Applying of Data Mining and Statistical Techniques to Analyze the Impact of Socioeconomic Background on University Admission

A Case Study Using the Iranian Educational Data

Zur Erlangung des akademischen Grades eines

Doktors der Wirtschaftswissenschaften

(Dr. rer. pol.)

von der Fakultät für Wirtschaftswissenschaften

des Karlsruher Instituts für Technologie (KIT)

genehmigte

DISSERTATION

Von

M.Sc. Seyed Bagher Mirashrafi

Tag der Mündlichen Prüfung: 20. July 2016

Referent: Prof. Dr. Gholamreza Nakhaeizadeh

Korreferent: Prof. Dr. Georg Bol

Korreferent: Prof. Dr. Ebrahim Khodaie

Karlsruhe, 2016

Acknowledgment

It is never sufficient how much I thank God for what He bestowed upon me. It is my utmost trust in Him that I achieved my goals including this work.

Had it not been the precious suggestions of and valuable discussions with my thesis supervisor, Prof. Dr. Gholamreza Nakhaeizadeh, this work would have never been realized. I would really thank him for all the support and unconditional encouragement at all the moments during my research. His academic qualification and practical approach to research are brilliant.

Many thanks to my co-supervisor, Prof. Dr. Georg Bol. Indeed he is a very encouraging person and was always available for a discussion whenever I faced a problem during my stay at KIT Karlsruhe.

I would further take the opportunity of thanking Prof. Dr. Ebrahim Khodaei who motivated me by various discussions and comments. His visit from Iran attending my exam means a lot to me.

I am also very grateful to the doctoral committee and would like to thank Prof. Dr. Karl-Heinz Waldmann, and Prof. Dr. Martin Ruckes. Many thanks to our ex-chairman Prof. Dr. S. T. Rachev for his absolute support and recommendations.

I am also grateful to our ex-chairman Prof. Dr. Wolf-Dieter Heller for all what he did. Meanwhile I would like to be thankful to the head of our chair Prof. Dr. Melanie Schienle.

Thanks to Dr. Ehsan Jamali, Dr. Behroz Kavehei and Ms. Fatemeh Zarin at NOET, for all the discussions during my visit of Sanjesh center in Iran. Besides, thanks to the IT center of NOET in Iran for providing the datasets for this work.

I am also grateful to my colleagues Prof. Dr. Steffi Höse, Prof. Dr. Young

Shin Kim, Dr. Markus Höchstötter, Dr. Edward W. Sun, Dr. Dirk Krause, Dr. Ruzana Davoian, Dr. Abdolreza Nazemi, Mr. Ralf Stegmueller and especially thanks to Mrs. Theda Schmidt for the support through this work.

Special thanks to Dr. Anees ul Mehdi and Dr. Mojtaba Ebrahimi for their support in the development of my research project. Without their help, it would have been not this comfortable.

Last but not least, I would like to express my gratitude to my dear wife Shahla, my son Mahdi and my daughter Fatemeh for their unconditional love and patience. Thank you my dear brother Akbar, all the relatives and friends and all those who contributed in this work.

Seyed Bagher Mirashrafi

Karlsruhe, July 2014

Abstract

The goal of this thesis was to conduct a focused and in-depth comprehensive study of the impact of socioeconomic status of the Iranian Wide Entrance Examination (WEE) applicants' family on the educational achievement of their children. To reach this goal we used various statistical methods, like variance and regression analysis as well as the data mining techniques, among them, various kinds of decision trees and artificial neural networks.

The data used in this study belongs to the National Organization of Educational Testing (NOET) in Iran and was collected over five years 2005 to 2009 including more than five million observations and about 40 attributes. The original sources of the data are two questionnaires which were completed by the WEE-applicants. The questions are about their personal characteristics, their high school performance records and socioeconomic background of their parents. The quality of data was relatively good. The data understanding and data preparation phases took relatively a lot of time. For a large empirical study like our project this was expectable, however, from beginning.

Having the data over five years made it possible for us to construct classification and forecasting models for each year, separately. Furthermore it gave us, the opportunity to examine the stability of the constructed models over the time. Concretely, it was possible to test the stability of a certain model by using the data in succeeding years. To the best of our knowledge, when dealing with the Iranian educational data, there is no comprehensive study that takes into account such dynamic aspects. Such aspects are simply ignored.

The main application aspects of our study are:

- Knowing the impacts of socioeconomic background of the applicants' family on the educational achievement of their children could be very useful for policy makers of higher education and for the families of the applicants

- Having models which can predict the performance of the WEE-applicants

in advance can help them to be better prepared for the WEE

- Being aware about the dynamic aspects of the data makes it possible to decide either to use a constructed model without any change for the future or adapt the model by using the new data

Contents

List of Abbreviation

ANN	Artificial Neural Networks
ANOVA	Analysis of Variance
BCA	Bachelor of Computer Applications
CART	Classification and Regression Trees
CHAID	Chi-square Automatic Detection
CRISP	Cross-Industry Standard Process
CRUISE	Classification Rule with Unbiased Interaction Selection and Estimation
FACT	Fast Algorithm for Classification Trees
GLM	Generalized Linear Model
GPA	Grade Point Average
GUIDE	Generalized, Unbiased, Interaction Detection and Estimation
IEA	International Association for the Evaluation of Educational Achievement
LSD	Least Significant Difference
KDD	Knowledge Discovery in Databases
LOTUS	Logistic Tree with Unbiased Selection
ML	Machine Learning
MLE	Maximum Likelihood Estimation
NCDS	National Child Development Study

NLSCY	National Longitudinal Survey of Children and Youth
NOET	National Organization of Educational Testing
OLS	Ordinary Least Squares
PSE	Post Secondary Education
QUEST	Quick, Unbiased, Efficient, Statistical Tree
RMSE	Root Mean Squared Error
SSE	Sum of Squares Error
SSR	Sum of Squares Regression
SST	Sum of Squares Total
WEE	Wide Entrance Examination

List of Figures

List of Tables

1. Introduction

The theory of rational action steps focuses on educational transitions, such as transition from elementary to high school and from high school to university. The demand for higher education comes from applicants with different socioeconomic status and is influenced by individual, environmental, economic and social factors. The choice at each stage is highly dependent on profits and losses. In this respect the social classes differ: for lower classes, the cost of the transition from secondary education to higher education is far more than it is for the privileged classes, thus there are less applicants from lower classes who enter into universities. Evaluating the impact of different factors on the demand for higher education as well as the probability of admission to university programs could assist the government to make efficient decisions about admission policies.

Recent studies conducted in different countries reveal that social and family background have a significant impact on the academic performance of an applicant, on his/her future profession, as well as on economic, social, cultural and political status. In [62], it is shown that the importance of different parameters varies among different countries; in particular, there is a significant difference between results from the developed and those from the developing countries. In addition, it has been shown that parents' education and profession have a direct impact on a child's academic performance [45, 37]. However, all previous studies mostly evaluate the static aspects of such dependencies and/or focus mainly on students' grades. To the best of our knowledge, there is no work dealing with time-related variables when evaluating the effect of socioeconomic status on academic performance. Hence, a deeper analysis of the impact of these parameters on the university entrance exam scores is required.

In order to find a suitable model, we perform a comprehensive study based on the high volume data by the various new approaches and methodologies. We first examine the designed model in the static situation by using various statistical analysis and data mining algorithms.

Furthermore, in this study, contrary to previous approaches, we focus also on the dynamic aspects of the observations. As of dataset, we use the data of around six million applicants gathered over 5 years (2005-2009) from university Wide Entrance Examination (WEE) in Iran. WEE is an exam held annually[1] consisting of multiple choice questions. An overall score for each candidate is calculated from his performance in this exam. In recent years the academic performance of students in Iranian high schools is additionally considered for rankings along with the WEE result. The applicants are then ranked according to their overall scores. Applicants with higher ranks have a higher chance to enter into a desired field/university, while those with lower rankings should either attend a secondary choice of study program or once again participate in the next year exam. Since some fields such as engineering and medical science have higher demand, applicants should have considerably higher ranking to be admitted in these fields. Therefore, success in this exam is defined as having a high overall score and, as a result, a higher ranking. We extensively analyze the probability of success in this exam with respect to the wide range of individuals and social factors. In a nutshell, our goal is to determine whether the socioeconomic background of families is reflected in the performance of their children. To pursue our goal, we identify the following two main objectives:

- Analyzing the effect of the socioeconomic status of applicants on their overall score.
- Determining the effect of time-dependent variables on the socioeconomic status and therefore on the overall score of candidates.

1.1. Effects of Socioeconomic Status

In this work, we consider a combination of the two categories of variables which represent the socioeconomic factors. We study the effects of these variables on the academic performance of candidates in the entrance exam. As the basis of performance analysis we consider the total score of the candidates. We believe that our approach is a suitable way and the total score is an appropriate criterion for analyzing academic performance of high school students qualifying for entering into universities. The considered socioeconomic variables are include:

- Individual factors like gender, age and field of study at high school

[1] For more details see chapter 4.

- Family background like education, job and income of the parents
- Environmental factors like city and state of residence

1.2. Exploration of Dynamic Aspects

As already discussed in the previous section, several socioeconomic variables do affect the overall score of the candidates. We believe that the impact of these variables is time-dependent. In our study we examine whether such dynamic aspects exist. In case we observe such variables, we suggest methods based on dynamic data mining for handling them. We are not aware of any related work which considers dynamic aspects when studying the effect of socioeconomic status on candidates' performance. The theoretical model employed in this work is dynamic as it considers time (year of WEE) as an additional important factor/variable.

1.3. Theoretical Model

Figure 1.1 presents a pictorial representation of our model. Note that in our model, an applicant's grade is affected by individual, family background and environmental factors. Moreover, it is also influenced by the time/year of their exam.

For this study, the research hypotheses are as follows:

- An increase in the level of some constituent variables of the socioeconomic status of the applicants increases the academic performance of the candidates (positive effects).
- An increase in the level of some constituent variables of the socioeconomic status of the applicants decreases their academic performance (negative effects).
- The impact of the socioeconomic status on the academic performance of candidates is time depended.
- In all testing groups[2], the education of the mother compared to that of the father has a greater effect on the academic performance.

[2]We used data consisting of roughly six million WEE applicants' information from three different main testing groups (Mathematics and Physics, Empirical Sciences, and Human Sciences). See Chapter 4.

In order to test our hypothesis, we apply various statistical methods and data mining algorithms such as decision tree, logistic regression, artificial neural networks, and regression. This is in contrast to the traditional studies where only a data sample is considered for analysis. We additionally compare the performance of the different algorithms considered in order to determine the most appropriate one.

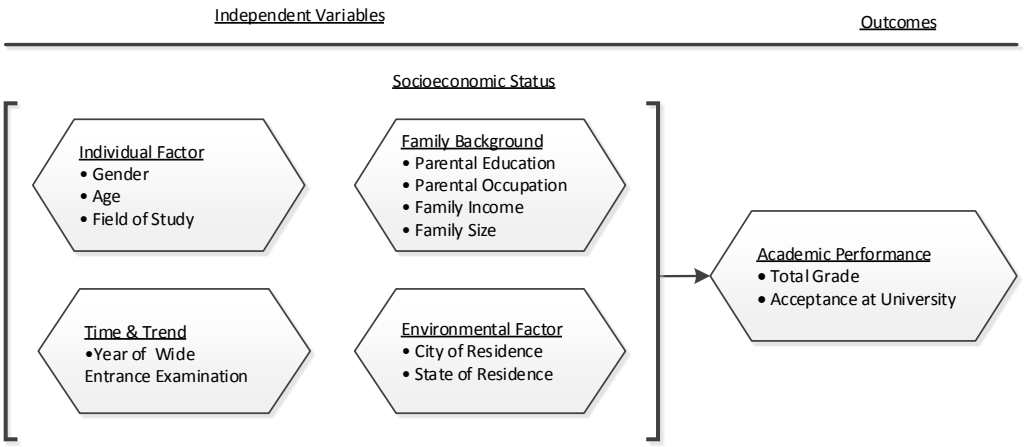

Figure 1.1: Framework of the Selected Factors with an Effect on Applicants' Academic Performance

1.4. Outline of the Thesis

This thesis is organized into three parts comprised in total of seven chapters. Part I presents theoretical aspects and is composed of the first three chapters. In the first chapter, we present an introduction to our work. Furthermore, it provides an introduction to WEE and the employed database. The second chapter presents a survey of the related works. The third chapter gives an overview of the applied statistical and data mining methods.

Chapter 4 to 6 are included in Part II. Chapter 4 briefly introduces the National Organization of Educational Testing (NOET) of Iran, and the data used, including dataset description, data selection, data quality and data cleaning. In the fifth and sixth chapters, the empirical analyses are discussed in detail and the analyses of different empirical experiments are presented. Chapter 5 is devoted to the investigation of static models by applying some statistical and data mining techniques, such as linear and logistic regression, analysis of variance, decision tree, and neural networks to different datasets. The results

are analyzed as well. Chapter 6 investigates the existence of dynamic aspects in our observations. The final part comprises Chapter 7 where we conclude this dissertation with an emphasis on our new findings and future works.

2. Related Works

This chapter presents an overview of several related works studying the influences of individual, environmental, and family background factors on the academic achievement of students. We categorize these work into two groups: 1) studies related to countries other than Iran; 2) (local) studies focusing on Iran only. Furthermore, the chapter presents some studies dealing with the application of data mining in research on education. After summarizing the results of these studies, we identify several issues yet to be investigated with which we later deal in the upcoming chapters.

2.1. Studies for Other Countries

We now discuss several studies on the influences of the aforementioned factors on the students' academic achievement. In this section we mainly focus on the studies covering countries other than Iran.

Social scientists have developed sophisticated models for educational attainment using different causal variables [8, 7, 25, 19]. Although, there are variations regarding race and sex, the same causal variables have been applied [32, 61].

Abesha et al. [31] conducted a wide study with 2116 university students in Ethiopia.They found that with regard to the interrelationships among academic self-efficacy, achievement motivation, and academic achievement, irrespective of students' gender, academic self-efficacy had a significant and positive direct effect on achievement motivation. In addition, it has a significant and positive mediated effect (i.e., through achievement motivation) on academic achievement. They also found that the parenting styles had a significant and positive direct effect on achievement motivation for female students, but not for male students.

Kodde and Ritzen [48] have done extensive research on the direct and indirect effects of parental education on the demand for higher education. Using multivariate methods and different confidence levels, they show that there is a

positive correlation between the level of parental education and their children's demands for higher education.

Past research has indicated an academic achievement gap between genders, with boys ahead of girls. However, recently it is shown that the achievement gap has been narrowing. In some instances girls even have higher academic achievement than boys [15]. Additionally, studies show that females perform better in reading than males. However, males are found to outperform females in mathematics and science [21].

In a case study on higher education in Spain, Albert et al. [2] showed that the education level of parents, especially of the mother, and a good economic status of the father, increases the chance of a student entering into university. However they also found that males are less likely to enter universities than females.

A research study in Argentina in 1998, showed that students studying at public universities had family income of 90% above the state average [76]. Meanwhile, parents of almost 50% of such students, had college education. Additionally, 70% of the students were from the richest families (comprising 30% of the total population), and only 11% were from below average families (50% of the total population).

During 1970-1994, the participation rates of applicants to university had increased for all income groups of society in the USA, but the problems for high-income and low-income groups, despite efforts, remained relatively constant. Gladieux and Swail argued that there exists a gap between high-income and low-income groups' access to universities. Additionally, the children of high-income groups have a better chance for entering into the best universities and academic disciplines [30].

Research has found that socioeconomic status, parental involvement, and family size are particularly important factors [59, 60]. Lochner & Belley [4] found that Post Secondary Education (PSE) attendance (and attendance at four-years PSE institutions) is positively related to parental income in the U.S., even after controlling for similar measures of family background and adolescent cognitive achievement. The effect of parental income PSE attendance in Canada is also positive, but substantially weaker.

The findings of Pedrosa et al. [71] indicate that students coming from a disadvantaged environment, in socioeconomic and educational terms, performed relatively better than those coming from higher socioeconomic and educational

strata. More interestingly, from an educational public policy perspective, students who came from public schools had a relatively better performance than those who had studied at private schools. On the other hand, the expansion of universities caused some social changes in middle and low class categories.

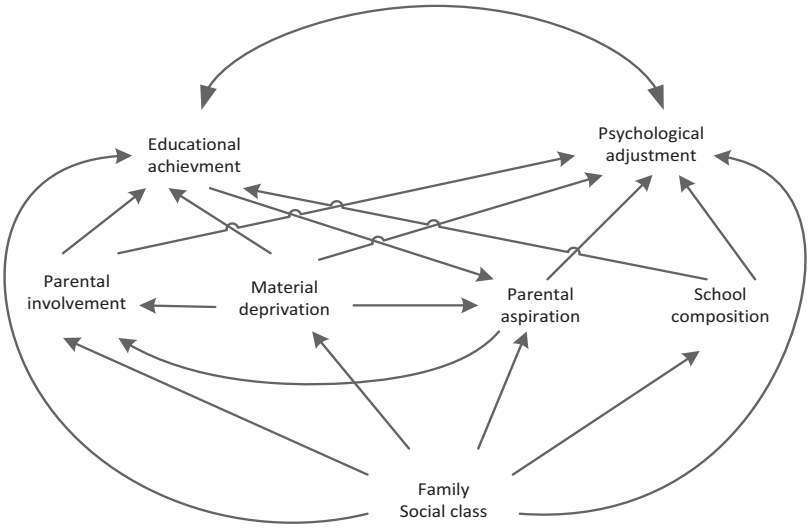

Figure 2.1: Sacker et al. [77] Model of the Relationship Between Family Social Class, and Pupil Achievement and Adjustment

Sacker and her colleagues set out to test the model shown in Figure 2.1. By using data from the National Child Development Study (NCDS), they examined how socioeconomic inequalities in students influence their educational achievement and psychosocial adjustment. Furthermore, they found that social class inequalities in educational achievement to be greater than inequalities in psychosocial adjustment.

Jimenez and Velasko [39] used the logarithmic models based on socioeconomic factors, such as income, occupation, education level of the parents, to demonstrate the effect of these factors on the demand for higher education. They concluded that students with a high level of socioeconomic status have opportunities of studying at better schools/universities. Consequently, such students have a higher probability of academic success.

Jeynes [38] argued that the socioeconomic conditions of a child are represented by a combination of educational level of parents, parental occupational status and family income level. As a result, these conditions are reflected in his/her academic performance. Moreover, he found that family size is associated with academic achievement. Students with fewer siblings are more likely

to perform better at schools/universities as they get more attention from their parents. A similar result is shown by Iacovov in [35].

Research shows that the socioeconomic and cultural status of the family, especially that of the parents, affect students during their academic career. For example, students from low income and lower social class families are often unable to continue their education. Probably students from such families are interested in entering the working world sooner to increase their contribution to the family income. Even when equal opportunities are provided, the social and cultural differences are extremely influential on students' choice of subjects and the duration of their study. Moreover, girls from families with high income, not only enter into high schools and colleges, but also continue with academic specialization (Psakharopoulos and Sanyal [73, 72]).

