arguing against, believed that the study of theology has no significance in making a student skillful to perform better as a computer science student, and believed that this adds to the volume of learning content and detracts students from learning of the actual discipline content. Those who argued for, believed that as Muslims, it is the moral responsibility of all students to undertake the study of Islamic Studies. The research observation reveals that the respondents' answers were influenced by personal emotions and beliefs rather than by providing an unbiased answer.
National Languages	This is weighted at 0%. Most of current content and curriculum is provided and is available in English although students are already proficient in their native languages and so this is not considered significant.
Math	For obvious reasons, this is weighted at 100%.
Physics	On average this weighted at 40%. Those who advocated less support for it, believed that only in particular branches of computer science such as hardware this may be more relevant but that the existing curriculum of computer science covers a general study of the field and this may not be that relevant.
Vocational Studies	This is weighted at 20% on average. Some believed it ignites creativity and innovation which is very important, and some believed it is not a prerequisite since it has no vital relationship in the primary skills required for a student of computer science discipline.
Trigonometry	This is weighted on average at 60% although respondents' views varied widely. It should be noted that those who showed less support for it, believed that its application may be more significant in particular branches of computer science than in the general discipline that is part of the current curriculum of MoHE.

Engineering:

Subject name	Rationale
Arabic Language	It is weighted at 0% since all teaching materials are in English and MoHE plans to change the medium of teaching to English as well.
Biology	It is weighted at 0%.

Chemistry	In civil engineering, it is weighted at 40%. Currently, this subject is offered with around 10 credits and thus is relevant. However, in other branches of engineering such as architecture it is not relevant.
Civil Studies	It is weighted at 10% due to the applicability and complexity of social relationships.
Computer	It is weighted at 50% and currently there are multiple computer related subjects including programs such as AutoCAD, Corel Draw, and 3D Max.
English Language	This is weighted at 60% on average. The delivery of the content is conducted through National Languages, but all the teaching material is in English.
Geography	This is weighted at 10% since understanding of its fundamentals is key to learning topics such as topography, climate and other areas related to geography.
Geology	This is weighted at 10% due its relation to understanding of earth crust, types of rocks, rivers, seas, etc.
Geometry	This is weighted at 100% since the entire discipline revolves around this subject as one of the key fundamentals.
History	It is weighted at 10% particularly for architecture students studying urban planning.
Islamic Studies	Although it is taught everywhere, all respondents argued that it should be weighted only 0%.
National Languages	It is weighted at 0% though respondents argued that in some cases it may determine whether one can get certain government jobs particularly when those positions are based in the capital or certain regions of the country where a language is spoken predominantly.
Math	Like geometry, it is weighted at 100%.
Physics	Like geometry and math, it is weighted at 100%.
Vocational Studies	It is weighted at 10% factoring in the element of innovation and creativity.
Trigonometry	It is weighted at 100% like geometry, math and physics as these are the fundamental subjects of the discipline.

Agriculture:

Subject name	Rationale
Arabic Language	It is weighted at 0%.
Biology	Since it is a compulsory subject and is a pre-requisite to several other subjects in this discipline, this is weighted at 90%.
Chemistry	Same as Biology, it is a compulsory subject and prerequisite for some other subjects in Agriculture. The respondents believed that all fertilizers, agricultural chemicals and the effects of such substances on plants and animals require knowledge of this subject. In average it is weighted at 90%.
Civil Studies	It is weighted at 20% due to the applicability and complexity of social relationships.
Computer	It is weighted at 50% because it needs to be regularly used when conducting research and other required academic activities.
English Language	It is weighted at 40% because it is needed every day, and at the international level.

Geography	It is weighted at 50% on average. The respondents who advocated for it argued that knowledge of geography will be key for students to know not only about the geopolitics of the land but also how geography plays a vital role in agronomy, water management and how climates of various geographies affect agricultural activities.
Geology	This is weighted at 100%. This is both essential and prerequisite to other subjects in Agriculture. It is needed for examining the soil and to determine the appropriate type of soil for agriculture as well as to understand the resistance of plants and animals in various geographical areas.
Geometry	It is weighted at 0%.
History	It is weighted at 0%.
Islamic Studies	The inclination to weigh the subject at 10% is because of recognizing the fact that the Government of Afghanistan has included the subject in all disciplines.
National Languages	It is weighted at 0%.
Math	Since it is an essential necessity for other professional subjects, it is weighted at 40% on average.
Physics	As it is partly essential to subjects such as veterinary, on average it is weighted at 30%.
Vocational Studies	Views among respondents vary. On average, this is weighted at 25% as engagement of students, especially those who purse direct work in practical agricultural activities, in various vocations will be helpful in their long-term careers.
Trigonometry	It is weighted at 10%.

Medicine:

Subject name	Rationale
Arabic Language	It is weighted at 10% on average.
Biology	It is weighted at 100%. All respondents mentioned this subject as compulsory and prerequisite for other subjects including histology, microbiology, pathology, physiology, anatomy and several other subjects.
Chemistry	Like Biology, it is a compulsory subject and prerequisite for biochemistry, pharmacology, pathology and other relevant subjects and is weighted at 100%.
Civil Studies	It is weighted at 40% due to the applicability and complexity of social relationships.
Computer	It is weighted at 50% because it is regularly used when conducting research and other required academic activities. Furthermore, respondents believed this subject is essential specially when operating advanced medical devices such as ultrasound, echocardiogram, endoscopy and other technologies.
English Language	It is weighted at 50% as most medical science terms are derived from Latin and English roots. Additionally, the MoHE plans to make English the universal medium for teaching Medicine.
Geography	It is weighted at 0%.
Geology	It is weighted at 0%.
Geometry	It is weighted at 0%.
History	It is weighted at 0%.