Sun et al. [87, 85, 86] analyzed a case study which included 636 students from 50 schools in the last year of primary education. They considered the type of school, the mean score of reading lessons during 3 years of schooling, and socioeconomic status as independent variables, whereas the score of reading in the 5th-grade was considered as a dependent variable. For this purpose they used various models for prediction. They found that the two-level model with students at the first level, and schools at the second-level of hierarchy was most appropriate. They further found that all input variables affected the scores of the 5th-grade students.

In [91] , Western and MacMillan assess Patrick Lynch's[58] research which used the zip code as the socioeconomic status. Despite the challenges, using zip-code is a simple and inexpensive but error-prone method. For instance, a person of higher socioeconomic status might be living in a neighborhood with people of lower status. Hence, the location or area of the residence is not a good variable to be considered as a socioeconomic factor. Consequently, they considered profession, education and income as the socioeconomic factors.

Several studies on the constituent variables of the socioeconomic status investigate the strength and weakness of these variables (c.f. [58], Kate MacDonald (1964), Dyvndan et al. (2000) and Anverona & Campbell (1997)). These studies used occupation and education of parents, and family income of the applicants as the constituent variables of socioeconomic status. A few studies considered only some of these variables, like occupation, income and/or education level, for determining the impact of socioeconomic status on the academic performance of the applicants (c.f. Lyvlakp (1974), Crest & Hawk (1986)).

However, later researchers (mainly after the work of Western and MacMillan in 2000) used all of these variables.

The effect of cognitive and non-cognitive factors on the academic performance of the students has been considered in some works. As for the non-cognitive factors, the status of capital such as social capital, economic capital and cultural capital are considered. Bourdieu[9] believes that children belonging to dominant socioeconomic backgrounds have a higher chance of entering into colleges/universities [41]. Le Thanh Khoi [44] argues that in the developing countries, since the educational infrastructure is usually not fully established and due to the limited opportunity for higher education, the socioeconomic status shall be considered as the main determinant of academic performance.

Socioeconomic status is a combined economical and sociological total measure of a person's work experience and of an individual or family's economic and social position in relation to others, based on income, education, and occupation. The achievement of students at graduate school from a variety of socioeconomic trajectories, is determined either by their parents' education and occupation or by their own occupational histories (Some students delay higher education in order to earn and save money, gain professional experience, or support their families. This in general affects their academic performance due to factors like age, study gap, etc.). Thus socioeconomic background is a largely "invisible" but important factor that influences students' mentoring needs.

In order to explain the PSE or to find the impact of family background on PSE participation, Lefebvre and Merrigang, [53] had a survey on Canadian youth aged 18 to 21 in 2005. Their research was based on National Longitudinal Survey of Children and Youth (NLSCY) statistics during 1996-2005. They found that a period of unemployment by a parent reduces the probability of PSE attendance. In other words, parental income has a positive effect on education achievement in the sense that a low family income reduces the chances of attaining PSE.

Several studies in the United States found income to have a small impact on educational attainment. Some literature considers ability, parental education and behavior variables in the regression model ([14], [17]). In such studies, it was found that regardless of gender, a person's level of knowledge gained through education is highly dependent on factors like marital status, the size

of the residential community, math proficiency, and hyperactivity. However, factors like family structure, immigration status, reading scores, health, and aggressive behavior impact females educational performance mainly [89, 34]. That means, females are affected by more factors than males. Further, it is worth mentioning that these factors have a higher impact on youths from low-income families in contrast to youths from middle and high-income families.

In order to accurately identify students from low socioeconomic backgrounds, measures based upon the characteristics of individual students are developed. Western et al. [91] findings suggest that individual-based measures relating to the parental job and parental education at the time when the student was in high school, are appropriate for both recent graduates and older students. These characteristics represent the family's socioeconomic status for the students completing their secondary schooling.

In a study in 2002, Buchmann [13] reviewed the measurement of family background and examined the methods used to assess the impact of family background factors on educational outcomes in several international studies. He mainly focused on the surveys conducted by the International Association for the Evaluation of Educational Achievement (IEA). We summarize his reviews in combination with other investigations during 1973 to 2012 in Tables A.1 to A.6. It gives an overview of more than 70 studies in different developed and developing countries (more than 50 countries).

Marjoribanks [60] argued that research on dynamics of socioeconomic background and achievement-related variables may lead to an enriched understanding of social inequality in students' educational outcomes.

Educational Data Mining (EDM)

One of the biggest challenges that educational institutes face is the improvisation of the quality of managerial decisions using educational data bases[57]. Due to the huge amount of data, data mining[1] enables such organizations to uncover and understand hidden patterns in such data-sets using their current reporting capabilities. EDM is an emerging discipline, concerned with developing methods for exploring the unique and increasingly large scale educational data which is introduced by International EDM Society[2]. Such techniques can be used to extract meaningful knowledge and useful information from the data[55].

[1] Data Mining will be explained later in Chapter 3.
[2] http://www.educationaldatamining.org/

Kumar and Pal [6] investigate an experimental methodology on students' records. They argue that performance in higher education in India is influenced by many factors such as grade in senior secondary education, place of residence, grade in BCA (Bachelor of Computer Applications) examination, mother's qualification, family income, and students' family status.

Parack and et al. [70] discuss the application of data mining in education for student profiling and grouping. They argued that in academic fields, data mining could be very useful in discovering valuable information which can be used for profiling students based on their academic records.

Erdogan and Timor [22] concentrated on the application of data mining in an educational environment. They used the cluster analysis and k-mean algorithm techniques for finding the relationship between students' results on university entrance examinations and their success. They found that the use of data mining techniques in education might provide more varied and significant findings, and may lead to an increase in the quality of education.

Kumar and Chanda [49] in their research used data mining techniques to extract meaningful knowledge and useful information from huge educational databases. They found that the application of data mining brings a lot of benefits to higher education learning institutions. They recommend the application of these techniques for the optimization of resources and the prediction of retainment of faculties in the university.

Luan [56] in her case study on higher education (which included 15000 students) by data mining techniques found that younger students with more privileged socioeconomic background often took high credit courses and graduated quickly. Moreover, she argued that data mining is a powerful analytical tool that enables educational institutions to better allocate resources and staff, and additionally can improve the effectiveness of alumni development.

In [81], the authors showed the use of a fine-tuned data mining approach for providing effective monitoring tools for faculty performance. They concluded that the methodology used for extracting useful patterns from the institutional or educational database is able to extract certain unidentified trends in faculty performance.

Susnea [88] in her study found that using classification algorithm can lead to discovering relevant knowledge contained in an educational database which can be used for providing feedback to learners in the educational process. Later she used the artificial neural network method in her studies since it has a better

computational performance. She found that this method can have the best configuration from the error point of view. She also noticed that in most of the cases, the radial basis function (RBF) network has error rates lower than the multilayer perceptron (MLP) network.

The above mentioned studies show that data mining is an adequate approach for analyzing educational databases.

2.2. Studies Related to the Iranian Higher Education

In this section, we present a review of existing research on the Iranian education system.

The impact of social stratification on educational opportunities is investigated by Mohammadi, [65]. Using the data from the university of Urmia, he concluded that there is a direct relationship between social stratification and the possibility of individual achievement together with the access to higher education. In other words, a person from a higher rank in the hierarchy of social stratification has a better chance of access to higher education. Further he found that personal attempt, family support, and advice of family and friends have a motivating effect on the success rate of people who have studied at the university.

Sarandy [78] investigated extensively the effects of several factors influencing the studying quality of the students of University of Tabriz. He concluded that occupation and education of parents are significantly associated with academic success. Moreover, he found that there is a significant relationship between marital status, quota[3] and educational attainment.

In [43], it is shown that the occupation of the father and the education of the mother are the best variables to predict educational achievement of students at school. However, the same is not observed for the students of high school at different levels of education.

For the years from 2001 to 2009, a comparison of the effects of socioeconomic factors on the demand for admission to higher education was intensively researched by Gharun [27]. She tried to estimate the demand for higher education according to gender, age and testing groups[4]. Furthermore, she studied

[3]The quota is some conditions or characteristics of WEE applicants for entering into a university in Iran.

[4]Testing groups are described in Chapter 4.

the effect of socioeconomic variables on the demand for higher education. For this work, unlike the usual approach, she used the number of nationwide examination applicants as a societal demand for education. She concluded that the level of parental education, assuming all the other conditions to be constant, has an influence on an applicant's entrance into university. Furthermore, the chance of entering into a university is less for applicants coming from large families and/or for older age groups.

In Bourdie's Theory, three notions of capital are presented: *1) economic capital*, which is immediately and directly convertible into money and may be institutionalized in the form of property rights *2) cultural capital*, which is convertible, on certain conditions, into economic capital and may be institutionalized in the form of educational qualifications *3) social capital*, made up of social obligations, which is convertible, in certain conditions, into economic capital and may be institutionalized in the form of a little of nobility [10]. Considering these notions, Noghani [69] examined the impact of cultural capital inequality in educational opportunity for pre-university students in higher education achievement. He found that the family's cultural-economic support is important in academic achievement. Furthermore, he concluded that social capital, economic capital and cultural capital have a positive and significant contribution on the probability of acceptance in a university and, as a result, on the total grade of the applicants.

In [75], Roshan and Salehi conducted research at the Institute of Research and Planning in Higher Education. In their case study in order to determine the socioeconomic conditions and factors, they focused on the family and individual factors such as gender, age, the number of family members, marital status, field of study, employment, Grade Point Average (GPA) of high school (diploma average), and type of diploma, based on a case study. They identify a pattern of three factors (parental job, parental education and student families' income) as the socioeconomic status of students influencing their educational achievement. They found that these factors cannot be replaced by each other. For example, a person with a high income and better education, does not necessarily have a high position job. Similarly, a person with a high position job does not necessarily have a higher income.

A significant difference between the scores of male and female in the Empirical Sciences and the Human Sciences applicants has been shown by Jamali [36] for the applicants of the nation wide exam of year 2008. He found that the

status of each of the constituent socioeconomic variables, has an effect on the applicants‘ achievement. The magnitude of these effects, however, varies for different testing groups. Further, he concluded that the families with better socioeconomic status, are propelling their children to study in the Mathematics and Physics group, and if not entering into this group, then in to the Empirical Sciences group.

Acceptance at universities is one of the possible ways of obtaining a better job and other economic opportunities. Acquiring a university degree may increase the chance of a change of social life from a lower class to a higher class. As mentioned, therefore, there is a variety of theories in literature on the effects of the family background on educational attainment. Khodaei [46] shows that parental education has positive effects on the children's success, from an educational point of view. Moreover, Jamali [36] shows that the increase of parent's educational level and the father's job causes the increase of educational performance.

According to the past researches, due to the field of socioeconomic status variables and their impact on academic performance, the roles of each of these variables vary in different studies [37]. Most of these variables, especially in developing countries, already obtain/derive from developed countries consideration. Despite of all of these researches, to the best of our knowledge, there is no study in the literature dealing with the impact of these variables on the entrance examination scores by data mining techniques. Additionally, there has not been much work on considering of dynamic aspects in observations. Our work is an attempt to fill this gaps.

3. Methodology

In this work we present several empirical analyses which are based on various statistical analysis and data mining techniques. This chapter is devoted to an overview of the applied statistical and data mining methods, and algorithms. We first present an overview of some of the statistical methods which includes analysis of variance, linear regression and logistic regression. For instance, descriptive statistics is used for the family and individual factors for the data of all the considered years (2005-2009). In order to find a prediction model for the total grade of candidates, we used linear regression analysis as a inferential statistics method. In Section 3.2, we provide some preliminaries on data mining and data mining process. We then provide an overview of data mining tasks as well as discuss several applied data mining techniques and algorithms such as decision trees and neural networks. Some criteria for measuring the performances and accuracy of data mining algorithms are described in this section as well.

3.1. Statistical Methods

3.1.1. Analysis of Variance

In order to find the effects of independent factors on the target variable, we use the Analysis of Variance (ANOVA) developed by R. A. Fisher (for a detailed description of the method, see D. C. Montgomery [66]). As we already mentioned, the total grade of WEE of applicants is treated as the target variable. The ANOVA is based on the variance of the target variable according to the different levels of the independent factors. In Chapter 1 we have seen that for our work, socioeconomic factors such as individual, environmental, and family background factors are taken as the independent variables. In the ANOVA setting, the observed variance of the target variable is partitioned into different components which are attributable to different sources of variation.

In the analysis of variance we have a many treatments or (different) levels of a single factor that we want to compare. Suppose y_{ij} represents the jth

observation taken under factor level or treatment i. The observational matrix Y is the general case of the data with a rows and n columns. The observations from an experiment are described as a model:

$$y_{ij} = \mu_i + \varepsilon_{ij} \; ; \quad i = 1, \, 2, \, \cdots , \, a \; ; \quad j = 1, \, 2, \, \cdots , \, n \tag{3.1}$$

where μ_i is the mean of the ith factor level, and ε_{ij} is a random error component as having mean zero. Hence $E(y_{ij}) = \mu_i$. In order to analyze the effect of the factor, an alternative way to write a model for the data is to define $\mu_i = \mu + \tau_i$; $i = 1, \, 2, \, \cdots , \, a$. Equation 3.1 thus becomes

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \; ; \quad i = 1, \, 2, \, \cdots , \, a \; ; \quad j = 1, \, 2, \, \cdots , \, n \tag{3.2}$$

where μ is a parameter common to all treatments (*overall mean*), and τ_i is a parameter unique to the ith treatment (*ith treatment effect*). Equation 3.2 is called the effect model, also known as a *linear statistical model.* Sometimes this equation is also called the *one-way* or *single-factor* analysis of variance. In this model ε_{ij}'s are assumed to be normally and independently distributed random variables with mean zero and variance σ^2 (which is constant for all levels of the factor). Consequently $y_{ij} \sim N(\mu + \tau_i \, , \sigma^2)$.

The statistical model (Equation 3.2) describes two different situations (*fixed effects model, random effects*). In the fixed model, the a treatments are specifically chosen by experimenter, whereas in the random effects model, the a treatments could be a random sample from a larger population of treatments. Note that for our set up we use the fixed effects model for the analysis of variance where family background, individual and environmental factors are the fixed effect factors.

For testing the equality of the a treatment means, the appropriate hypotheses are

$$\begin{aligned} &H_0 : \mu_1 = \mu_2 = \cdots = \mu_a \\ &H_1 : \mu_i \neq \mu_k \text{ for at least one pair}(i, k) \end{aligned} \tag{3.3}$$

The ith treatment means that μ_i includes two components μ (overall mean) and τ_i (treatment effects) with $\mu_i = \mu + \tau_i$. The overall mean μ is defined by $\sum_{i=1}^{a} \mu_i / a$, and thus $\sum_{i=1}^{a} \tau_i = 0$. Consequently, the above hypotheses is in terms of the treatments effects τ_i and thus we can test the ith treatment effects by

the following hypothesis

$$\begin{aligned} H_0 &: \tau_1 = \tau_2 = \cdots = \tau_a = 0 \\ H_1 &: \tau_i \neq 0 \ \text{ for at least one } i \end{aligned} \tag{3.4}$$

The appropriate procedure for testing the above hypotheses is derived from a partitioning of total variability (the total sum of squares) into its component parts. The total correct sum of squares (SS_T) can be written as

$$\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \overline{y}_{..})^2 = n\sum_{i=1}^{a}(\overline{y}_{i.} - \overline{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \overline{y}_{i.})^2 \tag{3.5}$$

or

$$SS_T = SS_{Treatment} + SS_E$$

$$\text{where } \ \overline{y}_{..} = \frac{1}{an}\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij} \ ; \ \overline{y}_{i.} = \frac{1}{n}\sum_{j=1}^{n} y_{ij} \ .$$

Then the *F-test* statistic for the testing the above hypotheses is defined by

$$F_0 = \frac{SS_{Treatment}/(a-1)}{SS_E/(N-a)} = \frac{MS_{Treatment}}{MS_E} \sim F_{a-1,\, N-a} \tag{3.6}$$

where $N = an$ is the total number of observations.

Therefore, we reject null hypotheses and conclude that there are differences in the treatment if $F_0 > F_{\alpha,\, a-1,\, N-a}$ where α is the significance level.

Equation 3.5 as a fundamental ANOVA identity provides us with two estimates of σ^2, one based on the inherent variability *within* treatments and the other based on the variability *between* treatments. These two estimates should be very similar if there are no differences in the treatment means. The estimators for the parameters in the *single-factor* model (Equation 3.2) are given by

$$\begin{aligned} \hat{\mu} &= \overline{y}_{..} \\ \hat{\tau}_i &= \overline{y}_{i.} - \overline{y}_{..} \quad ; \quad i = 1,\, 2,\, \cdots,\, a \end{aligned} \tag{3.7}$$

Already, in practical situations we wish to compare all pairs of a treatment means and it is applied only after the *F-test* in the ANOVA is significant at 5 percent. For this purpose the appropriate hypotheses are

$$\begin{aligned} H_0 &: \mu_i = \mu_j \\ H_1 &: \mu_i \neq \mu_j \,\text{for all}\, i \neq j \end{aligned} \tag{3.8}$$

The *Fisher* Least Significant Difference (LSD) method is used for testing these hypotheses. Also, this procedure uses the *t-statistic* for testing $H_0 : \mu_i = \mu_j$ by $t_0 = \frac{\overline{y}_{i.} - \overline{y}_{j.}}{\sqrt{2MS_E/n}}$.

Then, we reject H_0 and conclude that the pair of means μ_i and μ_j would be declared significantly different if

$$|t_0| > t_{\alpha/2\,,\,N-a} \quad or \quad \left|\overline{y}_{i.} - \overline{y}_{j.}\right| > t_{\alpha/2\,,\,N-a}\sqrt{2MS_E/n} \tag{3.9}$$

which in the quantity $LSD = t_{\alpha/2\,,\,N-m}\sqrt{2MS_E/n}$ is called the *least significant difference*.

As we already mentioned, in this work we study the effects of two or more factors (independent variables). For this purpose, the factorial designs are most efficient for this type of experiment. The two-factor factorial design is the simplest type of factorial design which involves only two factors or treatments A and B. There are a levels of factor A and b levels of factor B which are arranged in a factorial design; that is, each replicate of the experiment contains all ab treatment combinations, with n replicates.

In the general case, suppose y_{ijk} be the observed response when factor A is at the ith level ($i = 1, 2, \cdots, a$) and factor B is at the jth level ($j = 1, 2, \cdots, b$) for the kth replicate ($k = 1, 2, \cdots, n$). The abn observations are selected at random with two factors as a completely randomized design. The effects model for a factorial experiment with two factors is

$$\begin{aligned} y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk} \quad ; \quad i = 1, 2, \cdots, a \\ j = 1, 2, \cdots, b \qquad\qquad ; \qquad k = 1, 2, \cdots, n \end{aligned} \tag{3.10}$$

where μ is the overall mean effect, τ_i is the effect of the factor A, β_j is the effect of the jth level of the factor B, $(\tau\beta)_{ij}$ is the effect of the interaction between τ_i and β_j, and ε_{ijk} is a random error component. In this model ε_{ijk} are assumed to be normally and independently distributed random variables with mean zero and variance σ^2, then $y_{ijk} \sim N(\mu + \tau_i + \beta_j + (\tau\beta)_{ij}\,,\,\sigma^2)$.

For the fixed effects model, the treatment effects are defined as deviation from the overall mean, then $\sum_{i=1}^{a}\tau_i = 0$, $\sum_{j=1}^{b}\beta_j = 0$, and $\sum_{i=1}^{a}(\tau\beta)_{ij} = \sum_{j=1}^{b}(\tau\beta)_{ij} = 0$. The appropriate testing hypotheses about the equality of A treatment effects is

$$\begin{aligned} &H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0 \\ &H_1 : \tau_i \neq 0 \ \text{ for at least one } i \end{aligned} \tag{3.11}$$

and the equality of B treatment effects is

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \ldots = \beta_b = 0 \\ H_1 &: \beta_j \neq 0 \ \text{ for at least one } j \end{aligned} \tag{3.12}$$

and also in determining whether A and B treatments interact, the testing hypotheses is

$$\begin{aligned} H_0 &: \ (\tau\beta)_{ij} = 0 \ \text{ for all } i,\, j \\ H_1 &: \ (\tau\beta)_{ij} \neq 0 \ \text{ for at least one pair } (i,j) \end{aligned} \tag{3.13}$$

Similar to the *single-factor* analysis of variance, a partitioning of total variability into its component parts, the total corrected sum of squares (SS_T) could be written as

$$\begin{aligned} \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \overline{y}_{...})^2 &= bn\sum_{i=1}^{a}(\bar{y}_{i..} - \overline{y}_{...})^2 + an\sum_{j=1}^{b}(\bar{y}_{.j.} - \overline{y}_{...})^2 \\ &+ n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \overline{y}_{...})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \overline{y}_{ij.})^2 \end{aligned} \tag{3.14}$$

where

$$\bar{y}_{ijk} = \frac{1}{abn}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}y_{ijk}; \quad \bar{y}_{i..} = \frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n}y_{ijk};$$

$$\bar{y}_{.j.} = \frac{1}{an}\sum_{i=1}^{a}\sum_{k=1}^{n}y_{ijk}; \quad \bar{y}_{ij.} = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b}y_{ijk}.$$

Equation 3.14 is the fundamental ANOVA equation for the *two-factor* factorial design. Based on this equation and the number of degrees of freedom each sum of squares, the *mean square* for each treatment is calculated via sum of squares divided by its degrees of freedom.

As we assumed that the error terms (ε_{ijk}) of the model (Equation 3.10) are normally and independently distributed with constant variance σ^2, the ratios of *mean squares* MS_A/MS_E, MS_B/MS_E, and MS_{AB}/MS_E are distributed as F with $a-1$, $b-1$, and $(a-1)(b-1)$ numerator degrees of freedom, respectively, and $ab(n-1)$ denominator degrees of freedom. The test procedure in factorial designs is usually summarized in an ANOVA table, as shown in Table 3.1.

The values of the two columns in Table 3.1 (F_0, F_α) can be used for testing of the above hypotheses (Equation 3.11 to 3.13). For instance, if $F_A > F_{\alpha,\, a-1,\, ab(n-1)}$ then, we reject H_0 in hypotheses for testing the A treatments (Equation 3.11).

The general factorial design which we use in Chapter 5 is a natural extension

of the case with two factors.

Table 3.1: The ANOVA Table for the Two-Factor Factorial, Fixed Effects Model

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	F_α
A treatments	SS_A	$a-1$	$MS_A = \frac{SS_A}{a-1}$	$F_A = \frac{MS_A}{MS_E}$	$F_{\alpha,\, a-1,\, ab(n-1)}$
B treatments	SS_B	$b-1$	$MS_B = \frac{SS_B}{b-1}$	$F_B = \frac{MS_B}{MS_E}$	$F_{\alpha,\, b-1,\, ab(n-1)}$
Interaction	SS_{AB}	$(a-1)(b-1)$	$MS_{AB} = \frac{SS_{AB}}{(a-1)(b-1)}$	$F_{AB} = \frac{MS_{AB}}{MS_E}$	$F_{\alpha,\, (a-1)(b-1),\, ab(n-1)}$
Error	SS_E	$ab(n-1)$	$MS_E = \frac{SS_E}{ab(n-1)}$		
Total	SS_T	$abn-1$			

3.1.2. Regression Analysis

The goal of regression analysis is to create a valid model explaining the linear or non-linear relation between dependent and independent variables. For instance in our study, total grade is considered as dependent variable and family background and individual factors as independent variables. In the following by y we represent a dependent or target variable and x_1,x_2, ... represent the independent variables. Formula 3.15 shows a multiple regression model (with k regressors/independent variables), where ε represents the model error:

$$y = f(x_1, x_2, \cdots, x_k) + \varepsilon \tag{3.15}$$

The model could be of any kind, and so the validation, using tests, is an important part of regression analysis. We now only assume the model follows a linear relation. Thus this model can be formulated as:

$$y = \alpha + \sum_{i=1}^{k} \beta_i x_i + \varepsilon \tag{3.16}$$

Here α is the intercept of the regression function and β_i's are regression coefficients that have to be estimated in order to formulate the influence of every independent variable on the dependent variable. Here we call ε the *residual* and it denotes the error between model and reality which is usually assumed to be independent normally distributed ($\varepsilon = y - E(y|x_1, x_2, \ldots, x_k)$, $\varepsilon \sim N(0, \sigma^2)$). The unconditional and conditional means of the residual (ε) are zero and are uncorrelated with the independent variables $x_1, \ldots,\ x_k$ ($E(\varepsilon) = 0$ and $E(\varepsilon|x_1, \ldots, x_k) = 0$ and $E(\varepsilon x) = 0$). For a linear regression to be appropriate the joint distribution of the vector $\vec{z} = (y, x_1, \ldots, x_k)'$ has to be of a certain type (e.g. normal). In other words, if the response or dependent variable is normally distributed, a linear regression is appropriate. For simplicity,

we present the formula in matrix notation as:

$$\vec{y} = \mathbf{X}\vec{\beta} + \vec{\varepsilon} \tag{3.17}$$

Here $\vec{y}$ is the vector of different observations of the dependent variable, $\vec{\varepsilon}$ ($\vec{\varepsilon} = \vec{y} - \mathbf{X}\vec{\beta}$) is the vector of error terms, $\mathbf{X}$ is the matrix of independent variables for observations. In $\mathbf{X}$ the number of the columns corresponds to the number of regressors (k) used in the model whereas the number of rows is equal to the number of observations (n). However when the constant term α is different from zero then $\mathbf{X}$ has an additional column of 1's only . Finally the vector $\vec{\beta}$ contains all regression coefficients β_i and in the case of a constant term, α is same as β_0.

In order to estimate the regression coefficients, different methods are used, such as Maximum Likelihood Estimation (MLE) and Ordinary Least Squares (OLS) technique. For instance, in MLE method the estimators of $\vec{\beta}$ and $\sigma^2$are obtained by Equation 3.18 [50].

$$\begin{aligned} \hat{\vec{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\vec{y} \\ \hat{\sigma^2} &= \frac{(\vec{y}-\mathbf{X}\hat{\vec{\beta}})'(\vec{y}-\mathbf{X}\hat{\vec{\beta}})}{n} \end{aligned} \tag{3.18}$$

The regression analysis includes more than just the setup and calculation of the model. A few tests such as *t-test* and *F-test* should be run to validate the model. The t-test formulates the hypothesis as $H_0 : \beta_i = 0,\ H_1 : \beta_i \neq 0$. It tests the significance of each independent variables individually. That means, the t-test asks whether a certain variable is significant in the regression model or not. So the test statistic for β_i is:

$$t_0 = \frac{\hat{\beta}_i}{\hat{\sigma}_{\beta_i}} \quad with \qquad t_0 \sim t_{(n-k-1)} \tag{3.19}$$

The null hypothesis of significance of each independent variable (β_i) is rejected if $|t_0| > t_{\alpha/2,\, n-k-1}$.

The F-test includes all parameters which can test all regression coefficients at one time, but the result does not have to coincide with the results from the t-test. The conclusion of the t-test with k tests may be that all regression coefficients are significant. Meanwhile, the F-test could at the same time result in the opposite conclusion. The F-test for the significance of regression is a test to determine if there is a linear relationship between the dependent variable y and any of the independent variables $x_1, x_2, \ldots, x_k$ [67]. For this purpose the

appropriate hypotheses are

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \ldots = \beta_k = 0 \\ H_1 &: \beta_i \neq 0 \text{ for at least one } i \end{aligned} \tag{3.20}$$

The test procedure is a generalization of the analysis of variance. The *total sum of squares* SS_T is partitioned into a *sum of squares due to regression* SS_R, and a *residual sum of squares* SS_{Res}. That means, $SS_T = SS_R + SS_{Res}$ which are calculated by:

$$\begin{aligned} SS_T &= \sum_{i=1}^{n}(y_i - \bar{y})^2 = \vec{y}'\vec{y} - n\bar{y}^2 \\ SS_R &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\vec{\beta}}'\mathbf{X}'\vec{y} - n\bar{y}^2 \\ SS_{Res} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \vec{y}'\vec{y} - \hat{\vec{\beta}}'\mathbf{X}'\vec{y} \end{aligned} \tag{3.21}$$

Then the F-test statistic is defined by:

$$F_0 = \frac{SS_T - SS_{Res}}{SS_{Res}} \frac{(n-k-1)}{k} \quad with \qquad F_0 \sim F_{(k,\, n-k-1)} \tag{3.22}$$

The null hypothesis H_0 is rejected if $F_0 > F_{\alpha,\, k,\, n-k-1}$.

A comparison with other setups is inevitable to ensure the significance of the model. One needs to check whether the chosen parameters are really significant, or whether the linear relationship is indeed appropriate and so on. For this purpose, the stepwise regression method is discussed.

Stepwise regression is one of several iterative procedures for variable selection/elimination. In this regression, variables are added/removed one-by-one based on their contribution to R^2 which is calculated by $\frac{SS_R}{SS_T}$. Some statistical package report *t or F* statistics for entering or removing variables. At each step we determine whether any of the variables can be added/removed. This method has two types: the *forward selection* and the *backward elimination*. The forward method introduces new variables stepwise according to a certain criterion and stops when a final amount has been reached. One possible criterion can be R^2, for which the method includes the variables with the highest contribution to the model's R^2 from a set of independent variables. It then searches again for the next variables in the set of independent variables that are left over. The method stops when we arrive of a certain number of variables added to the model or when a specific R^2 value is achieved. The backward method does the very same thing but the other way round. It starts

with a model including the whole set of independent variables. It then removes them step by step until the desired number of variable is reached or a certain value for R^2 is attained. [67, 68].

3.1.3. Logistic Regression

This method is used to analyze the dependence of a probability of an event on some variables. In our case the event is the acceptance of a candidate i on some university ($Y_i = 1$), so the probability of acceptance may be denoted as

$$\pi_i = P(Y_i = 1). \tag{3.23}$$

The variables are denoted by $X_1, X_2, \cdots, X_k$, so we use the model equation

$$\pi_i = P(Y_i = 1) = f(X_1,\ X_2\ \cdots,\ X_k\ ) + \varepsilon. \tag{3.24}$$

Since the range of π_i is $[0, 1]$, for an application of Linear Regression the probability π_i has to be transformed [1, 3, 20]. One possibility is the transformation π_i

$$Log\frac{\pi_i}{1-\pi_i} = Logit\,(\pi_i) \tag{3.25}$$

which leads in combination with a linear function

$$f(X_1,\ X_2\ \cdots,\ X_k\ ) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \tag{3.26}$$

to the model equation

$$Log\frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_{1i} X_1 + \cdots + \beta_k X_{ki} + \varepsilon_i \tag{3.27}$$

which leads after some calculations to

$$\pi_i = \frac{e^{\beta_0+\beta_{1i}X_1+\cdots+\beta_k X_{ki}+\varepsilon_i}}{1+e^{\beta_0+\beta_{1i}X_1+\cdots+\beta_k X_{ki}+\varepsilon_i}} \tag{3.28}$$

Similar to the linear regression, the regression coefficient is estimated by the MLE, which usually is written in its logarithmic form as the product of the joint distribution [50, 47].

The test statistic of examining an independent variable in logistic regression is the *Wald* test which is calculated from the following equation:

$$W = \frac{\hat{\beta}}{SE(\hat{\beta})} \tag{3.29}$$

where $SE(\hat{\beta})$ is the standard error of the estimator of the regression coefficient

$(\hat{\beta})$, and W^2 yield a chi-square distribution with one degree of freedom.

3.2. Data Mining Methods

Data mining briefly is the art of extracting useful information and knowledge from a large amount of data. The structure of data mining is based mainly on different algorithms and methods from database technology, statistics, artificial intelligence, machine learning, data visualization, pattern recognition, etc. That means that data mining involves techniques from multiple disciplines. Meanwhile, data mining has been applied in different disciplines, sciences and industries as well as in finance and business [63, 29].

"Data mining is the process of selection, exploration, and modeling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of the database." [28]

Data Mining is closely related to Knowledge Discovery in Databases (KDD) and quite often these two processes are considered equivalent. Widely accepted definitions for KDD and Data Mining have been provided by Fayyad, Piatestsky-Shapiro, & Smyth [23, 24]:

"Knowledge Discovery in Databases is the process of extracting interesting, non-trivial, implicit, previously unknown and potentially useful information or patterns from data in large databases."

3.2.1. Data Mining Process

In this section, we give an overview of the data mining standard CRISP-DM. The Cross-Industry Standard Process for Data Mining (CRISP-DM) was developed in 1996 by Daimler Chrysler, SPSS, and NCR. The CRISP-DM process provides a life-cycle for a data mining project, consisting of six steps which includes business understanding, data understanding, data preparation, modelling, evaluation, and deployment. Moving back and forth is allowed and required (Figure 3.1) [52, 16].

- **1. Business Understanding:** At the initial step, the data miner focuses on understanding the project objectives and requirements from a business perspective, and converts this knowledge into a data mining problem definition. The data miner can then define his/her strategy to achieve the objectives.

- **2. Data Understanding:** In this step after collection of the initial

data, it is analyzed in order to become more familiar with the dataset. Further, the data quality problems are identified and first insights into the dataset are discovered.

- **3. Data Preparation:** The data preparation step covers all the activities and data analysis needed to construct the final dataset, which is to be applied in the next steps. Data preparation tasks are likely to be performed multiple times. Tasks include table, record, and attribute selection, as well as aggregating attribute values, data integration, transformation, e.g. replacing and estimation missing values, sampling methods, omit unusable variables to prepare data for modelling phase. At the end of this step a final dataset is constructed out of the initial raw data.

- **4. Modelling:** In this step, various modeling techniques are specified and applied depending on the type of data and project goals. Typically, there are several techniques for the same data mining problem type. Modelling is undertaken in selection of the modelling techniques, applying the modelling techniques, calibrating model setting, and initial assessment of the models. Usually, there is more than just one technique that can be applied. In fact, going back to the data preparation step is often necessary.

- **5. Evaluation:** Before proceeding to final deployment of the model, the results have to be evaluated and the steps that led to the model need to be reviewed to make sure that the model fits best the business objectives. Finally at the end of the step, a decision on the use of the data mining results is made.

- **6. Deployment:** Deployment is the final step of the data mining project whereas the creation of the model is generally not the end. Even if the purpose of the model is to increase knowledge gained from the data, the knowledge gained will need to be organized and presented in a way that the user can apply it. Additionally, monitoring and maintenance of the data mining project is planned at this stage. Conclusively, a final report is produced and the project is reviewed. However, even after deployment, it is important for the user to understand up front what actions need to be carried out in order to actually make use of the created models.

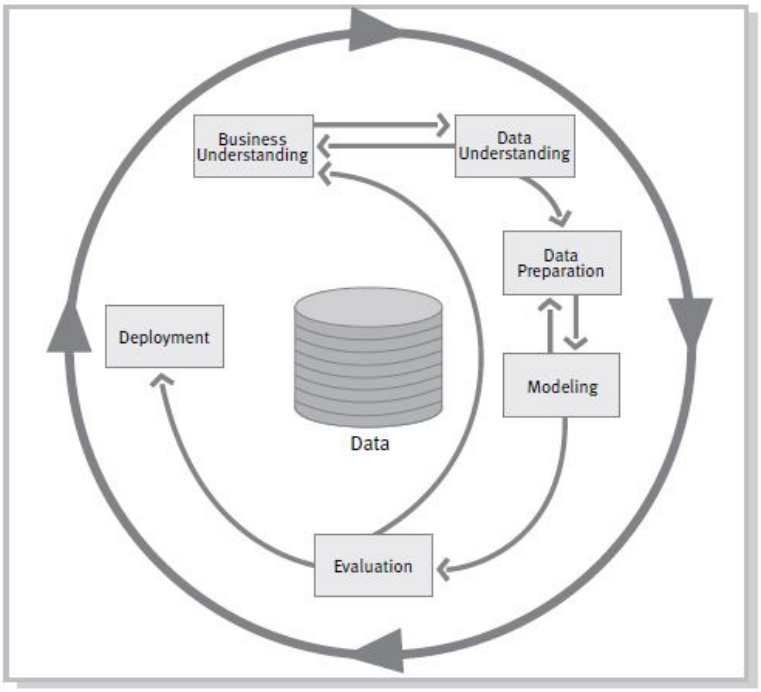

Figure 3.1: Steps of the CRISP–DM Life Cycle of a Data Mining Project [16]

3.2.2. Data Mining Tasks

The tasks of data mining are very diverse and distinct, as several patterns might exist in a huge database. Several methods and techniques are needed for identifying different kinds of patterns. Based on the type of the patterns we are looking for, tasks in data mining can be divided into classification, prediction, association, clustering and outlier analysis [26]. We now briefly discuss these tasks. However, in this work we mainly focus on the prediction and classification models.

Classification

Classification is the derivation of a function or model which determines the class of an object based on its attributes. A set of objects is given as the training set in which every object is represented by a vector of attributes along with its class. A classification function or model is constructed by analyzing the relationship between the attributes and the classes of the objects in the training set. Such a classification function or model can be used to classify future objects and develop a better understanding of the classes of the objects in the database.

Prediction

This is the task for predicting the value of the dependent variable from known values of the predictors, based on a prediction model like linear regression.

Association

Association is the discovery of togetherness or connection of objects usually termed as association rule. An association rule reveals the associative relationships among objects, i.e., the appearance of a set of objects in a database is strongly related to the appearance of another set of objects.

Clustering

Clustering is the identification of classes, also called clusters or groups, for a set of objects whose classes are unknown. The objects are clustered in a way that the intraclass similarities are maximized and the interclass similarities are minimized based on some criteria defined on the attributes of objects. Once the clusters are decided, the objects are labeled with their corresponding clusters. Common features of the objects in a cluster are summarized to form the class description. Usually this task is used in unsupervised learning cases [84] .

Outlier analysis

Outliers are data objects that do not comply with the general behavior of model of the data. In other words, the identification and exclusion of data that do not proceed by the behavior of the rest of the data records. Furthermore, outliers are observations that differ considerably from the majority of the data which can seriously disturb the least squares fit [67].

In this study, we need to detect the outliers of the some variables. For this purpose, we investigate the various detection and treatment of outliers methods[1]. One of the common and easy method is based on the three or four standard deviations from the mean. Stefansky [82, 83] has proposed an approximate test for finding outliers based on the normed residual$|e_i| / \sqrt{\sum_{i=1}^{n} e_i^2}$, where $e_i = y_i - \hat{y}_i$ is residual of i'th observation. In other words, the difference between the i'th observed and i'th fitted values is called residual e_i ; $i = 1, 2, \ldots, n$ the number of observations.