Islamic Studies	On average it is weighted at 10%. Those who argued for it believed that Islamic Studies has a lot to offer to strengthen the foundations of moral and ethical beliefs of students, for the common good, when they practice medicine.
National Languages	It is weighted at 0%.
Math	It is weighted at 20% on average since it is a pre-requisite and an essential necessity for other technical subjects.
Physics	It is weighted at 10% on average particularly when this subject is studied in the first year of the medical school as a pre-requisite compulsory subject.
Vocational Studies	It is weighted at 0%.
Trigonometry	On average, it is weighted at 15% since its applicability may become more relevant as graduates pursue further higher education to obtain specialty in certain fields that would require the knowledge of trigonometry.

Law and Political Sciences:

Arabic Language	It is weighted at 55% on average. Since the foundation of civil and national law is in Arabic, this subject plays a very important role in this discipline.
Biology	It is weighted 5% on average because it has no direct role in law and political sciences, however, it can be useful when students are learning topics on subjects like forensic medicine and criminology.
Chemistry	This is weighted at 5% as it has little significance on the field of law. However, at times it is considered helpful when students are dealing with matters of criminology and forensic medicine.
Civil Studies	It is weighted at 40% on average. Respondents' views were the same, but their weighting scores were widely varied and range from 5%, to 40% and 100%. This subject introduces social issues and good behavior and discusses the culture and communication of a country.
Computer	On average it is weighted 50%. It is useful for student when doing various field work.
English Language	On average it is weighted 40%. English plays a major role in scientific research and study of primary resources which are mostly available on in English.
Geography	On average it is weighted 20%. This is considered rather significant when dealing with topics such as criminology geopolitics and regional and global matters.
Geology	It has very little role and therefore on average it is weighted 5%.
Geometry	Same as geology.
History	On average it is weighted >=50%. The curriculum gives it a heavy weight by including around 16 credits throughout the academic program.
Islamic Studies	On average it is weighted >=70%. Most of law codes are based on religion and so this is applicable to the study of law across the discipline.
National Languages	On average it is weighted at 30% although the responses ranged from 10%, 30% and 100%. A major reasoning by those who argued for it is the relevance of languages to better understand unique cultural inferences.
Math	On average it is weighted 20%. Students need basic arithmetic skills to apply in data analysis, laws of inheritance and other related areas of the discipline.

Physics	It is weighted at 5% as basic skills at times will be required in forensic medicine and criminology.
Vocational Studies	It is weighted at 5% as professionals need solid personalities to ensure their relationship with the suspect or criminal is based on the law.
Trigonometry	It is though not directly relevant to the field; respondents weighted this at 5%.

After data collection was completed, the respondents' answers were analyzed by the author of this thesis to find out the average weighting score for each of the high school subjects in each discipline. The summary table is given below in **Table 6-10**.

Disciplines	High school subjects																	
	Arabic Language	Biology	Chemistry	Civil Studies	Computer	English Language	Geography	Geology	Geometry	History	Islamic Beliefs	Interpretation of holy Quran	Dari Language	Pashtu Language	Math	Physics	Vocational Studies	Trigonometry
Economics	5%	0%	0%	30%	60%	60%	20%	0%	5%	20%	5%	5%	15%	15%	50%	15%	5%	30%
Computer Science	10%	10%	10%	15%	100%	100%	25%	10%	50%	0%	10%	10%	0%	0%	100%	40%	25%	60%
Engineering	0%	0%	40%	10%	50%	60%	10%	10%	100%	10%	0%	0%	0%	0%	100%	100%	10%	100%
Agriculture	0%	90%	90%	20%	50%	40%	50%	100%	0%	0%	5%	5%	0%	0%	40%	30%	25%	10%
Medicine	10%	100%	100%	40%	50%	50%	0%	0%	0%	0%	5%	5%	0%	0%	20%	10%	0%	15%
Law and Political Sciences	55%	5%	5%	40%	50%	40%	20%	5%	5%	60%	75%	75%	15%	15%	20%	5%	5%	5%
Theology	100%	5%	0%	80%	40%	40%	20%	0%	0%	40%	100%	100%	40%	40%	20%	0%	30%	0%
Journalism	30%	10%	10%	50%	60%	60%	25%	10%	10%	25%	20%	20%	50%	50%	15%	5%	20%	5%
Literature	80%	5%	5%	35%	40%	100%	40%	10%	5%	40%	30%	30%	100%	100%	5%	5%	5%	5%

Table 6-10. (%) shows the average weighting of high school subjects in relation to the disciplines.

The '*weighting system*' developed in this section is supported by both qualitative and quantitative research. The actual high school marks of hypothetical candidates (given in the table below) and the average weighted scores (given in the table above) allow us to put together a system which better predicts for which field of study each candidate would be more likely to qualify.

High School Graduates	High School Subjects																	
	Arabic Language	Biology	Chemistry	Civil Studies	Computer	English Language	Geography	Geology	Geometry	History	Islamic Beliefs	Interpretation of holy Quran	Dari Language	Pashtu Language	Math	Physics	Vocational Studies	Trigonometry
Freshteh	58	65	67	70	100	100	72	64	100	52	55	55	75	60	100	100	76	100
Moslem	70	98	100	80	70	75	80	100	60	60	75	70	70	65	80	75	65	65
Maryam	98	60	59	90	75	65	68	70	64	62	96	95	80	60	70	65	70	63
Wahid	98	98	92	90	90	96	60	70	100	70	70	68	60	65	100	100	70	100
Solaiman	99	69	65	98	95	90	80	62	64	100	100	90	65	75	60	58	65	70
Fatemeh	63	100	100	80	84	88	99	100	56	59	64	61	60	65	80	75	80	78
Mustafa	100	60	65	96	95	80	78	59	61	85	90	92	84	82	70	70	70	71

Table 6-11. High school marks of hypothetical candidates in a 0-100 scale.

Let's assume "**S**" denotes marks of high school subjects of a student and "**W**" denotes the average weighting of those high school subjects for a discipline. Then the DOT PRODUCT of "**S**" and "**W**" shows an approximate preparedness of the candidate for the discipline. The following formula illustrates the above definition and the letter **"n"** denotes the total number of high school subjects.

$$approximate\ preparedness\ (student, discipline) = \sum_{i=1}^{n} S_i W_i$$

Taking the above formula into account, from the average weighting of high school subjects in relation to the disciplines (table 1) and students' marks in high school subjects (table 2) an approximate preparedness of each candidate required for each discipline was calculated, as reflected in the following table.

Disciplines	High School Graduates						
	Freshteh	Moslem	Maryam	Wahid	Solaiman	Fatemeh	Mustafa
Engineering	570	439	398	575	420	466	432
Computer Science	534	424	402	532	442	462	444
Literature	465	459	484	500	539	464	547
Theology	439	471	549	518	588	459	583
Agriculture	427	484	373	477	400	511	392
Journalism	363	347	355	384	397	360	399
Law and Political Sciences	347	361	405	401	451	357	438
Medicine	316	350	274	381	313	366	302
Economics	298	247	243	299	278	268	279

Table 6-12. Standardized Ranking Scale (SRS)

This Standardized Ranking Scale (SRS) is a tool developed to enable advisors to guide students in the selection of their discipline, using a recommendation system that is based on students' high school marks. The tool is designed only objectively, excluding any subjectivity in the process, in order to ensure the outcome of the recommendation is closely related to the actual skills possessed by the student and how much those skills and talents are directly relevant to a particular field of study. In other words, the SRS will serve as a key talent sourcing for Kankor system. For instance, table 3 exemplifies a scenario where the listed students' skills are inventoried and based on their skills the most relevant disciplines are recommended.

There are several challenges which require to be addressed prior to employing this suggested approach.

- Due to limitations in resources and the scope of this research, data collection for identifying the weighting of high school subjects and their relevance to the fields of study has been limited to mainly Herat University.
- The views and answers of respondents in every discipline vary widely. For example, some were very generous, and some were very restricted when determining weighting scores for high school subjects in relation to the disciplines.
- Biased and conflicting opinions of respondents in relation to the weighting of some subjects may have overshadowed a true weighting value particularly for religious studies, profession studies, geography, history and national languages.
- Results from the questionnaires and interviews also reveal that some high school subjects play an important role in more than one discipline. For instance, Biology and Chemistry are very important in Medicine, Agriculture and Veterinary disciplines. Likewise, Math, Physics, Trigonometry and Geometry are essential in Computer Science, Engineering and other similar disciplines.
- Furthermore, there are a few high school subjects such as Civil studies, Profession studies, Drawing and Calligraphy may have a productive role in some disciplines. However, in high schools these subjects are not given any practical importance and usually no proper attention is paid to them.
- The list of suggestions provided by the recommendation system may not always be in line with the actual talents of the tested student and thus a field of study for which the student would be qualified may not be recommended. For example, because the weightings of English language, Computer and Math in disciplines such as Computer Science and Engineering are higher than in Economics, the system may not put Economics on its list of recommendations.
- It is also worth mentioning that Education (pedagogy) is completely different from other disciplines. It is comprised of several specialized departments with the aim of training teachers for primary, secondary and high schools. For example, biology and chemistry subjects studied in high school are, respectively, very useful for candidates who go to the Biology and Chemistry Departments of Faculties of Education. Likewise, other unique subjects play unique roles in their corresponding departments. Thus, it is very difficult to determine the weighting scores of high school subjects for Education. However, in some provinces such as Kabul, this challenge does not exist

since this discipline has been promoted to a university. Officials in charge of Faculty of Education in Herat province have requested MoHE many times to promote it to a university but this is yet to be done.

After identifying the above challenges, this research proposes the followings to make proper modifications:

- In the current structure some high school subjects play similar roles in more than one disciplines. Therefore, it is deemed efficient to classify all the existing fields of study into clusters (main streams). Determining the role of high school subjects in the identified clusters can be efficient and systematic in comparison to the existing fields of study.
- Presently, as described, there are subjects such as Islamic and vocational studies which play the same role or have the same weighting across all natural sciences disciplines. However, respondents' views and answers were widely varied both in rationale and weighting. Thus, it makes more sense to assign the same weight to such subjects in clusters of similar disciplines.
- Since the study will be a national analysis, it will require the application of broader methods and coordination with professionals from across the disciplines to ensure quality data and inclusive analysis which are required to optimize the most relevant finding. Therefore, it is strongly recommended that any further research in this area should be a collaborative effort so that specialized teams identify, collect, design and analyze contextual data.

This approach opens up a whole range of possibilities and provides many benefits and other relevant opportunities. The followings are some major examples:

- Specialized studies at high school: As explained in chapter two, in the current Afghan education system high school students do not have the option to study in specialized streams, such as sciences and social sciences. If at any time the Ministry of Education decides to offer specialized studies to students at high schools, the results of this research and particularly the method explained here can be used by education experts to divide high school subjects into science, social science and other main streams. Furthermore, student advisors in high schools can benefit from this research and use the methods outlined here to recommend specializations to students in an efficient manner. This research also makes it possible to automate such recommendations using data mining and recommender system techniques.
- Classification of disciplines: Moreover, this last approach establishes the ground work to methodologically classify more than a hundred existing fields of study into sciences and social sciences or other relevant streams.
- Specialized Kankor: For all high school graduates the Kankor is used as an admission tool into higher education. In the current structure all candidates take one unique Kankor exam for all the existing disciplines regardless of the candidates' tendency, interest and competence. This means the candidates interested in Fine Arts and the candidates interested in Engineering take the same Kankor exam. This method enables scientists to make specialized Kankor for the identified streams so that the

candidates' skills and competences as well their tendencies and interests play a more direct role in their admission.

- Improve collaboration between MoE and MoHE: The suggested approaches in this study bring scholars and policy advocates from both Ministry of Education and Ministry of Higher Education closer together for collaboration to identify key features of data, which are essential in data mining applications, in order to make a better architecture to record the data.
- Recommender systems applications: Finally, this approach leads to the application of recommender systems to further equip high school advisors, so they can provide constructive guidance for the candidates on their academic career. The entire process of constructing profiles for disciplines (fields of study), constructing profiles for students and the application of recommender systems are presented in this paper (Castellano, Barranco, and Martínez 2011). The results of this paper can be contextualized and applied for Afghanistan.