[1]For detail of the methods for detecting and dealing with outliers, see D. C. Montgomery[67].

3.2.3. Data Mining Algorithms

Data modeling involves using certain data mining techniques in order to describe the already existing patterns within the database. For a given target a certain technique, like description or prediction, is chosen depending on the data type. For example, if the data are numerical or quantitative, then we have a regression problem or prediction. Similarly if the data is nominal or qualitative, we then have a classification problem.

As a multidisciplinary field, data mining adopted its techniques from many areas, including statistics, machine learning database systems, rough sets, and visualization. In other words, data mining involves an integration of several techniques from multiple disciplines[80].

We now present an overview of two of the data mining algorithms, namely Decision Trees and Neural Networks which are dealt with in this work.

3.2.3.1. Decision Trees

Decision trees are usually categorized as classification trees and regression trees. In classification trees, the target variable is categorical or qualitative. Regression trees are used when the dependent variable is continuous [29].

The basic idea of tree construction is to find subsets with maximum homogeneity or cases that are located in a subset belonging only to one class of target variable. At each step of splitting, tree algorithms split cases with independent variables that have maximum homogeneity.

In trees construction, leaves are final nodes and the first node of a tree is a root node. Impurity of a node t is defined as a function of the probability of a different class in considering a node:

$$I(t) = \phi(p_1, p_2, \ldots, p_k) \quad ; \quad p_i \geq 0, \quad i = 1,\ 2,\ \ldots,\ k \quad ; \quad \sum_{i=1}^{k} p_i = 1$$

where each p_i is a probability of cases belong to class i. Each quantification of impurity should have the following characteristics:

- An impurity function has a minimum value, when all cases have the same class, i.e., when the probability of being in a certain class is 1 and 0 for all the other classes.
- An impurity function must be maximum, when relative frequencies of all classes are the same, that is all the p_i are equal (uniform distribution).
- An impurity function should be symmetric, i.e., if we permute p_i , ϕ

remains constant..

There are different kinds of impurity functions with these characteristics. One of the main differences in tree algorithms is related to the impurity functions. Following are some of the important functions figured out/ invented by the researchers [28, 11]:

- **Misclassification Rate**

$$MI(t) = 1 - \underset{i}{max}(p(i|t)) \, , \quad i = 1, \, 2, \, ... \, , \, k$$

where $p(i|t)$ is the probability of class i in node t.

- **Gini**

$$GI(t) = \sum_{i}^{k} p(i|t)(1 - p(i|t)) = 1 - \sum_{i}^{k} p^2(i|t)$$

For instance, considering two classes only, the formula can be rewritten as following:

$$GI(t) = p(0|t)p(1|t) + p(1|t)(1 - p(1|t)$$

- **Entropy**

$$EI(t) = -\sum_{i}^{k} p(i|t)log_2(p(i|t))$$

For two class we have:

$$EI(t) = -p(0|t)log_2(p(0|t)) - p(1|t)log_2(p(1|t))$$

- **Maximize Half-Sum of Squares**

$$Chi = \frac{n(l).n(r).[p(0|l) - p(0|r)]^2}{n(l) + n(r)}$$

where $n(r)$ and $n(l)$ are the number of observations in the right and left nodes. In this function, a large value of χ^2statistic (Chi) means that the two proportions are not the same.

The reduction of impurity that the split obtained can be defined as quality and is given as following:

$$\triangle I = I(t) - [\pi(l)I(l) + \pi(r)I(r)]$$

where $\pi(l)$ and $\pi(r)$ are the observed proportions of observations in classification. In fact, tree algorithms select the variable that has the best quality of a split. Finally, tree algorithms label leaf nodes based on the majority of target variables. In regression trees, tree fitted $\hat{y}_i$ is equal to the mean of dependent variables for observations when considering leaf nodes.

Classification and regression trees (CART) are the most common tree algorithms [12]. In CART the target variable can be categorical as well as continuous. The impurity function of CART is assumed to be Gini or entropy. Chi-square Automatic Detection (CHAID) in another decision tree algorithm was developed by Kass [42] where the impurity is assumed to be chi-square. Other popular tree algorithms are C4.5 and its later version, C5.0 [74]. C4.5 and C5.0 were applied for classification only. Another algorithm used in our study is QUEST (for Quick, Unbiased, Efficient, Statistical Tree) [54].

The most important advantage of trees is ease of interpretation and understanding. Furthermore, decision trees are robust to outliers. In addition, these models are nonparametric and there are no distribution restrictions. However, the disadvantage of tree models is that they are unstable which means they are sensitive to training data [68].

3.2.3.2. Artificial Neural Networks

Artificial Neural Networks (ANN) is an artificial intelligence model that was originally developed based on the processing of information in the human brain. This algorithm can be used for classification, prediction and descriptive purposes. ANN gives a good performance to fit observed data, especially with high dimensional data and datasets with missing values, errors or inaccuracies [51, 29].

A neural network consists of a set of computational cells/units called neurons. The neurons of an ANN are organized in layers. The layers can be of three types: *input*, *output*, or *hidden*. The input layer receives information of input variables, which include n neurons or input units (x_1, x_2, ... , x_n). The rule of input layer is to transmits information to the next level. The output layer's rule is to produces the final results. Each of its neurons corresponds to target variables (y_1, y_2, ... , y_p). Between the input and the output layers there can be one or more hidden layers as intermediate levels. The hidden layers are exclusively for analysis, which in their function is to take the relationship between the input and the output neurons (variables). The number of layers in an ANN is counted by weighted neurons which are to be learned from the data.

Furthermore, the neurons are linked together through axons that are weighted connections in the ANN. These units are located in layers, and every neuron in a layer is linked to the neurons of the next and previous layers. A neuron, as a computational cell/unit, receives input as activation signals and then forwards

an output signal. Although a neuron receives more than one input signal, it only forwards one output signal. The input signal is related to a connection weight that shows the importance of the input unit. In the learning process, the weights are adapted. There is a threshold value, called bias, which is the same as the intercept in a regression analysis. Suppose that $x_1, \ldots x_n$ are input units (variables), then $w_{0j}x_0$ is known as bias and $w_{0j}, w_{1j}, \ldots, w_{nj}$ are connection weights where n indicates the number of the variables (input units) and j shows the number of the neurons to which the weight applies [29, 63, 90]. A single layer neural network, is represented in Figure 3.2.

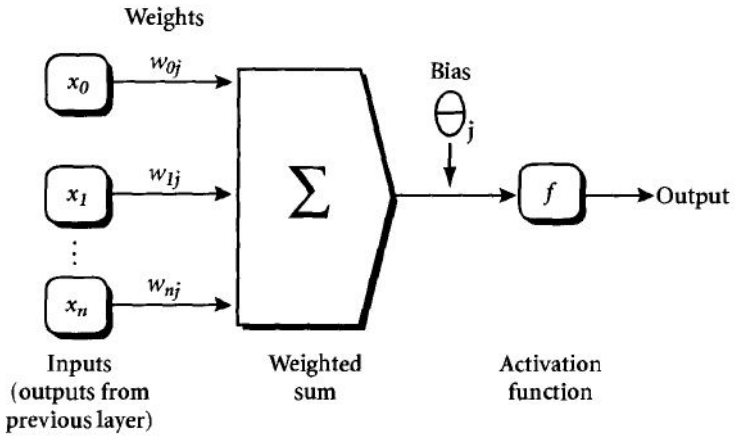

Figure 3.2: Single Layer Neural Network [33]

The algorithm (model) mentioned above can be expressed as follows:

$$q_j = w_{0j}x_0 + w_{1j}x_1 + \ldots + w_{nj}x_n = \sum_{i=0}^{n} w_{ij}x_i$$

$$y_j = f(q_j) = f(\sum_{i=0}^{n} w_{ij}x_i) \quad , \quad j = 1,\ 2,\ \ldots\ ,\ p$$

The function f transforms q_j, is called *activation function*[2]. There are different kinds of activation functions in ANN modelling. In [29], some of the important activation functions are given as the following:

- Threshold or sign activation function:

$$F(q) = \begin{cases} 1 & q \geq 0 \\ 0 & q < 0 \end{cases}$$

[2]See Figure 3.2

- Stepwise activation function:

$$F(q) = \begin{cases} \alpha & q \geq \theta \\ \beta & q < \theta \end{cases}$$

  Note that when $\alpha = 1,\ \beta = 0$ and $\theta = 0$, this is the same as the sign activation function.

- Logistic activation or sigmoidal activation function:

$$F(q) = \frac{1}{1 + e^{-aq}} \qquad , \qquad a > 0$$

The sigmoidal activation function is the most popular activation function. Another kind of activation function is the soft-max function that is applied when the categorical response variable has more than two categories.

An ANN with more than one layer of weighted neurons, which contain one or more hidden layers is a multilayer perceptron. For instance, a two-layer network has one hidden layer; there are n neurons in the input layer, h neurons in hidden layer and p in the output layer. Moreover, weights w_{ik} ($i = 1,\ 2,\ \ldots,\ n \quad ; \quad k = 1,\ 2,\ \ldots,\ h$) connect the input layer nodes with the hidden layer nodes, and also, weights z_{kj} ($k = 1,\ 2,\ \ldots,\ h \quad ; \quad j = 1,\ 2,\ \ldots,\ p$) connect the hidden layer with the output layer. In this network, the hidden layer weighted by the weights w_{ik} from the input layer, that produce outputs $h_k = f(\mathbf{x},\ \mathbf{w}_k)$. On the other hand, the neurons of the output layer receive the outputs from the hidden layer weighted by the weights z_{kj} which produces the final network output (results) $y_j = g(\mathbf{h},\ \mathbf{z_j})$. Then the output of the neuron j as a final result is

$$y_j = g(\sum_{k=1}^{h} h_k z_{kj}) = g(\sum_{k=1}^{h} z_{kj}(f(\sum_{i=0}^{n} x_i w_{ik}))) \,, \qquad j = 1,\ 2,\ \ldots,\ p\,.$$

Furthermore, a perceptron is an ANN with a single neuron and sign activation function, whereas a multilayer perceptron has an output layer, an input layer, and some hidden layers. In hidden layers, each neuron has a weight that is used on its input. It is clear that the value of weights in layers can be different. The outputs from each neuron in a hidden layer have weights and become inputs for the next hidden layer, and they become inputs to the output layer if there is only one hidden layer. The outputs of output layers are used for classifying each sample with comparison to the cutoff values.

The ANN, as a machine learning method, does not have distributional as-

sumption. However, this model can be used for classification and regression analysis. In other words, ANNs are also able to deal with continuous and categorical variables as independent variables [79, 51].

3.2.4. Performance Evaluation

The following criteria have been used for comparison of classification and prediction models which summarized by Han et al. [33] :

- **Accuracy:** In classification, this refers to the accuracy of models in predicting the class of new/unseen observations. In regression, this specifies the quality of models in predicting the dependent variable. In fact, the dataset is divided into two sections, learning and testing, as when we apply the learning data to calculate the accuracy rate than we might have a more realistic estimator. Thus, we use the second part of our data (testing set) for calculation of the accuracy rate or the misclassification rate. The proportion of testing data that is correctly classified using the model is the accuracy rate of the classification model.
- **Speed:** This criterion is related to the time taken in training and applying the model.
- **Robustness:** Robustness refers to the behavior of the model towards missing and noisy data.
- **Scalability:** This characteristic is related to the model´s capacity to deal with huge datasets.

3.2.4.1. Cross-validation

The cross-validation randomly divides the dataset into subsets D_1, $D_2, \ldots, D_k$ such that

- $D_i \cap D_j = \emptyset \qquad \forall i \neq j$
- $\cup_{i=1}^{k} D_i = D$
- The sizes of subsets are relatively similar.

The subset j is used as a testing data set, and other subsets are used as learning data sets in iteration j. In fact, each subset D_i is applied $k-1$ times for learning and only once for testing. The accuracy rate is equal to means of measure in k'th iteration. This is in contrast to 'leave-one-out' cross-validation, in which one of the subsets is fixed for testing purposes only.

3.2.4.2. Confusion Matrix

In order to compare different classification models consider the confusion matrix given in Table 3.1 [28, 5].

Table 3.2: Confusion Matrix

Actual Class \ Predicted Class	**True (1)**	**False (0)**	**Total**
True (1)	a	b	a+b
False (0)	c	d	c+d
Total	a+c	b+d	a+b+c+d

The *sensitivity* is then defined as the proportion of cases correctly predicted as true, whereas the *specificity* is defined as the proportion of cases correctly predicted as false. More precisely, $Sensitivity = \frac{a}{a+b}$, and $Specificity = \frac{d}{c+d}$, where a and d are the number of observations in the test data which the model has predicted their class correctly. On the other hand, b and c are the number of observations which the model predicted their class wrongly.

In addition, we have:

- $False\ positive\ rate = \frac{b}{a+b} = 1 - sensitivity$ known as type I error Statistics.
- $False\ negative\ rate = \frac{c}{c+d} = 1 - specificity$ known as type II error in Statistics.

3.2.4.3. Coefficient of Determination (R-squared)

The coefficient of determination is applied for the identification of regression model's quality. This coefficient, denoted by R^2 or r^2, can be interpreted as the percentage of variance of the dependent variable explained by the model. In other words, the coefficient is a number that indicates how well data fit a statistical model. It is calculated by the division of the variation of the regressors by the variation of the dependent variables. The coefficient of determination is defined as:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

where $SST = \sum_{t=1}^{n}(y_i - \bar{y})^2$ Total Sum of Squares, $SSR = \sum_{t=1}^{n}(\hat{y}_i - \bar{y})^2$ Regression Sum of Squares and $SSE = \sum_{t=1}^{n}(y_i - \hat{y}_i)^2$Error Sum of Squares. R^2 can be assumed as the proportion of variation of the dependent variable that can be explained by independent variables and is between 0 and 1. When $R^2 = 1$, it means that the independent variables can completely explain the

variation of dependent variable. A major drawback of this measure is its insensitivity to the quantity of regressors. If one simply follows the value of R^2 to decide which model to prefer, a model with a lot of regressors will always be the choice over a model with fewer regressors. The larger is the number of integrated regressors, the greater is the explained variance. If the number of independent variables is equal the number of observations, R^2 is then equal to 1. But the more regressors used, the lower the explanatory power out-of-sample will get.

Adjusted R^2: A new measure with penalty for the number of regressors in the regression model is defined as follows:

$$Adjusted\ R^2 = \frac{(n-1)\,SSR}{(n-k-1)\,SST}$$

where n is the number of observations, and k the number of independent variables

3.2.4.4. Root Mean Squared Error

Root Mean Squared Error (RMSE) is equal to the square root of the average of the squared difference between estimated value by model and real value of observations:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

where n is the number of observation, $\hat{y}_i$'s are estimated value by the model and y_i's are real values of observations. The domain of this measure is between 0 and the maximum squared error. The outliers have an influence on this measure; also, the unit of this metric is the same target variable unit [68].

4. Data Description

This chapter is devoted to data description. In Section 4.1 we first briefly introduce the NOET, and give general information about the used data. Data understanding and data preparation which include data quality and data cleaning are presented in Section 4.2. In Section 4.3 we provide a thorough explanation of the applicants' data followed by a description of their family background in Section 4.4. Finally, a description of the chance of entrance along with a statistical descriptive analysis[1] are presented in Section 4.5.

4.1. General Information about the Used Data

The NOET in Iran which is the owner of the data used in our study, (NOET Iran, our data source) was established by the ministry of Science and Higher Education in 1968. It is responsible for the entrance examination to test prospect students for different academic institutes. In cooperation with universities, several regulations were developed and later implemented. With the increasing number of applicants, NOET was extended in 1975 to a broader organization.

Academic performance in this study is calculated by the weighted average of the normalized scores of examination subjects, which include general and specialized subjects in all the testing groups at the WEE. We now describe the several factors we consider in this study.

Socioeconomic status: The socioeconomic status of applicants in this study will be evaluated by the family and environmental factors[2] such as parental education, father's occupation, family income, place of residence (region and province) with regard to gender.

Status of county/region[3]: Due to various indicators of educational, cultural,

[1] For this purpose, we use the Statistical Package for the Social Sciences (SPSS) version 23.

[2] The categorization of variables such as status of province, family income, parental education, parental occupation and the number of family members were coded by the NOET.

[3] Each candidate/applicant in the WEE, corresponding to the her/his place of residence, categorized to 3 county/region by the NOET.

social, economic, and medical in different parts of the country, each region is categorized as high, mid or low county.

Status of province/state: Iran is administratively divided into 30 provinces[4]. Each province is assigned a number which is the weighted average of status of county of the candidates of the province where high, mid and low correspond to 1, 2, and 3 respectively. The status of each province is provided by NOET.

Our dataset contains two parts: part one consists of applicants' specifications such as gender, testing group, age, total grade and acceptance as well as information about their applications during 2005-2009 from NOET's original data file. The second part of the data is extracted from a questionnaire with four/six questions[5] regarding applicants' family background, which is filled out by the majority of applicants during the nationwide examination. These two files are merged into one dataset according to the applicant's ID.

The Iranian university and college admission system require a ranking up to one hundred majors/code-fields by a student according to his/her preference after receiving examination grades. The NOET processes admission applications for all kind of universities at all levels. The nationwide exam is held for five groups: *mathematics and physics, empirical sciences, human sciences, art, and foreign language* [6]. Typically, five to ten subjects are examined in each group. The special subjects for each testing group are listed in Table 4.1. Four general subjects are common in all groups: Farsi Literature, Arabic Language, Islamic Literature, and Foreign Language. Once assessed, for each subject the examination results are used to produce a score between -33 to 100 per candidate. The structure of exams is based on multiple choice questions, and every three wrong answers are considered as one negative point. In other words, if a candidate has three wrong answers and one correct answer, his or her mark is equal to zero. Subsequent to examination grading, the NOET normalizes a total mark for each candidate following a certain process based on standardization of the total mark of candidates.

[4]The list of provinces and the weighted averages will be presented in Section 4.4.

[5]The family background factors are filled in by applicants as a questionnaire annually, which are the parental education and occupation, family income, and the number of family members.The number and percentage of answers these questions/factors will be presented in Section 4.4.

[6]These groups are either main (Mathematics and Physics, Empirical Sciences, Human Sciences) or floating (Art, Foreign Language) groups. Each candidate have to participate in one of the main groups. He/She can also participate in one of the floating/optional groups.

Table 4.1: Special Subjects in Each Testing Group

Testing Group		**Special Subject**
Main Group	Mathematics and Physics	Mathematics, Physics, Chemistry
	Empirical Sciences	Mathematics, Physics, Chemistry, Biology, Geology
	Human Sciences	Mathematics, Economics, History and Geography, Social Sciences, Philosophy and Logic, Psychology
Floating/Optional Group	Art	Mathematics, Art Information, Technical Drawing, Music, Film and Visual Creativity, Play Skills
	Foreign Language	Specialize Language

For each major/field the number of students, who can be accepted is fixed. The total grade for acceptance to a major is set in a way that eligible applicants are offered a position in their highest preferred major. In case there are several candidates with the same total grade for the last position in a field, the capacity of the field is increased in a way that all these candidates can be accepted. The point requirement for a special major is not known prior to the exam. It can be influenced by the examination results of the candidates who applied for that major and by the number of available places in that field.