Chapter 7

Chapter 7. CONCLUSION AND FUTURE WORKS

The collected data and the literature show that the challenges and problems of Kankor candidates in Afghanistan are profoundly deeper in comparison with the problems of candidates in neighboring and regional countries. A few of the problems are as follows:

- lack of specialized studies at secondary and high school levels as well as lack of specialized Kankor,
- lack of pre-university courses to prepare the candidates for Kankor,
- lack of academic counseling organizations within/outside the education system to support the candidates in the critical process of making decisions about their careers,
- lack of proper facilities and systems to train the candidates on the methods and techniques of exam cram,
- not enough qualified teachers at secondary and high school levels,
- lack of proper classification of fields of study into main streams,
- the limitation to attend *Kankor* only twice in a lifetime,
- low literacy or illiteracy of parents,
- and the sharp and continuous increase in the number of *Kankor* participants.

These challenges cause most applicants to choose their field of study in complete randomness without systematic preparation and understanding of the consequences. Therefore, many candidates drop out of higher education, or they discover they are not interested in the fields they are studying. This negatively impacts the quality of higher education, students, and the society in general. The collected data show Afghan Universities on average have a seating capacity which is less than one-fourth of the total number of candidates. Almost one-third of the candidates who were admitted into universities were either dropped out, left, changed their disciplines or it took them longer to complete their higher education studies. This is noticed in other countries too. However, in Afghanistan this problem is profoundly deeper due to the following reasons: (1) candidates have only two chances to take the Kankor, and (2) some candidates do well in Kankor but do not get admitted into their desired disciplines due to limited seating capacity while some candidates who have taken the spots either drop out, leave or change their disciplines after some time.

The focus of this research was to investigate practical solutions and approaches based on data mining, recommender systems and assessment methods to support policymakers to reform the education system and to facilitate counselling services aimed at guiding of Afghan Kankor candidates so that they can plan their careers in the light of their skills and interests. Improving Afghanistan's education system, mainly through enhancing Kankor methodologies, as outlined in this research, will not only alleviate the associated challenges but will also be a beginning for a multiplier effect in various other dimensions, be it social, economic, political and environmental.

The findings of this research are mainly based on 1.5 million records of Kankor candidates (2003-2015), 6,000 records of high school students' marks, 2000 questionnaires and interviews, the author's observations as well as empirical studies. The collected data were systematically cleansed and analytically explored and analyzed from different perspectives. Findings are systematic methods enabling educational institutions to preprocess their data effectively and efficiently for example by properly auto filling the missing gender values, missing locations for high schools, and matching high school marks of Kankor candidates with their Kankor results through structured processes. Furthermore, these preprocessing and transformation methods can be used to cleanse the new datasets.

The concept of data mining in education and other sectors in Afghanistan is quite new, and in particular the concept I chose is completely new and, therefore, there is no literature to refer to. The existing studies and literature carried out in other countries 1) dealt with better organized and richer datasets, 2) their classifier models were built to classify instances into a very small number of classes such as sciences and social sciences, and 3) instances from different classes did not overlap much. None of these is true in the context of Afghanistan.

This research investigated approaches to systematically narrow down the existing disciplines while still remaining practical in the context of Afghanistan. The research outcomes include the construction of prediction models classified into the following four scenarios:

1. Pass or Fail: All Kankor participants were classified into two classes and were labeled as "Pass" or "Fail".
2. Educational Institution Types: All Kankor participants were classified into the following three classes "Public University", "Private University & Technical and Vocational Institutions" or "Fail".

In both scenarios 1 and 2 the classifier models worked well since they are responsible to classify instances into a few limited classes. The result from both classifiers are useful and support advisors to provide guidance to the candidates at the very general level. It, however, is not robust and comprehensive.

1. University-wise or Province-wise: Most candidates prefer to study in their home province university or in well reputed universities with better infrastructure in provinces with better life-standards and safety and security. Taken this option of candidates into account, this research methodically narrowed down more than a hundred disciplines to the disciplines offered by university or province. The accuracy of the classifier models was very low and not representative since some disciplines have similar admission scores. Therefore, this research, prior to developing the classifier models, classified all the disciplines of every university or province into a limited number of clusters. This led to representative accuracy of the classifier models.
2. Discipline-wise: For candidates who know what discipline they want to study; this research took their desired field of study into account identified all the universities which offer the discipline. Like in step 3, the admission scores vary widely from one university to another and they closely overlap for some universities. Prior to building

classifier models, this research classified the universities into a few clusters and the accuracy of the models turned out to be representative and illustrative.

The academic career guidance to the candidates must be prior to their Kankor exam. Hence, this research proposed e-Kankor which simulates the actual Kankor to estimate the performance of candidates through several pilot tests, and provides the results to the classifier models as unseen data.

The above classifier models were built using the Kankor data. It is worth noting that few countries use Kankor exam as a medium to higher education and in most countries students' high school marks are used to evaluate their admission into university. The findings in this research show that in Afghanistan, students' marks at high school alone are not reliable to be used as admission criteria into higher education. However, high school marks added value to previous classifier models and are valuable to determine the strengths and weaknesses of candidates in various disciplines. Therefore, this research also proposed another approach which is to use high school marks to recommend relevant disciplines to candidates, since different high school subjects play significant roles in different disciplines. Therefore, this research proposed associating weighting scores to high school subjects in relation to the disciplines. The output of this approach was found to be good. It turns out that determining the weight of subjects in relation to the clusters of disciplines can be more useful. But the challenge remains that disciplines in Afghanistan are not categorized into sciences, social sciences and or other relevant streams.

All the above approaches used either Kankor data or high school data and cannot truly reveal the interest and tendency of the candidates. Surveys and observations also show that simply asking the candidates about their interest is not enough because most candidates and their parents are unfamiliar with most of the disciplines. Mainly, they choose Medicine, Stomatology, Engineering and a few other "prestigious" disciplines as these are well-known in the society. It happens often that a candidate's choice of discipline does not represent their skills, competence or interest. Therefore, this research investigated methods to enable advisors to better understand and gauge the interest of the candidates. This research proposed an assessment test comprising of a non-technical set of Yes/No questions carefully designed to expose the candidate's tendency in relation to the disciplines.

Advisors can benefit from exploring and using all the proposed models since Kankor data lacks some significant features and some high schools are very generous in scoring their students and also because high school studies and Kankor exam are not specialized. The models proposed in this research, if executed systematically, will on average support 250,000 Kankor participants every year in choosing their fields of study. Moreover, they will support over 1,000,000 students who are currently in high school. That would transform the selection of careers for the generations to come. Moreover, the proposed models provide a roadmap for systematic student advising to become an integral part of the education structures. Furthermore, this research establishes the foundations to streamline standardized testing in a country which is in the very initial stages of experiencing the power of data. Such a transformation, powered by insight from data, may be the beginning of a paradigm shift. Finally, this work opens up numerous opportunities for further investigation.

The contribution of this research concept is not just limited to Afghanistan but can be generalized and adopted in other countries with Kankor examination. Finally, even though not every country has a Kankor exam, the results of this research will have the potential to be generalized and applied to other contexts similar to that of Afghanistan.

7.1. Further work and Opportunities

The Kankor data currently lacks some key variables listed below, which might potentially increase the accuracy of the classifier models. Future research could improve the classifier models to produce more useful and realistic patterns and rules that can further help high school advisors to guide applicants in choosing their academic career in an informed manner.

- Scientific research has not been carried out by scholars to show how many students switched disciplines from fields of study that are part of a lower cluster to ones that belong to a higher cluster and therefore the dataset lacks such information. Thus, the constructed classifier models in this research predict and classify these candidates to a higher cluster because they were built and trained without such information.
- The information on the classification of high schools into public and private and information on whether they are located in rural or urban areas is entirely missing from the data architecture. A dataset with such key features could be very substantial in constructing and training the classifier models that would certainly lead to better and higher accuracy.
- There are limited seating capacities for every field of study and priority is given to candidates who pick these disciplines as their first choices out of the five choices they have. The Kankor dataset lacks information on seating capacities of the disciplines. Moreover, it lacks the candidates' choices and the order they chose their five disciplines. Hence, the constructed models were not built with such information to consider the seating capacity as a factor and therefore classify some candidates with a higher score in Kankor to a higher cluster while they are in a lower cluster, and vice-versa.

Furthermore, prior to using high school marks for making discipline recommendations, there are several challenges which need to be addressed.

- The views and answers of university officials who were surveyed in every discipline vary widely. These biased and conflicting opinions of respondents in relation to the weighting of some subjects may have overshadowed a true weighting value.
- Some high school subjects play important roles in more than one discipline with the same weighting score. Therefore, the list of suggestions provided by the recommendation system may not always be in line with the actual talents of the tested student and thus a field of study for which the student would be qualified may not be recommended.

As solutions, this research proposes the followings for future work:

- Presently fields of study are not classified into sciences, social sciences and other streams/clusters. Rather than assigning weighting scores to high school subjects in

relation to individual fields of study, it will be more efficient to assign such weights in relation to streams. The classification of disciplines into streams/clusters and assignment of weights to high school subjects in relation to clusters of similar disciplines requires national level research and analysis with input from professionals from across all the disciplines to ensure quality data and inclusive analysis.

- Next step would be to construct profiles for disciplines, construct profiles for students and finally to use content-based, collaborative and other types of recommendations to further equip high school advisors, so they can provide constructive guidance for candidates on their academic career.

This research opens up a whole new range of possibilities in a wide range of areas and establishes the ground work in offering specialized studies at high schools; classification of disciplines into streams; helping officials and policy advocates to offer specialized Kankor for the identified streams; bringing scholars and policy advocates from both Ministry of Education and Ministry of Higher Education closer together for collaboration to identify key features of data which are essential in data mining applications in order to make a better architecture to record the data; evaluation of candidates' five-choice to see if they are selected methodically; assisting university students in selection of their major studies after two years of general studies; and finally, the proposed techniques make it possible to institutionalize advisory organizations inside or outside of educational institutions to guide students concerning their academic careers.

REFERENCES

ACM RecSys. 2017. “Submission Statistics.” Submission Statistics. June 7, 2017. https://recsys.acm.org/statistics/.

Adomavicius, Gediminas, and Alexander Tuzhilin. 2005. “Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions.” *IEEE Trans. on Knowl. and Data Eng.* 17 (6): 734–749. https://doi.org/10.1109/TKDE.2005.99.

Aggarwal, Charu C. 2015. *Data Mining: The Textbook*. 2015 edition. New York, NY: Springer.

Allington, Matt. 2015. *Learn to Write DAX: A Practical Guide to Learning Power Pivot for Excel and Power BI*. Uniontown, OH: Holy Macro! Books.

———. 2016. “Self Referencing Tables in Power Query.” *Excelerator BI* (blog). June 21, 2016. https://exceleratorbi.com.au/self-referencing-tables-power-query/.

Asadzadeh, Maedeh. 2014. “Education in Turkey.” World Education Services. WENR. August 13, 2014. https://wenr.wes.org/events/event/education-in-turkey.

———. 2015. “Education in Iran.” World Education Services. WENR. September 18, 2015. http://knowledge.wes.org/Archive-Education-in-Iran.html.

Baker, Ryan S. J. d, and Kalina Yacef. 2009. “The State of Educational Data Mining in 2009: A Review and Future Visions.” *JEDM - Journal of Educational Data Mining* 1 (1): 3–17.

Bondarenko, Ivan. 2015. “Multiple Replacements of Words in Power Query.” *Ivan Bond’s Blog* (blog). April 16, 2015. https://bondarenkoivan.wordpress.com/2015/04/16/multiple-replacements-of-words-in-power-query/.

Buyya, Rajkumar, Rodrigo N. Calheiros, and Amir Vahid Dastjerdi, eds. 2016. *Big Data: Principles and Paradigms*. 1 edition. Cambridge, MA: Morgan Kaufmann.