Some fields have minimum entry standards, for example, good knowledge of mathematics. A few fields in the Art group have interviews, but these are not common. The candidates who apply within the first three groups are also allowed to apply for the Art and Foreign Language groups.

4.2. Data Understanding and Data Preparation

Data quality is a serious concern in any data driven enterprise and in data mining. Solving data quality problems usually requires a large investment of time and energy. Up to 80% of a data mining project deals with preparation and making the data reliable enough in order to ensure the trustworthiness of the data. These problems can be addressed by a multidisciplinary approach, such as management science, statistics, database technique, and metadata management [40, 18].

In this work we used descriptive statistical and database methods such as data exploration, aggregation, transformation and data merging to prepare the data. Furthermore, after the exploration and preparation of the original datasets, the datasets of different data files (from dataset of Year 2005 to that of 2009) are then merged into a single dataset.

During WEE registration, the candidates have to answer various questions

Table 4.2: The Number of Applicants and Actual Candidates in WEE During 2005 to 2009

Year:		**2005**	**2006**	**2007**	**2008**	**2009**
Applicants		1,488,040	1,424,492	1,562,968	1,507,372	1,389,715
Actual Candidates		1,186,650	1,147,895	1,101,324	1,157,483	1,077,749
Duplicated candidates		301,390 (20.3%)	276,597 (19.4%)	461,644 (29.5%)	349,889 (23.2%)	311,966 (22.4%)
Trend of the actual number of the applicants	2005	—				
	2006	-3.3%	—			
	2007	-7.2%	-4.1%	—		
	2008	-2.5%	+0.8%	+5.1%	—	
	2009	-9.2%	-6.1%	-2.1%	-6.9%	—

about family background factors. The information provided by the candidates are precise and guarantee high quality dataset. There were some cases with missing answers which are eliminated from the dataset. In addition, we have eliminated the outlier candidates (e.g. candidates older than 35 years old) to be sure that such cases do not influence the results. The ratio of eliminated candidates due to missing information or being outlier was always below 1% across different years.

As we explained in the previous section, the testing groups are divided in two groups which include the main group (Mathematics and Physics, Empirical Sciences, Human Sciences) and the floating group (Language and Art). The Language and Art groups are floating groups in the sense that the candidates could optionally choose either one or both of these groups along with one of the main required groups (Mathematics and Physics, Empirical Sciences, Human Sciences). This means that there were candidates who had chosen more than one group simultaneously. For instance, in 2005 the total number of applicants in the Iranian nationwide university entrance examination was 1,488,040 persons. After removing duplications[7], only 1,186,650 non-duplicating candidates were left which we consider in our analysis.

Similarly, in 2006 the total number of applicants was 1,424,492. Therefore, after removing the duplications, 1,147,895 unique candidates remained which were used in our analysis. Note that the number of candidates in 2006 is reduced by approximately 3.3% as compared to 2005.

A summary of all the datasets along with the trend of applicants in WEE is shown in Table 4.2. It can be seen that the number of non-duplicating

[7]For this purpose, NOET used some database techniques such as sorting by unique ID, creating new variable.

candidates is 1,101,324 for year 2007, 1,157,483 for year 2008 and 1,077,749 for year 2009. Hence, the proportion of duplicated applicants was 20.3%,19.4%. 29.5%, 23.2% and 22.4% for year 2005, 2006, 2007, 2008 and 2009 respectively. Note that the total number of candidates has decreased from 2005 to 2009 approximately by 10%.

4.3. Description of the Applicants Data

4.3.1. General Aspects

In some countries the number of applicants for higher education is significantly higher than the number of available seats. In such countries, governments have to come up with a fair solution to select the most suitable persons for entering into the universities. In most cases, the selection process is based on a nation wide examination. The rapid growth of the population in Iran after Islamic revolution (1979) lead to a significant increase in the number of university applicants. Therefore, the examination became very important for both government and applicants.

In this section, we discuss the impact of family background on the examination results of the applicants. This study is a contribution to the general debate of educational attainment, which shows that parental education has positive effects on the children's success, from an educational point of view.

4.3.2. Information about the Applicants Characteristics

For characterizing the applicants, we use *age, gender, testing group, total grade, acceptance* as variables. Table 4.3 shows the number and percentage of candidates by gender and testing group during 2005 to 2009. It shows that 25.83% of candidates were from Group I (Mathematics and Physics) which is the smallest group, whereas Group III (Human Sciences) comprises the largest number (39.34%) of candidates. Also, as it can be seen in Table 4.3, the ratio of candidates in Human Science group is relatively constant across 2005 to 2009. In contrast, by moving from 2005 to 2009, the ratio of candidates in Mathematics and Physics reduces while that of Empirical Science grows.

Table 4.4 shows the description of variables along with the quantity and quality attributes for all years 2005 to 2009. For instance, the age and total grade are numeric data. The other variables are nominal, such as gender, testing groups, acceptance, and family background factors.

The minimum and maximum age of participants in 2005 are 13.0 and 84.0

Table 4.3: Frequency Table of Dataset Used in our Analysis According to the Testing Group and Gender

Year	**Gender**	**Testing Group**						**Total**	
		Math. & Phys.		Empi. Scie.		Huma. Scie.			
		No.	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**
2005	Male	181,943	40.0	115,492	25.4	157,295	34.6	454,730	38.3
	Female	147,820	20.2	268,657	36.7	245,086	33.5	731,920	61.7
	Total	329,763	27.8	384,149	32.4	472,738	39.8	1,186,650	100
2006	Male	170,803	39.1	115,507	26.4	150,732	34.5	437,042	38.1
	Female	135,335	19.0	270,452	38.0	305,066	42.9	710,853	61.9
	Total	306,138	26.7	385,959	33.6	455,798	39.7	1,147,895	100
2007	Male	151,118	37.4	112,816	27.9	140,091	34.7	404,025	36.7
	Female	122,459	17.6	277,470	39.8	297,370	42.6	697,299	63.3
	Total	273,577	24.8	390,286	35.4	437,461	39.7	1,101,324	100
2008	Male	157,579	36.4	120,800	27.9	155,050	35.8	433,429	37.4
	Female	124,000	17.1	296,228	40.9	303,826	42.0	724,054	62.6
	Total	281,579	24.3	417,028	36.0	458,876	39.6	1,157,483	100
2009	Male	154,849	37.1	119,197	28.6	143,376	34.3	417,422	38.7
	Female	118,785	18.0	278,563	42.2	262,979	39.8	660,327	61.3
	Total	273,634	25.4	397,760	36.9	406,355	37.7	1,077,749	100
Total 2005-9	Male	816,292	38.0	583,812	27.2	746,544	34.8	2,146,648	37.9
	Female	648,399	18.4	1,391,370	39.5	1,484,684	42.1	3,524,453	62.1
	Total	1,464,691	25.83	1,975,182	34.83	2,231,228	39.34	5,671,101	100

respectively, where the mean of age is 20.17 years. After eliminating the outlier candidates (i.e., older than 35 years old), the mean of the age becomes 20.04 years. Moreover, 61.7% of participants are female and 38.3% are male. The results also show that only 24.1% of candidates are accepted at universities. That means the chance of candidates for entrance into a university is approximately 1/4. As shown in Table 4.4, the acceptance rate in different years varies between 24.1% and 45.3%.

The value variation of the applicants' characteristics is not too high. On the other hand, the mean of age has increased slightly as well as the rate of acceptance at universities.

Note that the chance of entrance to universities has increased to 42.9% in 2009 from 24.1% in 2005. Given the fact that number of applicants did not change from 2005 to 2009, this indicates an acceptable growth of the academic organization and universities.

Table 4.4: Description of Variables Used in Data Analysis During 2005 to 2009

Variable Name	Values	Year						Value Type
		2005	2006	2007	2008	2009	2005-9	
Age	Min	13.00	12.00	15.00	14.00	13.00	12.00	Numerical
	Max	84.00	76.00	74.00	79.00	74.00	84.00	
	Mean	20.17	20.23	20.43	20.77	20.91	20.44	
	Median	19.00	19.00	19.00	19.00	19.00	19.00	
	Std. Dev.	2.927	3.01	3.441	3.84	4.145	3.424	
Gender	Female	61.7%	61.9%	63.3%	62.6%	60.2%	62.1%	Nominal
	Male	38.3%	38.1%	36.7%	37.4%	39.8%	37.9%	
Testing Group	Mathematics	27.8%	26.7%	24.9%	24.3%	25.4%	25.8%	Nominal
	Empirical Sciences	32.4%	33.6%	35.4%	36.1%	36.6%	34.8%	
	Human Sciences	39.8%	39.7%	39.7%	39.6%	38%	39.3%	
Total Grade	Min	-4472.0	-4151.0	-4919.0	-4834.0	-4271.0	-4919.0	Numerical
	Max	13517.0	13614.0	13746.0	13688.0	14196.0	14196.0	
	Mean	5236.13	5244.00	5588.36	5592.04	5647.63	5457.11	
	Median	4975.0	5036.0	5340.0	5346.0	5387.0	19.0	
	Std. Dev.	1481.15	1546.36	1562.89	1574.73	1556.24	1554.70	
Acceptance	No	75.9%	64.3%	54.7%	63.0%	57.1%	62.2%	Nominal
	Yes	24.1%	35.7%	45.3%	37.0%	42.9%	37.8%	

4.4. Considering the Family Background

The following family background factors are used by NOET: Father's education, Mother's education, Father's occupation and family income and additionally in 2005 Mother's occupation and the number of family members.

According to NOET's encoding, every variable representing a family background factor has four values[8]. As already mentioned, this data is extracted from a questionnaire[9] with four/six questions regarding these factors. We now provide the description statistics of these factors. First we present the results for each individual year from 2005 to 2009, and then we present the combined result for five years from 2005 to 2009.

Year 2005

Table 4.5 shows the description of the family background factors in 2005. These factors include parental education, parental occupation, number of fam-

[8]Unfortunately the encoding in 2006 differs from that in the other years, (see Table 4.5, ..., 4.9).

[9]Information about the response frequency and response rate can be find in the Table 4.5, ..., 4.9.

ily members, and family income. The percentage of parental education shows that the level of father's education is higher than mother's education.

Table 4.5: Description of Family Background Factors in 2005

(Total Number of Applicants: 1,186,650)

Variable Name	Response Frequency (Rate)	Values	Percent
Father's Education	1095808 (92.3%)	No Education	16.6
		Primary School	36.3
		High School	32.2
		University Degree	14.9
Mother's Education	1100621 (92.8%)	No Education	25.4
		Primary School	40.1
		High School	28.1
		University Degree	6.4
Father's occupation	1088024 (91.7%)	Work-less	25.7
		Private Sector Employee	41.6
		Government Employee	25.3
		Teacher or Lecturer	7.4
Mother's occupation	1099511 (92.7%)	Housewife	89.5
		Private Sector Employee	2.1
		Government Employee	3.1
		Teacher or Lecturer	5.3
Family Income (Yearly in US$)	1090226 (91.9%)	Weak (<2340)	30.8
		Average (2340-3120)	34.0
		Good (3120-4680)	22.2
		Very Good (>4680)	13.0
The Number of Family Members	1102038 (92.9%)	Four or Less	16.3
		Five	21.2
		Six	22.2
		Seven or More	40.3

Year 2006

Table 4.6 shows the description of the family background factors in 2006. Note that for this year, as well as for the consequent years, these factors now include parental education, father's occupation, and family income only. This is in contrast to the year 2005 where the results include mother's occupation as well as the family size of the candidates. Meanwhile, the variable representing parental education in the year 2006 is slightly differently categorized compared to the subsequent years in the sense that "No Education" and "Primary School" are considered as a single category, whereas "Graduate"[10] is consid-

[10]Master or PhD

ered as the forth category. In the later years, we have "No Education" and "Primary School" as separate categories whereas "Graduate" is not considered as a category at all.

Table 4.6: Description of Family Background Factors in 2006
(Total Number of Applicants: 1,147,895)

Variable Name	Response Frequency (Rate)	Values	Percent
Father's Education	1075889 (93.7%)	No Education/Primary School	44.1
		High School	34.9
		University Degree	18.0
		Graduate	3.0
Mother's Education	1083964 (94.4%)	No Education/Primary School	57.5
		High School	30.9
		University Degree	10.3
		Graduate	1.3
Father's occupation	1087951 (94.8%)	Work-less	28.6
		Private Sector Employee	26.6
		Government Employee	31.1
		Teacher or Lecturer	13.7
Family Income (Yearly in US$)	1084695 (94.5%)	Weak (<2860)	22.1
		Average (2860-5720)	30.6
		Good (5720-8580)	30.4
		Very Good (>8580)	16.9

Year 2007

Table 4.7 shows the description of the family background factors in 2007. Similar to the previous years, these factors now include parental education, father's occupation, and family income only. As it can be seen from the Table 4.7, similar to 2005, the level of father's education is much more higher than mother's education.

Year 2008

The description of family background factors for the year 2008 are shown in Table 4.8. Similar to the previous years, again the level of father's education is still higher than mother's education.

Year 2009

Table 4.9 describes the percentage of candidates' family background factors in year 2009.

Table 4.7: Description of Family Background Factors in 2007

(Total Number of Applicants: 1,101,324)

Variable Name	**Response Frequency (Rate)**	**Values**	**Percent**
Father's Education	1053822 (95.7%)	No Education	17.5
		Primary School	35.8
		High School	31.3
		University Degree	15.4
Mother's Education	1059585 (96.2%)	No Education	27.0
		Primary School	39.6
		High School	26.8
		University Degree	6.6
Father's occupation	1026265 (93.2%)	Work-less	17.1
		Private Sector Employee	50.0
		Government Employee	25.8
		Teacher or Lecturer	7.1
Family Income (Yearly in US$)	1048643 (95.2%)	Weak (<3575)	45.5
		Average (3575-6435)	38.9
		Good (6435-9295)	10.6
		Very Good (>9295)	5.0

Table 4.8: Description of Family Background Factors in 2008

(Total Number of Applicants: 1,157,483)

Variable Name	**Response Frequency (Rate)**	**Values**	**Percent**
Father's Education	1106511 (95.6%)	No Education	18.0
		Primary School	35.7
		High School	31.1
		University Degree	15.2
Mother's Education	1112477 (96.1%)	No Education	27.8
		Primary School	39.2
		High School	26.3
		University Degree	6.7
Father's occupation	1078690 (93.2%)	Work-less	17.4
		Private Sector Employee	50.5
		Government Employee	25.9
		Teacher or Lecturer	6.2
Family Income (Yearly in US$)	1102096 (95.2%)	Weak (<3575)	38.5
		Average (3575-6435)	42.1
		Good (6435-9295)	13.1
		Very Good (>9295)	6.3

Table 4.9: Description of Family Background Factors in 2009
(Total Number of Applicants: 1,077,749)

Variable Name	Response Frequency (Rate)	Values	Percent
Father's Education	1041416 (96.6%)	No Education	17.8
		Primary School	35.1
		High School	31.6
		University Degree	15.5
Mother's Education	1046653 (97.1%)	No Education	27.7
		Primary School	38.2
		High School	27.0
		University Degree	7.1
Father's occupation	1015228 (94.2%)	Work-less	18.4
		Private Sector Employee	49.5
		Government Employee	26.1
		Teacher or Lecturer	6.0
Family Income (Yearly in US$)	1036155 (96.1%)	Weak (<3250)	31.6
		Average (3250-5850)	42.3
		Good (5850-8450)	18.1
		Very Good (>8450)	8.0

Trend of Family background factors during 2005-2009

We now present combined results of the family background factors for the years 2005 to 2009. Interestingly, the results of Table 4.10 show a promising trend in the case of parental education in the sense that from 2005 to 2009, the percentage of mothers/fathers with university degrees has increased. As mentioned earlier, the categorization of parental education in 2006 is slightly different from other years which make the comparison complicated.

The ratio of uneducated Iranian males and females in 2006 were 15.3% and 19.7%, respectively[11]. A comparison of these values with the ratio of uneducated WEE applicant's parents indicates that the qualification of parental education of applicants is less than that of the entire of population. This is mainly due to the fact that the parents typically belong to an older generation which has less educational background.

Socioeconomic Status of Province

We now present a description of socioeconomic status of WEE applicants as an environmental factor which is mentioned in the theoretical model in Section 1.3.

[11]The information come from the census report of the statistical center of Iran in year 2011.

Table 4.10: Description of Variables Used in Data Analysis in 2005-2009

Variable Name	Values	Year				
		2005	2006	2007	2008	2009
Father's Education	No Education	16.6%	44.1%	17.4%	18.0%	17.8%
	Primary School	36.3%		35.8%	35.7%	35.1%
	High School	32.1%	34.8%	31.4%	31.1%	31.6%
	University Degree	14.9%	21.0%	15.4%	15.2%	15.5%
Mother's Education	No Education	25.4%	57.5%	27.0%	27.7%	27.7%
	Primary School	40.1%		39.6%	39.2%	38.2%
	High School	28.1%	30.9%	26.8%	26.3%	27%
	University Degree	6.4%	11.7%	6.6%	6.7%	7.1%
Father's occupation	Work-less	25.7%	26.5%	17.1%	17.4%	18.4%
	Private Sector Employee	41.6%	28.6%	50.1%	50.5%	49.5%
	Government Employee	25.3%	31.1%	25.8%	25.9%	26.1%
	Teacher or Lecturer	7.4%	13.7%	7.0%	6.2%	6%
Family Income (Yearly in US$)	Weak (<3120)	30.8%	22.1%	45.5%	38.5%	31.6%
	Average (3120-5510)	34.0%	30.6%	38.9%	42.1%	42.3%
	Good (5510-8060)	22.2%	30.4%	10.6%	13.1%	18.1%
	Very Good (>8060)	13.0%	17.0%	5.0%	6.3%	7.9%
Unemployed Rate, the Statistical Center of Iran - Report	Male	10.0%	10.0%	9.3%	9.1%	10.8%
	Female	17.1%	16.2%	15.8%	16.7%	16.8%
	Total	11.5%	11.3%	10.5%	10.4%	11.9%

Table 4.11 shows the socioeconomic status of applicants in 30 provinces in Iran during 2005 to 2009 which is calculated based on the weighted average of status of county of the candidates where value 1, 2, and 3 correspond to high, mid and low status of county respectively. It can be seen that four provinces, Tehran, Isfahan, Khorasan Razavi, and Azerbaijan East during all five years are better than others, respectively. In contrast, Ilam, Kohgiluyeh and Boyer Ahmad, and Bushehr have (in order) minimum levels of socioeconomic status. As already mentioned, the status of province are coded by NOET which is based on the socioeconomic status of the WEE applicants. That means the status of economy, cultural, medical, and industrial status of the 30 province differ.