Castellano, Emilio J., Manuel J. Barranco, and Luis Martínez. 2011. “Academic Orientation Supported by Hybrid Intelligent Decision Support System.” https://doi.org/10.5772/17019.

Chair, Joseph Beck, Ryan Baker, Albert Corbett, Judy Kay, Diane Litman, Tanja Mitrovic, and Steve Ritter. 2004. “Workshop on Analyzing Student-Tutor Interaction Logs to Improve Educational Outcomes.” In *Intelligent Tutoring Systems*, 909–909. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-30139-4_121.

Clark, Nick. 2006. “Education in India.” World Education Services. WENR. February 1, 2006. http://wenr.wes.org/2006/02/wenr-feb-2006-education-in-india.

Collie, Rob, and Avichal Singh. 2016. *Power Pivot and Power BI: The Excel User’s Guide to DAX, Power Query, Power BI & Power Pivot in Excel 2010-2016*. 2 edition. Uniontown, OH: Holy Macro! Books.

Cook, Darren. 2015. *Introduction To Manipulating Data Programmatically In Microsoft Excel With VBA*. 2 edition. Abstractive Media.

CSO, Central Statistics Organization. 2014. “2013-2014 - Central Statistics Organization.” 2015 2014. http://cso.gov.af/en/page/1500/4722/2013-2014.

Dahan, Haim, Shahar Cohen, Lior Rokach, and Oded Maimon. 2014. *Proactive Data Mining with Decision Trees*. 2014 edition. New York: Springer.

Das, T. K., and P. Mohan Kumar. 2013. “BIG Data Analytics: A Framework for Unstructured Data Analysis.” *International Journal of Engineering and Technology (IJET)* 5 (1): 153–156.

Dean, Jeffrey, and Sanjay Ghemawat. 2008. “MapReduce: Simplified Data Processing on Large Clusters.” *Commun. ACM* 51 (1): 107–113. https://doi.org/10.1145/1327452.1327492.

Đurović, Gordan. 2016. “Educational Recommender Systems.”

Đurović, Gordan, Martina Holenko Dlab, and Nataša Hoić-Božić. 2016. “Using Recommender System to Motivate Electrical Engineering Course Students to Use Web 2.0 Tools in Their Learning Process.” In *International Conference on E-Learning*, 16:189. http://www.elearning-conf.eu/docs/cp16/paper-29.pdf.

EDM community website. 2017. “Home.” International Educational Data Mining Society. 2017. http://www.educationaldatamining.org.

Excel team. 2015. “Integrating Power Query Technology in Excel 2016.” *Office Blogs* (blog). September 10, 2015. https://blogs.office.com/en-us/2015/09/10/integrating-power-query-technology-in-excel-2016/.

Feldmann, Imke. 2015. “Tip for Parameter Tables in Power Query and Power BI.” *The BIccountant* (blog). October 31, 2015. http://www.thebiccountant.com/2015/10/31/tip-for-parameter-tables-in-power-query-and-power-bi/.

———. 2016. “Multiple Replacements or Translations in Power BI and Power Query.” *The BIccountant* (blog). May 22, 2016. http://www.thebiccountant.com/2016/05/22/multiple-replacements-in-power-bi-and-power-query/.

Freedman, Liz. 2013. “The Developmental Disconnect in Choosing a Major: Why Institutions Should Prohibit Choice until Second Year.” *The Mentor. Penn State Division of Undergraduate Studies*, June. https://dus.psu.edu/mentor/2013/06/disconnect-choosing-major/.

Goebel, Michael, and Le Gruenwald. 1999. “A Survey of Data Mining and Knowledge Discovery Software Tools.” *SIGKDD Explor. Newsl.* 1 (1): 20–33. https://doi.org/10.1145/846170.846172.

Gordon, Virginia N., Wesley R. Habley, and Thomas J. Grites, eds. 2008. *Academic Advising: A Comprehensive Handbook*. 2 edition. San Francisco, CA: Jossey-Bass.

Gu, Mini. 2016a. “The Gaokao.” World Education Services. WENR. May 2, 2016. https://wenr.wes.org/2016/05/the-gaokao-history-reform-and-international-significance-of-chinas-national-college-entrance-examination.

———. 2016b. “The Gaokao: Gateway to Higher Education in China.” World Education Services. Webinar. September 16, 2016. https://wenr.wes.org/events/event/14956.

Han, Jiawei, Micheline Kamber, and Jian Pei. 2011. *Data Mining: Concepts and Techniques, Third Edition*. 3 edition. Haryana, India; Burlington, MA: Morgan Kaufmann.

Hartshorn, Scott. 2016. *Machine Learning with Random Forests and Decision Trees: A Mostly Intuitive Guide, but Also Some Python*.

Ho, Robert. 2013. *Handbook of Univariate and Multivariate Data Analysis with IBM SPSS, Second Edition*. 2 edition. Boca Raton: Chapman and Hall/CRC.

Hurwitz, Judith, Alan Nugent, Fern Halper, and Marcia Kaufman. 2013. *Big Data for Dummies*. 1 edition. Hoboken, NJ: For Dummies.

Hussain, Tauqeer, and Adib Farah. 2015. “Big Data -Tools and Technologies.” *Computer Science and Education in Computer Science* 11 (1): 132–39.

InternetLiveStats. 2017. “One Second.” Internet Live Stats. May 18, 2017. http://www.internetlivestats.com/.

Jannach, Dietmar, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. 2010. *Recommender Systems: An Introduction*. 1 edition. New York: Cambridge University Press.

Karau, Holden, Andy Konwinski, Patrick Wendell, and Matei Zaharia. 2015. *Learning Spark: Lightning-Fast Big Data Analysis*. 1 edition. Beijing ; Sebastopol: O’Reilly Media.

KDnuggets. 2016. “Polls.” Industries/Fields Where You Applied Analytics, Data Mining, Data Science in 2015. January 2016. http://www.kdnuggets.com/polls/2016/industries-analytics-data-mining-data-science.html.