4.5. Chance of Entrance at University

We now present and discuss statistically the global chance of entrance at university. The global chance is defined as the number of entrants in each testing group divided by the total number of candidates in that group. Similar to previous subsections, we first present the results for the individual years from

Table 4.11: Socioeconomic Status of Province in 2005 to 2009 [37]

Province of Residence	2005	2006	2007	2008	2009	2005-9
Azerbaijan East	1.87	1.87	1.90	1.89	1.89	1.89
Azerbaijan Western	2.22	2.22	2.26	2.26	2.27	2.25
Ardebil	2.45	2.50	2.51	2.52	2.51	2.51
Isfahan	1.67	1.65	1.66	1.66	1.66	1.66
Ilam	3.00	2.99	3.00	3.00	3.00	3.00
Bushehr	2.84	2.73	2.80	2.78	2.78	2.80
Tehran	1.46	1.46	1.49	1.51	1.53	1.50
Chahar Mahal Bakhtiari	2.36	2.39	2.41	2.42	2.44	2.41
South Khorasan	2.46	2.45	2.51	2.59	2.59	2.55
Khorasan Razavi	1.75	1.78	1.80	1.79	1.82	1.79
North Khorasan	2.62	2.63	2.63	2.63	2.64	2.63
Khuzestan	2.55	2.56	2.56	2.57	2.59	2.57
Zanjan	2.30	2.32	2.34	2.37	2.38	2.35
Semnan	2.18	2.16	2.19	2.20	2.18	2.19
Sistan and Baluchestan	2.68	2.71	2.70	2.71	2.74	2.71
Fars	1.96	1.96	2.01	2.00	2.01	2.00
Qazvin	2.36	2.40	2.43	2.42	2.44	2.41
Qom	2.02	2.02	2.03	2.03	2.03	2.03
Kurdistan	2.65	2.64	2.66	2.64	2.68	2.67
Kerman	2.55	2.56	2.57	2.58	2.60	2.58
Kermanshah	2.53	2.54	2.54	2.54	2.53	2.54
Kohgiluyeh And Boyer Ahmad	2.99	2.98	3.00	3.00	2.99	2.99
Golestan	2.67	2.69	2.71	2.69	2.71	2.70
Gilan	2,40	2.40	2.43	2.44	2.45	2.43
Lorestan	2.52	2.53	2.54	2.54	2.56	2.54
Mazandaran	2.15	2.15	2.16	2.15	2.16	2.15
Central	2.13	2.12	2.13	2.14	2.15	2.13
Hormozgan	2.66	2.68	2.72	2.72	2.71	2.71
Hamedan	2.25	2,25	2,28	2,28	2,29	2.27
Yazd	2.29	2.29	2.32	2.31	2.31	2.31

2005 to 2009 followed by a combined result for all these years.

Year 2005

Table 4.12 shows the number and percentages of candidates and entrants (successful candidates) in each testing group. It can be seen that from 1,186,751 candidates 286,071 persons (24.1%) were accepted at universities; of which 38.6% were from Group I (Mathematics and Physics), 26.0% from Group II (Empirical Sciences) and 35.5% from Group III (Human Sciences). Moreover, the last row shows that the global chance of entrance for Group I, II and III were 33.4%, 19.3% and 21.5%, respectively. In short, Group I candidates had

a higher global chance of entrance at university compared to the other candidates. Furthermore, Figure 4.1 shows that the candidates of Group I were more likely[12] getting admitted to universities.

Table 4.12: The Frequency of Candidates and Entrants by Testing Group in 2005

Testing Group	Mathematics and Physics		Empirical Sciences		Human Sciences		Total	
	No.	%	No.	%	No.	%	No.	%
Candidates	329,829	27.8	384,184	32.4	472,738	39.8	1,186,751	100
Entrants	110,284	38.6	74,303	26.0	101,484	35.5	286,071	100
Glo. Cha.		33.4		19.3		21.5		24.1

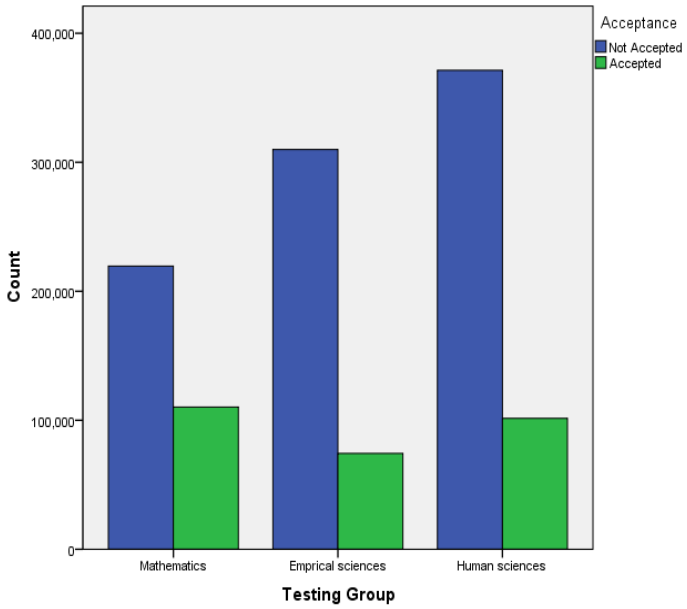

Figure 4.1: The Number of Accepted Students at University by Testing Groups in 2005

Year 2006

In 2006, the actual number of candidates was 1,147,895. Table 4.13 shows the number and percentages of candidates and entrants by each testing group. It can be seen from this table that from the actual number of candidates 409,733 persons (35.7%) were accepted in universities; of which 37.5% were from Group I, 23.4% were from Group II and 39.1% were from Group III respectively. Moreover, the last row of Table 4.13 shows that the chance of entrance for Group I, II and III were 50.2%, 24.8% and 35.2% respectively. It shows that a Group I candidate had a higher chance for entrance at university compared to the other group candidates which can be seen from Figure 4.2 as

[12]The frequencies in Table 4.12 show that among these testing groups, Group I has the maximum capacity (the number of places for students). On the other hand, the number of the candidates in this group is the least.

well.

Table 4.13: The Frequency of Candidates and Entrants by Testing Group in 2006

Testing Group	Mathematics and physics		Empirical sciences		Human sciences		Total	
	No.	%	No.	%	No.	%	No.	%
Candidates	306,138	26.7	385,959	33.6	455,798	39.7	1,147,895	100
Entrants	153,625	37.5	95,758	23.4	160,350	39.1	409,733	100
Glo. Cha.		50.2		24.8		35.2		35.7

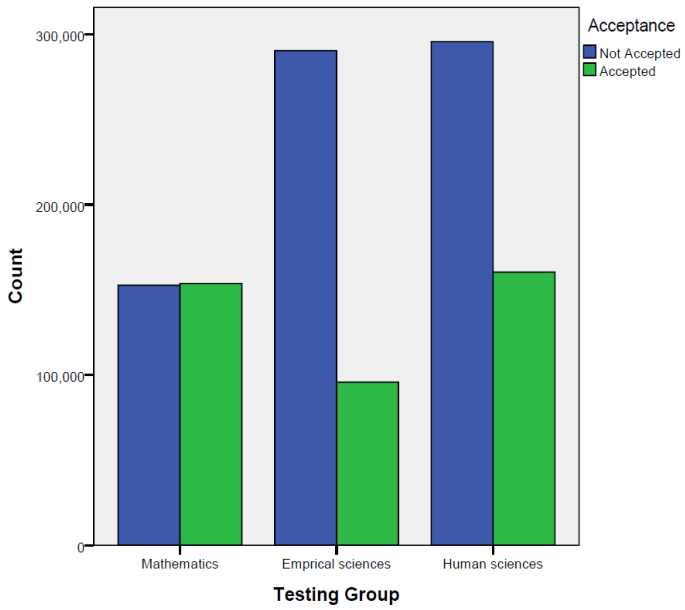

Figure 4.2: The Number of Accepted Students at University by Testing Groups in 2006

Year 2007

In this year, 1,101,324 candidates participated in entrance examination. Table 4.14 shows a number and percentages of candidates and entrants by each group. It can be seen from this table that from 1,101,324 candidates 499,234 persons (45.3%) were accepted at universities; of which 35.4% were from Group I, 25.4% were from Group II and 39.2% were from Group III. This means that the chance of entrance for Group I, II and III were 64.6%, 32.5% and 44.7% respectively. Figure 4.3 shows the Group I candidate had a higher chance for entrance at university compared to the other group candidates as in previous years.

Year 2008

In this year from 1,507,372 candidates we consider 1,157,483 non-duplicated ones only. Table 4.15 shows the number and percentages of candidates and entrants by each group. It can be seen that from 1,157,483 candidates 428,902

Table 4.14: The Frequency of Candidates and Entrants by Testing Group in 2007

Testing Group	Mathematics and physics		Empirical sciences		Human sciences		Total	
	No.	%	No.	%	No.	%	No.	%
Candidates	273,577	24.8	390,286	35.4	437,461	39.7	1,101,324	100
Entrants	176,615	35.4	126,892	25.4	195,727	39.2	499,234	100
Glo. Cha.		64.6		32.5		44.7		45.3

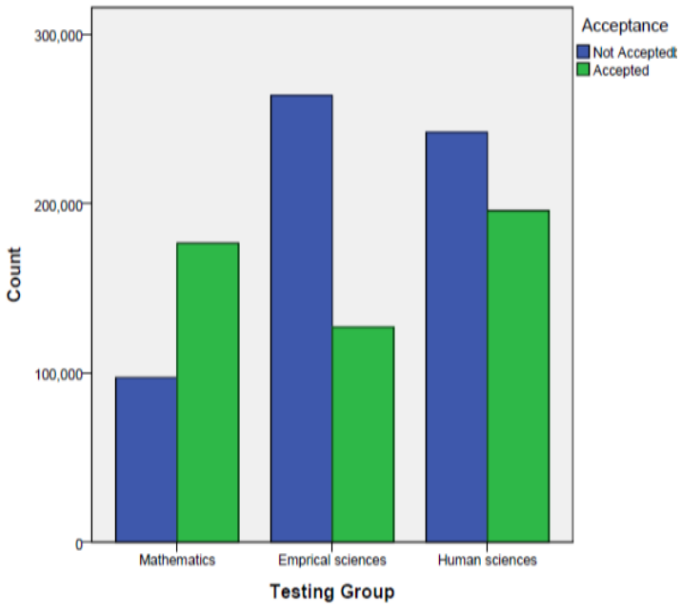

Figure 4.3: The Number of Accepted Students at University by Testing Groups in 2007

persons (37.1%) were accepted at universities; of which 38.5% were from Group I, 25.3% were from Group II and 36.2% were from Group III. It follows that the chance of entrance for Group I, II and III were 58.7%, 26.0% and 33.8% respectively. The Figure 4.4 as well as Table 4.15 shows that a Group I candidate had a higher chance for entrance at university compared to the other group candidates in this year.

Table 4.15: The Frequency of Candidates and Entrants by Testing Group in 2008

Testing Group	Mathematics and physics		Empirical sciences		Human sciences		Total	
	No.	%	No.	%	No.	%	No.	%
Candidates	281,579	24.3	417,028	36.0	458,876	39.6	1,157,483	100
Entrants	165,227	38.5	108,425	25.3	155,250	36.2	428,902	100
Glo. Cha.		58.7		26.0		33.8		37.1

Year 2009

Table 4.16 shows the number and percentages of candidates and entrants by each group. It can be seen from this table that from 1,077,749 candidates

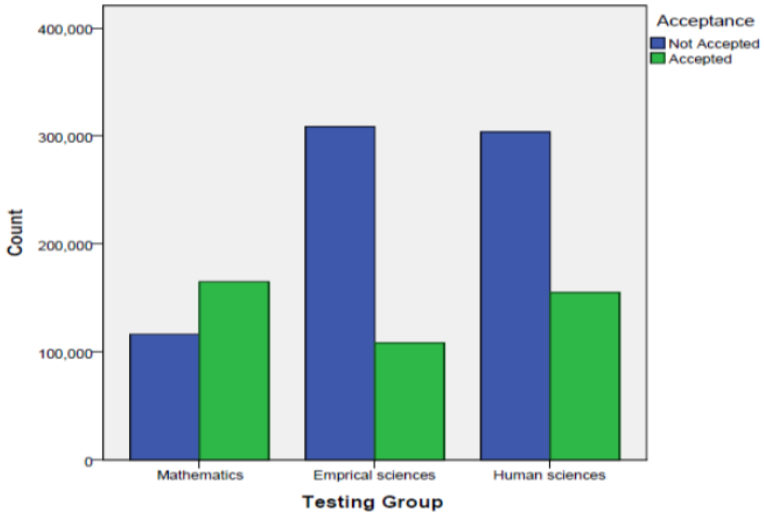

Figure 4.4: The Number of Accepted Students at University by Testing Groups in 2008

517,979 persons (48.1%) were accepted at universities; of which 34.1% were from Group I, 28.2% were from Group II and 37.7% were from Group III. It means that relatively the chance of entrance for Group I, II and III were 64.5%, 36.7% and 48.1% respectively. Figure 4.5 shows that a Group I candidate had a higher chance for entrance at university compared to the other group candidates, similar to the other years.

Table 4.16: The Frequency of Candidates and Entrants by Testing Group in 2009

Testing Group	**Mathematics and physics**		**Empirical sciences**		**Human sciences**		**Total**	
	No.	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**
Candidates	273,634	25.4	397,760	36.9	406,355	37.7	1,077,749	100
Entrants	176,596	34.1	145,941	28.2	195,442	37.7	517,979	100
Glo. Cha.		64.5		36.7		48.1		48.1

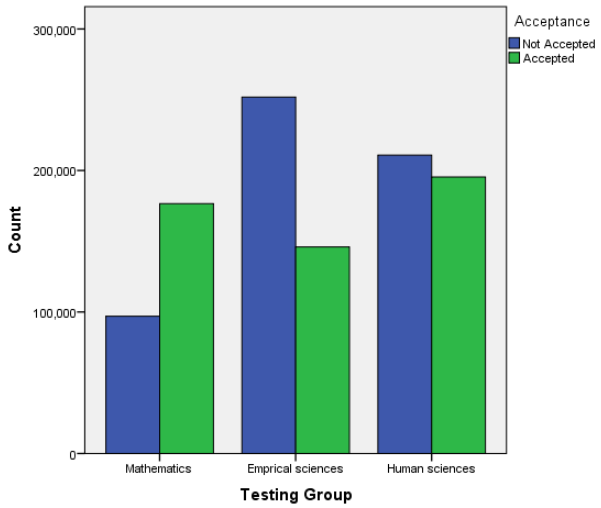

Figure 4.5: The Number of Accepted Students at University by Testing Groups in 2009

Trend of Entrants at University 2005-2009

In this subsection we describe the chance and entrants of WEE applicants in five years (2005-2009) as a trend or time-related factor which is already mentioned in the theoretical part of this study. Table 4.17 shows an overview of the candidates, entrants and chance of entrance for five year datasets according to the three testing groups. It shows that the number of candidates decreases during 2005 to 2009. Meanwhile, the capacity of universities has increased by 80%. Therefore, the chance of acceptance has increased from 24.1% to 48.1%, that means a growth of nearly 100%.

There are three kinds of universities in Iran which use NOET service as entrance exam: public, private, and semi-centralized universities. The public universities are normal universities funded by government and have no tuition. All well-known and highly ranked universities belong to this category. In contrast, the private universities offer the same level of service as public universities, but the students have to pay a tuition. The third group is semi-centralized universities which offer virtual courses and students have to learn courses through self-study Universities in this category have relatively low tuition.

Azad university is another university in Iran which have lots of branched across the country. This university does not use the NOET service, and hence, its entrance examination is out of scope of this thesis. Beside this, there are also another group of institutes which provide a short-tern programs (typically two years) on various topics. The idea is to make the student ready to work in minimum amount of time. The degree provided by these universities is only valid inside Iran and typically indicates low-quality educations. These institutes also have their own entrance exam, and hence are not included in this study.

Table 4.18 shows the number and percentage of entrants into the different kinds of universities during 2005 to 2009. It shows that the semi-centralized universities have maximum capacity for WEE applicants. The reason is that these universities are virtual universities with various educational levels and branches. Establishing a new branch is easily achieved requiring minimum physical structure and staff, hence resulting in greater capacity.

Table 4.17: The Number and Percentage of Candidates and Entrants by Testing Group in 2005 to 2009

	Testing Group	**Mathematics and physics**		**Empirical sciences**		**Human sciences**		**Total**	
	Years	**No.**	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**
Candidates	2005	329,763	27.8	384,149	32.4	472,738	39.8	1,186,650	100
	2006	306,138	26.7	385,959	33.6	455,798	39.7	1,147,895	100
	2007	273,577	24.8	390,286	35.4	437,461	39.7	1,101,324	100
	2008	281,579	24.3	417,028	36.0	458,876	39.6	1,157,483	100
	2009	273,634	25.4	397,760	36.9	406,355	37.7	1,077,749	100
Entrants	2005	110,284	38.6	74,303	26.0	101,484	35.5	286,071	100
	2006	153,625	37.5	95,758	23.4	160,350	39.1	409,733	100
	2007	176,615	35.4	126,892	25.4	195,727	39.2	499,234	100
	2008	165,227	38.5	108,425	25.3	155,250	36.2	428,902	100
	2009	176,596	34.1	145,941	28.2	195,442	37.7	517,979	100
Global Chance	2005	33.4		19.3		21.5		24.1	
	2006	50.2		24.8		35.2		35.7	
	2007	64.6		32.5		44.7		45.3	
	2008	58.7		26.0		33.8		37.1	
	2009	64.5		36.7		48.1		48.1	

Table 4.18: The Frequency of Kind of Acceptance at University in 2005 to 2009

Kind of Uni.	**2005**		**2006**		**2007**		**2008**		**2009**		**2005-9**	
	No.	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**
Public University	113819	9.6	122219	10.6	120237	10.9	110756	9.6	121613	11.3	382840	6.8
Private University	50847	4.3	70492	6.1	87675	8.0	87758	7.6	121441	11.3	418213	7.4
Semi-Cen. University	121405	10.2	217021	18.9	291322	26.5	230388	19.9	274925	25.5	1135061	20.0
Accepted	285971	24.1	409732	35.7	499234	45.3	428902	37.1	517979	48.1	2141918	37.8
Not Acc.	900579	75.9	738163	64.3	602090	54.7	728581	62.9	559770	51.9	3529183	62.2
Total	1186550	100	1147895	100	1101324	100	1157483	100	1077749	100	5671101	100

4.6. Summary

The results provided in this chapter show a statistical description and qualification of the used data in this work. The general information and results are provided to analyze the impact of socioeconomic background factors of the WEE applicants in Iran. These are the first steps of the process[13] of this comprehensive study, which are shown in the upcoming chapters. Moreover, the descriptive statistical outputs show the trends of entrance at university and the status of the socioeconomic factors for the five years (2005 to 2009).

[13]Such as business and data understanding, data preparation.

5. Static Descriptional Aspects

In the previous chapter we provided a descriptive statistical analysis of our data. We now describe the data from a inferential statistics point of view by discussing several statistical and data mining methods[1] described in Chapter 3.

5.1. Analysis of Variance

In the following section, we use the total grade[2] of each candidate as a dependent variable and calculate cross-tabulations between the total grade and the other independent factors by gender. The analysis of variance methods are used to find the different effects of family background factors on the educational achievement of candidates. In the subsections 5.1.1 to 5.1.6 we use one-way ANOVA and each time consider only one factor as independent variable. In subsection 5.1.7 we use factorial ANOVA and consider all factors as independent factors[3]. Among the different factors, we first consider parental education.

5.1.1. Parental Education

In this section we present the result of the analysis of variance on parental education from 2005 to 2009. As already described in the previous chapter, we consider four levels of parental education; *no education, primary education, high school education and university education*[4]. Depending on the education level of parents the candidates can be categorized into four groups where each group corresponds to one level of parental education.

[1] For all evaluation and calculation in this work we used several tools and software. For the statistical analysis we used the SPSS version 21. For the data mining techniques the Clementine package is used as well.

[2] The distribution of target variable is investigated by the goodness of fit test as a Normal distribution. Moreover, the other ANOVA assumptions are investigated as well, such as homogeneity of variance test.

[3] See Section 4.4 for more information about the used factors.

[4] Except for the year 2006, for the details see Section 4.4.

Table 5.1: Frequency Table of Candidates Corresponding to the Parental Education (2005-2009)

Year	Gender	Father's Education									
		No Education		Primary School		High School		University Degree		Total	
		No.	%	No.	%	No.	%	No.	%	No.	%
2005	Male	77,751	19.0	137,060	33.5	122,443	29.9	72,381	17.7	409,635	37.4
	Female	104,193	15.2	261,061	38.0	229,688	33.5	91,231	13.3	686,173	62.6
	Total	181,944	16.6	398,121	36.3	352,131	32.1	163,312	14.9	1,095,808	100
2006	Male	175,879	43.9	125,430	31.3	81,845	20.4	17,341	4.3	400,495	37.2
	Female	298,799	44.2	249,448	36.9	111,676	16.5	15,471	2.3	675,394	62.8
	Total	474,678	44.1	374,878	34.8	193,521	18.0	32,812	3.0	1,075,889	100
2007	Male	76,120	20.0	124,010	32.6	109,780	28.9	70,590	18.6	380,500	36.1
	Female	107,825	16.0	252,998	37.6	220,424	32.7	92,075	13.7	673,322	63.9
	Total	183,945	17.5	377,008	35.8	330,204	31.3	162,665	15.4	1,053,822	100
2008	Male	85,474	21.0	134,610	33.0	115,622	28.4	71,784	17.6	407,490	36.8
	Female	113,796	16.3	260,693	37.3	228,023	36.6	96,509	13.8	699,021	63.2
	Total	199,270	18.0	395,303	35.7	343,645	31.1	168,293	15.2	1,106,511	100
2009	Male	82,681	20.6	132,085	33.0	115,303	28.8	70,458	17.6	400,527	38.5
	Female	102,666	16.0	233,847	36.5	213,784	33.4	90,592	14.1	640,889	61.5
	Total	185,347	17.8	365,932	35.1	329,087	31.6	161,050	15.5	1,041,416	100
Total 2005 - 2009	Male	497,905	24.9	653,195	32.7	544,993	27.3	302,554	15.1	1,998,647	37.2
	Female	727,279	21.6	1,258,047	37.3	1,003,595	29.7	385,878	11.4	3,374,799	62.8
	Total	1,225,184	22.8	1,911,242	35.6	1,548,588	28.8	688,432	12.8	5,373,446	100
Mothers's Education											
2005	Male	115,564	28.1	148,887	36.2	112,829	27.5	33,712	8.2	410,992	37.3
	Female	164,169	23.8	292,985	42.5	196,293	28.5	36,182	5.2	689,629	62.7
	Total	279,733	25.4	441,872	40.1	309,122	28.1	69,894	6.4	1,100,621	100
2006	Male	226,238	56.1	117,446	29.1	51,248	12.7	8,457	2.1	403,389	37.2
	Female	396,697	58.3	217,040	31.9	60,956	9.0	5,482	0.8	680,175	62.8
	Total	622,935	57.5	334,486	30.9	112,204	10.4	13,939	1.3	1,083,564	100
2007	Male	114,507	30.0	135,013	35.3	98,971	25.9	33,599	8.8	382,090	36.1
	Female	171,996	25.4	284,258	42.0	184,764	27.3	36,477	5.4	677,495	63.9
	Total	286,503	27.0	419,271	39.6	283,735	26.8	70,076	6.6	1,059,585	100
2008	Male	127,738	31.2	144,573	35.3	101,875	24.9	34,879	8.5	409,065	36.8
	Female	180,906	25.7	291,808	41.5	190,943	27.1	39,755	5.7	703,412	63.2
	Total	308,644	27.7	436,381	39.2	292,818	26.3	74,634	6.7	1,112,477	100
2009	Male	125,582	31.3	140,202	34.9	101,415	25.2	34,654	8.6	401,853	38.4
	Female	164,678	25.5	259,693	40.3	181,290	28.1	39,139	6.1	644,800	61.6
	Total	290,260	27.7	399,895	38.2	282,705	27.0	73,793	7.1	1,046,653	100
Total 2005 - 2009	Male	709,629	35.4	686,121	34.2	466,338	23.2	145,301	7.2	2,007,389	37.2
	Female	1,078,446	31.8	1,345,784	39.6	814,246	24.0	157,035	4.6	3,395,511	62.8
	Total	1,788,075	33.1	2,031,905	37.6	1,280,584	23.7	302,336	5.6	5,402,900	100

Table 5.1 shows the number and rate of candidates in each year corresponding to the four categories[5] of father and mother education, respectively, according to the gender.

In the following, we present the results of the analysis of variance corresponding to the different years. For each year we compare the mean of total grade of applicants with the level of parental education, as well as gender.

Year 2005

As can be seen from Table 5.1, for the year 2005, the percentages of candidates according to the father's education in four categories are 16.6%, 36.3%, 32.1% and 14.9%, respectively. The corresponding values in the same year with respect to mother's education levels are 25.4%, 40.1%, 28.1% and 6.4%, respectively. These percentages show that the level of father's education is slightly higher than the mother's education level.

Table 5.2 presents a comparison between father's education and mother's education for the year 2005. The analysis of variance[6] shows that the father's and mother's education have a positive effect on the total grades of candidates. With an increase in the father's/mother's level of education, the total grades of applicants increase. For example, the means of total grades in the four categories of father's education are 5060.8, 5146.0, 5229.4, and 5789.8. This pattern is the same for both sexes and the only difference is that the mean of total grades for female applicants is slightly higher than for male applicants.

Although in this study we analyze all applicants in one year, the final goal is to come up with a model to predict the behavior for the new applicants in the next year(s). Therefore, we have to estimate the marginal mean according to the information obtained from all applicants in this year by assuming that these are just some samples of the entire population. Then we can predict the marginal mean for the rest of the applicants which participate in WEE in the upcoming year(s).

The estimated marginal mean can be obtained according to the methodology described in Section 3.1. The estimated marginal means of the total grade according to the father's and mother's education in 2005 are illustrated in

[5]The labels of these categories for the year 2006 is deffer than the other years which is shown in Table 4.6

[6]The F values and significance levels of the Table 5.3 indicate the positive effects of the parental education on the total grade of applicants.

Figure 5.1. Due to the modeling inaccuracy, these results are slightly different from those presented in Table 5.2. However, the amount of inaccuracy is always below 5%. This clearly reveals the effectiveness of the GLM in estimating the total grade.

Table 5.2: The Mean of Total Grades of Applicants by Parental Education and the Gender of Candidates in 2005

	No Educations		**Primary School**		**High School**		**Uni. Degree**		***F Value***		**Sig.**	
	Father	**Mother**	**Father**	**Mother**	**Father**	**Mother**	**Father**	**Mother**	**Father**	**Mother**	**Father**	**Mother**
Male	5074.9	5101.9	5114.4	5143.1	5141.9	5241.3	5644.0	5742.1	2293.6	1580.4	0.000	0.000
Female	5050.2	5071.0	5162.6	5204.2	5276.1	5407.4	5905.5	6181.2	7601.9	7017.1	0.000	0.000
Total	5060.8	5083.8	5146.0	5183.6	5229.4	5346.8	5789.8	5969.5	9086.2	7529.1	0.000	0.000

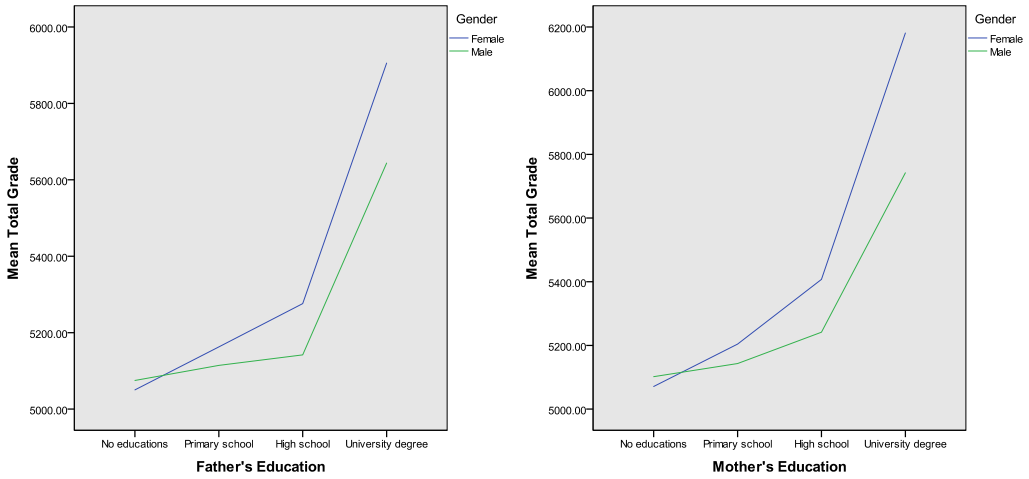

Figure 5.1: The Marginal Mean of Total Grades by Parental Education and the Gender of Candidates in 2005

Year 2006

For this year, Table 5.3 shows the results of ANOVA according to the father's and mother's education. As already mentioned the categories of the parental education levels for the year 2006 is different from the other years[7]. The marginal means of the total grade of the candidates according to parental education for year 2006 can be seen in Figure 5.2.

[7]For the detail see Section 4.4

Table 5.3: The Mean of Total Grades of Applicants by Parental Education and the Gender of Candidates in 2006

	No Educations/ Primary School		High School		Uni. Degree/ Graduate		*F Value*		Sig.	
	Father	Mother	Father	Mother	Father	Mother	Father	Mother	Father	Mother
Male	5040.6	5065.2	5084.6	5175.8	5470.8	5496.3	2424.0	1664.2	0.000	0.000
Female	5138.1	5168.3	5284.1	5395.8	5808.3	5952.2	9436.4	8586.3	0.000	0.000
Total	5102.0	5130.9	5217.3	5318.6	5660.4	5736.4	10428.5	8559.1	0.000	0.000

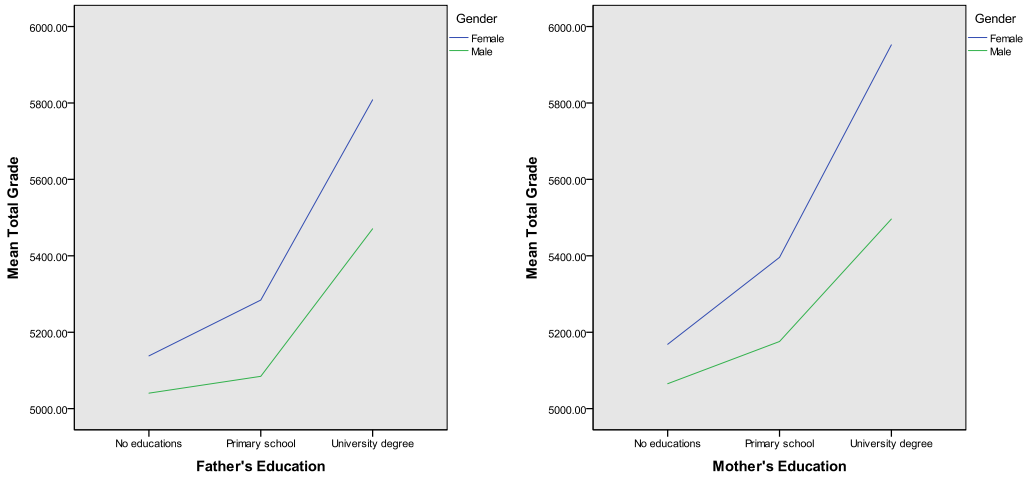

Figure 5.2: The Marginal Mean of Total Grades by Parental Education and the Gender of Candidates in 2006

Year 2007

For the year 2007, Table 5.4 shows the analysis of variance results according to the father's and mother's education. Figure 5.3 shows the marginal means of the total grade according to the parental education in 2007.

Table 5.4: The Mean of Total Grades of Applicants by Parental Education and the Gender of Candidates in 2007

	No Educations		Primary School		High School		Uni. Degree		*F Value*		Sig.	
	Father	Mother	Father	Mother	Father	Mother	Father	Mother	Father	Mother	Father	Mother
Male	5186.3	5264.8	5304.5	5376.7	5454.3	5602.6	6009.8	6122.2	4006.6	2803.9	0.000	0.000
Female	5359.0	5387.2	5560.9	5598.6	5751.4	5896.0	6419.3	6702.5	10214.4	9634.9	0.000	0.000
Total	5284.9	5338.3	5473.6	5527.1	5651.4	5793.6	6244.6	6424.3	13469.9	11297.7	0.000	0.000

Year 2008

The result of analysis of variance according to the parental education for 2008 is summarized in Table 5.5. Similarly, Figure 5.4 shows the marginal means

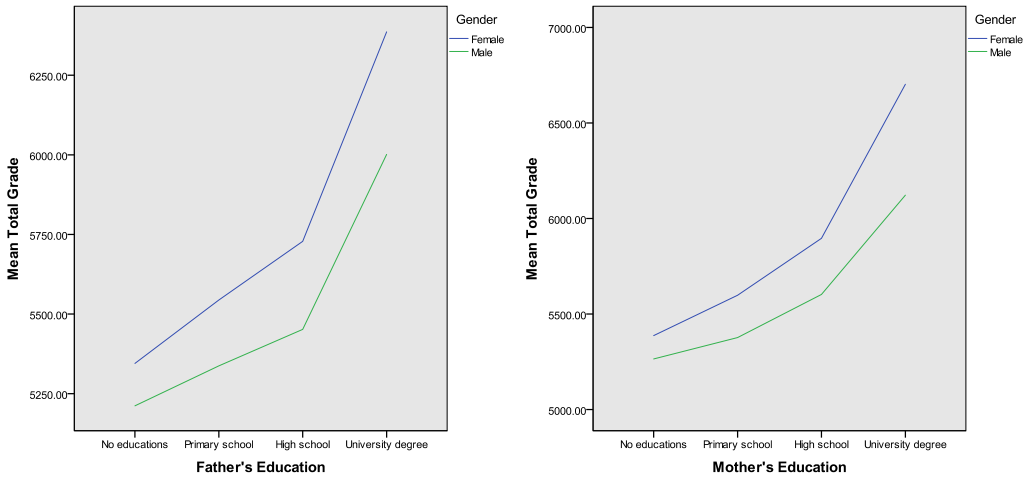

Figure 5.3: The Marginal Mean of Total Grades by Parental Education and the Gender of Candidates in 2007

of the total grade of candidates according to the parental education for year 2008.

Table 5.5: The Mean of Total Grades of Applicants by Parental Education and the Gender of Candidates in 2008

	No Educations		Primary School		High School		Uni. Degree		*F Value*		Sig.	
	Father	Mother	Father	Mother	Father	Mother	Father	Mother	Father	Mother	Father	Mother
Male	5186.3	5237.6	5304.5	5346.5	5454.3	5601.1	6009.8	6162.5	4006.6	3490.0	0.000	0.000
Female	5359.0	5411.1	5560.9	5610.1	5751.4	5918.9	6419.3	6737.8	10214.4	10318.1	0.000	0.000
Total	5284.9	5339.3	5473.6	5522.7	5651.4	5808.4	6244.6	6468.9	13469.9	12865.4	0.000	0.000

Year 2009

Table 5.6 presents the result of analysis of variance according to the father's and mother's education, respectively, whereas Figure 5.5 shows the marginal means of the total grade according to both father's and mother's education.

Table 5.6: The Mean of Total Grades of Applicants by Parental Education and the Gender of Candidates in 2009

	No Educations		Primary School		High School		Uni. Degree		*F Value*		Sig.	
	Father	Mother	Father	Mother	Father	Mother	Father	Mother	Father	Mother	Father	Mother
Male	5260.5	5303.0	5352.7	5396.7	5471.4	5595.2	6015.7	6165.1	3501.9	2982.1	0.000	0.000
Female	5407.8	5456.7	5611.4	5665.1	5798.2	5952.6	6455.5	6756.9	9474.4	9653.7	0.000	0.000
Total	5342.1	5390.2	5518.0	5571.0	5683.7	5824.4	6263.1	6479.0	12142.4	11640.5	0.000	0.000

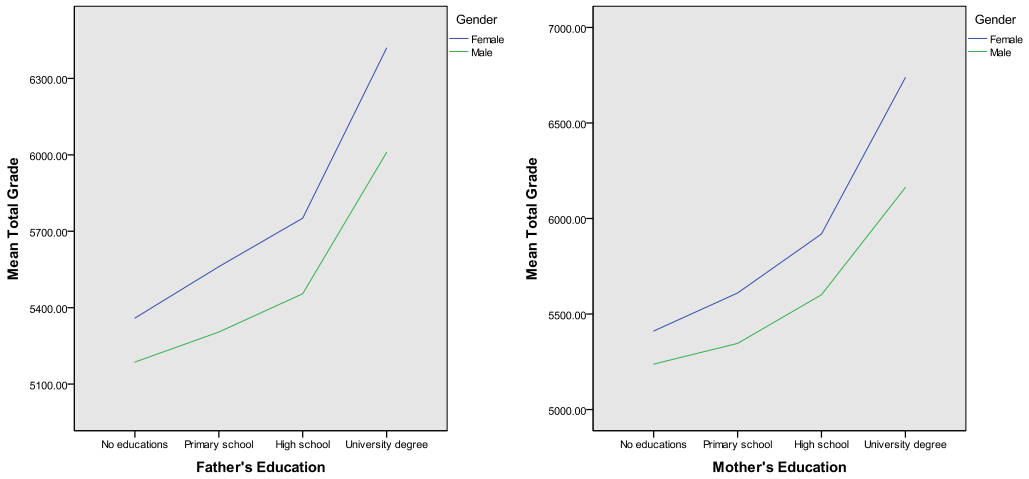

Figure 5.4: The Marginal Mean of Total Grades by Parental Education and the Gender of Candidates in 2008

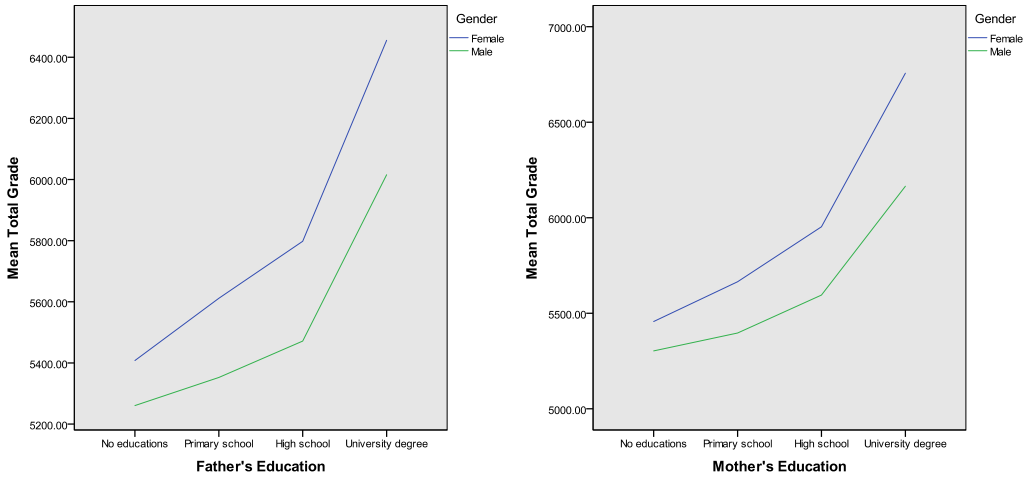

Figure 5.5: The Marginal Mean of Total Grades by Parental Education and the Gender of Candidates in 2009

Conclusion

From the above results, it can be seen that in every year the education of both parents has a positive effect on the total grades of the candidates. In other words, the education level of parents affects the means of the total grades of both male and female candidates. We thus conclude that the higher the education level of a candidate's parent, the greater is the chance of the candidate's entering into universities. However, interestingly, the mean of the total grades for female candidates is higher than that for male candidates. Specially, this difference is clear in years 2008 and 2009.