Kevin, Kamal, and WES Staff. 2017. “Education in Turkey.” World Education Services. WENR. April 4, 2017. http://wenr.wes.org/2017/04/education-in-turkey.

Kotu, Vijay, and Bala Deshpande. 2014a. *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*. 1 edition. Morgan Kaufmann.

———. 2014b. *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*. 1 edition. Waltham, MA: Morgan Kaufmann.

Kumbhare, Trupti A., and Santosh V. Chobe. 2014. “An Overview of Association Rule Mining Algorithms.” *International Journal of Computer Science and Information Technologies* 5 (1): 927–30.

Larose, Daniel T., and Chantal D. Larose. 2015. *Data Mining and Predictive Analytics*. 2 edition. Hoboken, New Jersey: Wiley.

Leskovec, Jure, Anand Rajaraman, and Jeffrey David Ullman. 2014. *Mining of Massive Datasets*. 2 edition. Cambridge: Cambridge University Press.

Linden, Greg, Brent Smith, and Jeremy York. 2003. “Amazon.Com Recommendations: Item-to-Item Collaborative Filtering.” *IEEE Internet Computing* 7 (1): 76–80. https://doi.org/10.1109/MIC.2003.1167344.

Linoff, Gordon S. 2015. *Data Analysis Using SQL and Excel*. 2 edition. Wiley.

Lowe, Anna, and Michael Toney. 2000. “Academic Advising: Views of the Givers and Takers.” *Journal of College Student Retention: Research, Theory & Practice* 2 (2): 93–108. https://doi.org/10.2190/D5FD-D0P8-N7Q2-7DQ1.

MacLennan, Jamie, ZhaoHui Tang, and Bogdan Crivat. 2008. *Data Mining with Microsoft SQL Server 2008*. 1 edition. Indianapolis, IN: Wiley.

Maghsoudi, Behroz, Sadeq Soleimani, Ali Amiri, and Mohsen Afsharchy. 2012. “Improving education quality in e-learning systems using data mining.” *Journal of Technology Education* 6 (4): 277–86.

Mandegar Daily Newspaper. 2015. “Ministry of Higher Education: The complaints of Kankor participants are considered.” News. *Ministry of Higher Education: The complaints of Kankor participants are considered* (blog). April 4, 2015. http://mandegardaily.com/reports/وزارت-تحصیلات-عالی-به-شکایات-داوطلبان/.

Merceron, Agathe, and Kalina Yacef. 2008. “Interestingness Measures for Association Rules in Educational Data.” In *Educational Data Mining 2008*. Montreal, Canada.

Michael, Rachel. 2016. “Education in China.” World Education Services. WENR. March 7, 2016. https://wenr.wes.org/2016/03/education-in-china-2.

———. 2017. “China’s Huikao, the Advanced Placement Test, and Shanghai Variations.” World Education Services. WENR. August 16, 2017. https://wenr.wes.org/2017/08/the-huikao-the-advanced-placement-test-and-shanghai-variations-a-look-at-chinas-high-school-exams.

Microsoft Research. 2014. “Fuzzy Lookup Add-In for Excel.” Microsoft Download Center. November 6, 2014. https://www.microsoft.com/en-us/download/details.aspx?id=15011.

Mihael, Ari. 2014. “Education in Turkey: Exploring a Top Emerging Market.” World Education Services. WENR. February 4, 2014. https://wenr.wes.org/events/event/education-in-turkey-exploring-a-top-emerging-market.

Mikut, Ralf, and Markus Reischl. 2011. “Data Mining Tools.” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1 (5): 431–43. https://doi.org/10.1002/widm.24.

Montaner, Miquel, Beatriz López, and Josep Lluís de la Rosa. 2003. “A Taxonomy of Recommender Agents on the Internet.” *Artificial Intelligence Review* 19 (4): 285–330. https://doi.org/10.1023/A:1022850703159.

Mostow, Jack, and Joseph Beck. 2006. “Some Useful Tactics to Modify, Map and Mine Data from Intelligent Tutors.” *Natural Language Engineering* 12 (02): 195. https://doi.org/10.1017/S1351324906004153.

Ng, Annalyn. 2016. “Tutorials, Overviews.” Association Rules and the Apriori Algorithm: A Tutorial. April 14, 2016. http://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html.

Osborne, Jason W. 2012. *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data*. 1 edition. Thousand Oaks, Calif: SAGE Publications, Inc.

Padilla, Danilo. n.d. “Enhancement of Literacy in Afghanistan.” Education. Enhancement of Literacy in Afghanistan (ELA) Program. http://www.unesco.org/new/en/kabul/education/enhancement-of-literacy-in-afghanistan-ela-program/.

Peck, Roxy, and Jay L. Devore. 2011. *Statistics: The Exploration & Analysis of Data*. 7 edition. Australia; United States: Brooks / Cole.

Piatetsky, Gregory, and KDnuggets. 2014. “CRISP-DM, Still the Top Methodology for Analytics, Data Mining, or Data Science Projects.” October 28, 2014. http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html.

Prasad, Y. Lakshmi. 2016. *Big Data Analytics Made Easy*. 1 edition. Notion Press, Inc.

Puls, Ken. 2015. “Clean WhiteSpace in Power Query.” *The Ken Puls (Excelguru) Blog* (blog). October 8, 2015. https://www.excelguru.ca/blog/2015/10/08/clean-whitespace-in-powerquery/.

Puls, Ken, and Miguel Escobar. 2015. *M Is for (Data) Monkey: A Guide to the M Language in Excel Power Query*. Uniontown, OH: Holy Macro! Books.

Reynolds, Vince. 2016. *Big Data For Beginners: Understanding SMART Big Data, Data Mining & Data Analytics For Improved Business Performance, Life Decisions & More!* CreateSpace Independent Publishing Platform.

Ricci, Francesco, Lior Rokach, Bracha Shapira, and Paul B. Kantor, eds. 2010. *Recommender Systems Handbook*. 2011 edition. New York: Springer.

Richard L., Miller, and Irons Jessica G. 2014. *Academic Advising: A Handbook for Advisors and Students Volume 1: Models, Students, Topics, and Issues*. Vol. 1. 2 vols. Society for the Teaching of Psychology. http://teachpsych.org/ebooks/academic-advising-2014-vol1.

Richert, Willi, and Luis Pedro Coelho. 2013. *Building Machine Learning Systems with Python*. Birmingham, UK: Packt Publishing.

Romero, C., and S. Ventura. 2007. "Educational Data Mining: A Survey from 1995 to 2005." *Expert Systems with Applications* 33 (1): 135–46. https://doi.org/10.1016/j.eswa.2006.04.005.

———. 2010. "Educational Data Mining: A Review of the State of the Art." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 40 (6): 601–18. https://doi.org/10.1109/TSMCC.2010.2053532.

Romero, Cristobal, Sebastian Ventura, Mykola Pechenizkiy, and Ryan S. J. d Baker, eds. 2010. *Handbook of Educational Data Mining*. 1 edition. Boca Raton: CRC Press.

Sagiroglu, Seref, and Duygu Sinanc. 2013. "Big Data: A Review." In *2013 International Conference on Collaboration Technologies and Systems (CTS)*, 42–47. IEEE. https://doi.org/10.1109/CTS.2013.6567202.

Santos, Olga C., and Jesus G. Boticario. 2001. *Educational Recommender Systems and Technologies: Practices and Challenges*. IGI Global. http://www.igi-global.com/book/educational-recommender-systems-technologies/55284.

Sayad, Saed. 2010. "Data Mining Map." An Introduction to Data Mining. 2017 2010. http://www.saedsayad.com/data_mining_map.htm.

———. 2011. *Real Time Data Mining*. Cambridge Ont: Self-Help Publishers.

———. 2017. "Data Mining Map." Association Rules. 2017. http://www.saedsayad.com/association_rules.htm.

———. n.d. "KNN Classification." KNN Classification. Accessed September 8, 2016. http://www.saedsayad.com/k_nearest_neighbors.htm.

Sedgwick, Robert. 2000. "Education in Post-Revolutionary Iran." World Education Services. WENR. May 1, 2000. https://wenr.wes.org/2000/05/ewenr-mayjune-2000-education-in-post-revolutionary-iran.

Sedgwick, Robert, and Xiao Chen. 2002. "Education in China." World Education Services. WENR. March 1, 2002. http://wenr.wes.org/2002/03/education-in-china.

Sherzad, Abdul Rahman. 2016. "Applicability of Educational Data Mining in Afghanistan: Opportunities and Challenges." In *Proceedings of the 9th International Conference on Educational Data Mining*, 634–35. Raleigh, NC, USA. http://www.educationaldatamining.org/EDM2016/proceedings/paper_1.pdf.

Thai-Nghe, Nguyen, Tomas Horvath, and Lars Schmidt-Thieme. 2011. "Context-Aware Factorization for Personalized Student's Task Recommendation." In *Proceedings of the International Workshop on Personalization Approaches in Learning Environments*, 732:13–18. http://www.academia.edu/download/30803583/PALE2011_proceedings.pdf#page=19.

The Power User. 2017. *BULK Replace Values in Power BI / Power Query*. YouTube. https://www.youtube.com/watch?v=MLrRlPh_ZFQ.

Tierney, Brendan. 2014. *Predictive Analytics Using Oracle Data Miner: Develop & Use Data Mining Models in Oracle Data Miner, SQL & PL/SQL*. 1 edition. New York: McGraw-Hill Education.

Webb, Chris. 2014. "Using List.Generate() to Make Multiple Replacements of Words In Text in Power Query." *Chris Webb's BI Blog* (blog). June 25, 2014. https://blog.crossjoin.co.uk/2014/06/25/using-list-generate-to-make-multiple-replacements-of-words-in-text-in-power-query/.

Webb, Christopher, and Crossjoin Consulting Limited. 2014. *Power Query for Power BI and Excel*. 1st ed. edition. Berkeley, CA: Apress.

WES Staff. 2003. "Turkey." World Education Services. WENR. November 1, 2003. https://wenr.wes.org/2003/11/wenr-novemberdecember-2003-turkey.

———. 2011. "Admitting Chinese Undergraduates." World Education Services. WENR. November 1, 2011. https://wenr.wes.org/2011/11/wenr-novemberdecember-2011-feature-admitting-chinese-undergraduates.

———. 2013. "Education in Iran." WENR. April 1, 2013. https://wenr.wes.org/2013/04/wenr-april-2013-an-overview-of-education-in-iran.

———. 2017. "Education in Iran." World Education Services. WENR. February 7, 2017. http://wenr.wes.org/2017/02/education-in-iran.

White, Tom. 2015. *Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale*. 4 edition. Beijing: O'Reilly Media.

Williams, Graham. 2011a. *Data Mining with Rattle and R*. 1 edition. Springer New York.

———. 2011b. *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery*. 2011 edition. New York: Springer.

Winston, Wayne. 2016. *Microsoft Excel Data Analysis and Business Modeling*. 5 edition. Redmond, Washington: Microsoft Press.

Witten, Ian H., Eibe Frank, and Mark A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition*. 3 edition. Burlington, MA: Morgan Kaufmann.

Wu, Xindong, and Vipin Kumar, eds. 2009. *The Top Ten Algorithms in Data Mining*. 1 edition. Boca Raton: Chapman and Hall/CRC.

Wu, Xindong, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, et al. 2008. "Top 10 Algorithms in Data Mining." *Knowledge and Information Systems* 14 (1): 1–37. https://doi.org/10.1007/s10115-007-0114-2.

Zaíane, Osmar R. 2002. "Building a Recommender Agent for E-Learning Systems." In *Computers in Education, 2002. Proceedings. International Conference on*, 55–59. IEEE. http://ieeexplore.ieee.org/abstract/document/1185862/.

Zhao, Yanchang. 2012. *R and Data Mining: Examples and Case Studies*. 1 edition. Amsterdam; Boston: Academic Press.