Table 5.7: The Mean of Total Grades of Applicants by Parental Education and the Gender of Candidates (2005-2009)

Gender	Parental Education	Years									
		2005		2006		2007		2008		2009	
		Father	Mother	Father	Mother	Father	Mother	Father	Mother	Father	Mother
Male	No Educations	5074.9	5101.9	5040.6	5065.2	5211.5	5264.8	5186.3	5237.6	5260.5	5303.0
	Primary School	5114.4	5143.1			5337.3	5376.7	5304.5	5346.5	5352.7	5396.7
	High School	5141.9	5241.3	5084.6	5175.8	5451.7	5602.6	5454.3	5601.1	5471.4	5595.2
	Uni. Degree	5644.0	5742.1	5470.8	5496.3	6001.5	6122.2	6009.8	6162.5	6015.7	6165.1
	F Value	2293.6	1580.4	2424.0	1664.2	3483.9	2803.9	4006.6	3490.0	3501.9	2982.1
	Sig.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Female	No Educations	5050.2	5071.0	5138.1	5168.3	5344.9	5387.2	5359.0	5411.1	5407.8	5456.7
	Primary School	5162.6	5204.2			5543.9	5598.6	5560.9	5610.1	5611.4	5665.1
	High School	5276.1	5407.4	5284.1	5395.8	5728.2	5896.0	5751.4	5918.9	5798.2	5952.6
	Uni. Degree	5905.5	6181.2	5808.3	5952.2	6386.3	6702.5	6419.3	6737.8	6455.5	6756.9
	F Value	7601.9	7017.1.	9436.4	8586.3	9603.1	9634.9	10214.4	10318.1	9474.4	9653.7
	Sig.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Total	No Educations	5060.8	5083.8	5102.0	5130.9	5289.7	5338.3	5284.9	5339.3	5342.1	5390.2
	Primary School	5146.0	5183.6			5476.0	5527.1	5473.6	5522.7	5518.0	5571.0
	High School	5229.4	5346.8	5217.3	5318.6	5636.3	5793.6	5651.4	5808.4	5683.7	5824.4
	Uni. Degree	5789.8	5969.5	5660.4	5736.4	6219.3	6424.3	6244.6	6468.9	6263.1	6479.0
	F Value	9086.2	7529.1	10428.5	8559.1	12194.5	11297.7	13469.9	12865.4	12142.4	11640.5
	Sig.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 5.7 provides a collective view of the results for all these years along with the values of the F statistic. As indicated by these values, for the father's eduction level is higher than the mother's education level in male and total group candidates. However, in the female group during 2007-2009 the effects is vice versa. Although, the means of total grades effect of mother's education is a higher than of that effect of the father's education in female group candidates which is one of the supported hypotheses of this work.

5.1.2. Parental Occupation

In this section we present the results of the analysis of variance on parental occupation for five years (2005-2009). Similar to the parental education, we consider four levels of occupation for the parents namely, *unemployed, private sector, government employee,* and *teacher or lecturer*. We categorize the candidates into four groups according to their parents' occupations.

The number of applicants (corresponding to the four groups) participating in WEE for each year is summarized in Table 5.8. In this section, we analyze the variance of the means of candidates total grades according to the parental

occupation in each year.

Year 2005

Unlike in other years, there are two additional family background variables in 2005, i.e. mother's occupation and family size. Table 5.9 shows the results of ANOVA where father's occupation and mother's occupation is considered.

The variable representing mother's occupation in contrast to the father's occupation is differently categorized in the sense that "Housewife" is considered instead of "unemployed" category.

Figure 5.6 provides another view of the results by displaying the estimated marginal means of total grades with respect to parental occupation in 2005.

Table 5.8: Frequency Table of Candidates Corresponding to the Father's Occupation in 2005-2009

Year	Gender	Father's Occupation									
		Unemployed		Private Sector		Government Employee		Teacher or Lecturer		Total	
		No.	%	No.	%	No.	%	No.	%	No.	%
2005	Male	109,351	26.9	158,944	39.0	103,052	25.3	35,694	8.8	407,041	37.4
	Female	170,679	25.1	293,131	43.0	172,292	25.3	44,881	6.6	660,983	62.6
	Total	280,030	25.7	452,075	41.6	275,344	25.3	80,575	7.4	1,088,024	100
2006	Male	105,017	25.9	122,368	30.2	114,575	28.2	63,837	15.7	405,797	37.3
	Female	206,463	30.3	166,424	24.4	223,818	32.8	85,449	12.5	682,154	62.7
	Total	311,480	28.6	288,792	26.5	338,393	31.1	149,286	13.7	1,087,951	100
2007	Male	73,430	19.7	173,144	46.6	94,830	25.5	30,444	8.2	371,848	36.2
	Female	102,101	15.6	340,294	52.0	170,096	26.0	41,926	6.4	654,417	63.8
	Total	175,531	17.1	513,438	50.0	264,926	25.8	72,370	7.1	1,026,265	100
2008	Male	79,523	19.9	188,672	47.3	102,420	25.7	28,242	7.1	398,857	37.0
	Female	107,719	15.8	355,974	52.4	176,998	26.0	39,142	5.8	679,833	63.0
	Total	187,242	17.4	544,646	50.5	279,418	25.9	67,384	6.2	1,078,690	100
2009	Male	82,681	20.9	182,373	46.5	101,471	25.9	25,963	6.6	391,888	38.6
	Female	104,955	16.8	319,944	51.3	163,925	26.3	34,516	5.5	623,340	61.4
	Total	187,036	18.4	502,317	49.5	265,396	26.1	60,479	6.0	1,015,228	100
Total 2005 - 2009	Male	449,402	22.7	825,501	41.8	516,348	26.1	184,180	9.3	1,975,431	37.3
	Female	691,917	20.8	1,475,767	44.4	907,129	27.3	245,914	7.4	3,320,727	62.7
	Total	1,141,319	21.5	2,301,268	43.5	1,423,477	26.9	430,094	14.6	5,296,158	100

Table 5.9: The Mean of Total Grades of Applicants by Parental Occupation and the Gender of Candidates in 2005

Father's Occupation						
	Unemployed	**Private Sector**	**Gov. Emp.**	**Teacher or Lecturer**	***F Value***	**Sig.**
Male	5160.8	5130.1	5238.9	5632.9	1042.6	0.000
Female	5190.1	5212.3	5360.4	5829.6	2886.0	0.000
Total	5178.6	5183.4	5314.9	5742.5	3673.1	0.000
Mother's Occupation						
	Housewife	**Private Sector**	**Gov. Emp.**	**Teacher or Lecturer**	***F Value***	**Sig.**
Male	5166.0	5102.9	5336.3	5743.1	1156.5	0.000
Female	5233.1	5186.1	5533.9	6125.0	4280.8	0.000
Total	5208.7	5147.9	5447.7	5952.2	4917.4	0.000

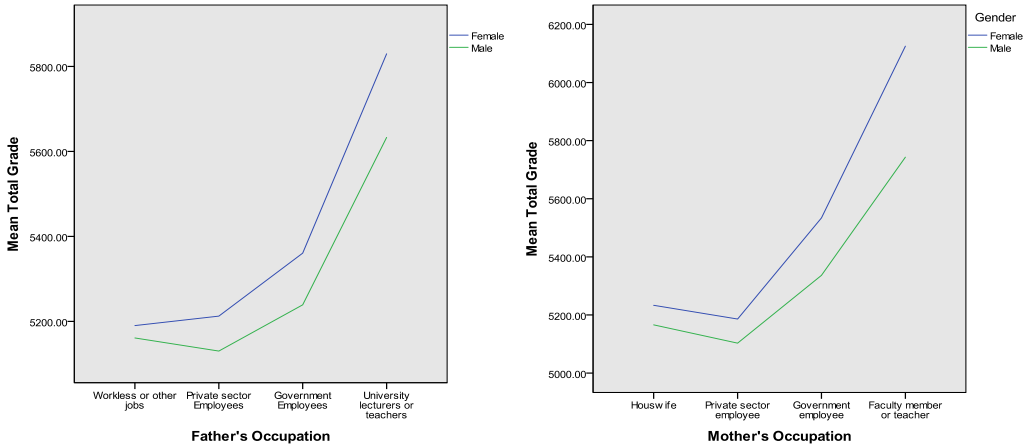

Figure 5.6: The Marginal Mean of Total Grades by Parental Occupation and the Gender of Candidates in 2005

Year 2006

For 2006 and consequent years we consider father's occupation only, as we didn't have any data regarding the mother's occupation. Table 5.10 shows the mean of total grades of applicants, whereas Figure 5.7 shows the marginal mean of the total grades, both relative to father's occupation.

Table 5.10: The Mean of Total Grades of Applicants by Father's Occupation and the Gender of Candidates in 2006

	Unemployed	**Private Sector**	**Gov. Emp.**	**Teacher or Lecturer**	***F Value***	**Sig.**
Male	5025.3	5132.2	5185.2	5395.7	704.4	0.000
Female	5213.2	5308.8	5311.2	5620.2	1533.2	0.000
Total	5149.9	5233.9	5268.5	5524.2	2024.0	0.000

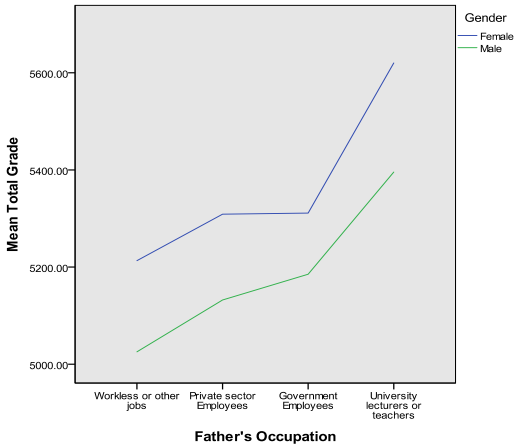

Figure 5.7: The Marginal Mean of Total Grades by Father's Occupation and the Gender of Candidates in 2006

Year 2007

Table 5.11 shows the mean of total grades of applicants relative to father's occupation in this year. A marginal means of the total grade of the candidates according to the father's occupation for 2007 can be seen in Figure 5.8.

Table 5.11: The Mean of Total Grades of Applicants by Father's Occupation and the Gender of Candidates in 2007

	Unemployed	**Private Sector**	**Gov. Emp.**	**Teacher or Lecturer**	***F Value***	**Sig.**
Male	5145.9	5483.7	5562.5	5910.4	1821.1	0.000
Female	5297.2	5690.9	5803.8	6241.5	4546.3	0.000
Total	5233.9	5621.0	5717.4	6102.2	6301.0	0.000

Year 2008

The mean of total grades of applicants relative to their father's occupation for 2008 is given in Table 5.12. Figure 5.9 shows the marginal means of total grades with respect to the father's occupation.

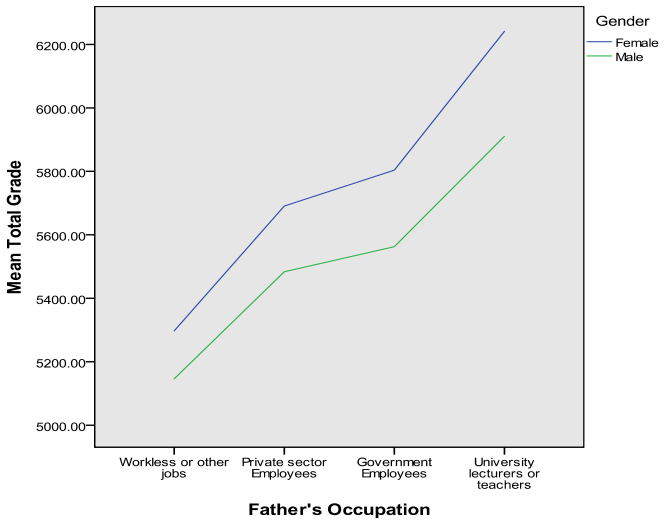

Figure 5.8: The Marginal Mean of Total Grades by Father's Occupation and the Gender of Candidates in 2007

Table 5.12: The Mean of Total Grades of Applicants by Father's Occupation and the Gender of Candidates in 2008

	Unemployed	**Private Sector**	**Gov. Emp.**	**Teacher or Lecturer**	***F Value***	**Sig.**
Male	5148.1	5437.1	5566.2	5833.5	1608.0	0.000
Female	5333.0	5716.0	5830.0	6243.7	4188.4	0.000
Total	5254.4	5626.3	5733.3	6071.8	5762.3	0.000

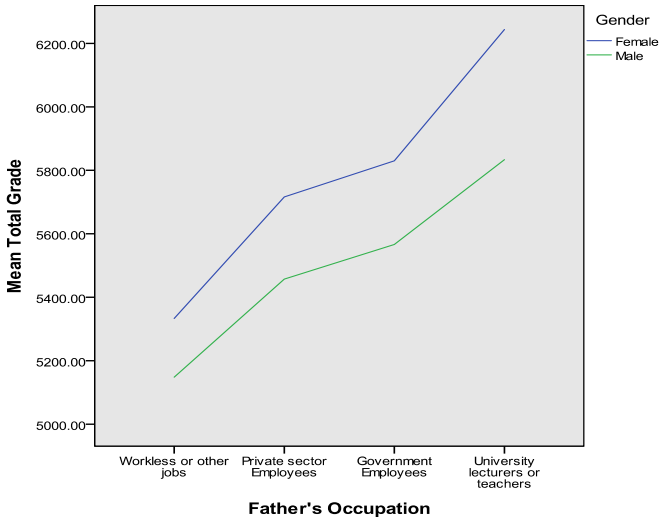

Figure 5.9: The Marginal Mean of Total Grades by Father's Occupation and the Gender of Candidates in 2008

Year 2009

Table 5.13 and Figure 5.10 respectively show the mean and marginal means of total grades of applicants relative to their father's occupation in this year.

Table 5.13: The Mean of Total Grades of Applicants by Father's Occupation and the Gender of Candidates in 2009

	Unemployed	**Private Sector**	**Gov. Emp.**	**Teacher or Lecturer**	***F Value***	**Sig.**
Male	5216.9	5505.9	5576.9	5871.4	1399.2	0.000
Female	5391.4	5773.8	5881.8	6281.8	3878.4	0.000
Total	5314.8	5676.5	5765.2	6105.6	5230.7	0.000

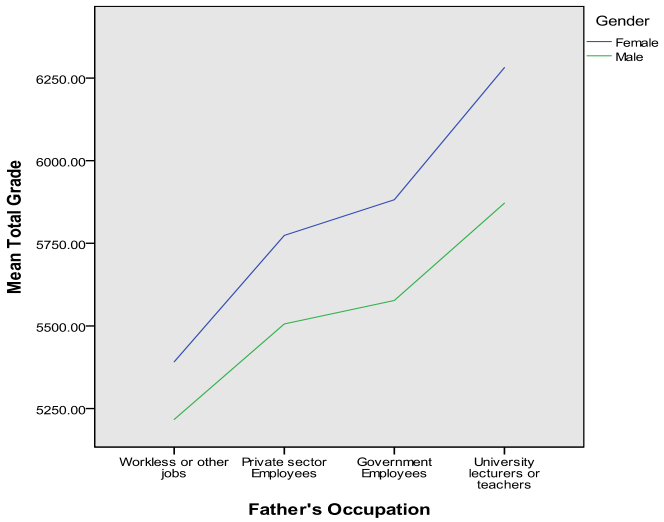

Figure 5.10: The Marginal Mean of Total Grades by Father's Occupation and the Gender of Candidates in 2009

Conclusion

For the year 2005, similar to other years, the mean of total grades for applicants whose fathers are teachers or university lecturers is more than other categories. However, the mean of total grades for applicants whose fathers are government employees is better compared to the remaining two groups. Probably government employees provide better quality of care to their children and try to construct a supportive environment by encouraging them to study. The same pattern can be observed for the mean of total grades of applicants relative to their mothers' occupation in 2005. This means that the mean of total

grades for applicants with lectures/teacher or government employee mother is higher than the other groups.

From the above results it can be seen that for the years 2005 to 2009, the mean of total grades of applicants whose fathers are teachers or university lecturers is more than the other categories. Meanwhile, the mean of total grades for the applicants whose fathers are governmental employees is greater compared to the other two groups. As we already mentioned, the reason probably is that the government employees provide better quality of care and supportive environment for their children.

In Table 5.14, we present a collective view of our results of ANOVA according to the father's occupation by gender for the five years. This table shows that the mean of total grade of applicants is increasing according to the father's occupation level during the years 2005 to 2009. One could conclusively say that the results indicate a growth in the educational performance of WEE applicants during these five years. As we already mentioned, this pattern is the same for both sexes. The only difference is that the mean of total grades for females is slightly higher than males according to parents' occupation.

5.1.3. The Number of Family Members in 2005

As already mentioned, unlike other years, we have for the year 2005 additionally the number of family members. Table 5.15 shows the mean of total grades in 2005 relative to the number of family members and the gender of applicants. Family members include parents and siblings. This variable is categorized into four categories: *four and less, five, six,* and *seven or more* family members. The means of total grades of applicants are 5367.9, 5367.6, 5250.5, and 5150.6 for applicants, respectively.

The analysis of variance shows that the means of total grades of applicants is lower for the candidates from a larger family. That means, the family size has a negative effect on educational achievement.

Figure 5.11 provides yet another view of the result by displaying the estimated marginal means of total grades with respect to the number of family members in year 2005.

Table 5.14: The Mean of Total Grades of Applicants by Father's Occupation and the Gender of Candidates (2005-2009)

Gender	Father's Occupation	Years				
		2005	**2006**	**2007**	**2008**	**2009**
Male	Unemployed	5130.1	5025.3	5145.9	5148.1	5216.9
	Private Sector	5160.8	5132.2	5483.7	5437.1	5505.9
	Gov. Emp.	5238.9	5185.2	5562.5	5566.2	5576.9
	Teacher or Lecturer	5632.9	5395.7	5910.4	5833.5	5871.4
	F Value	1042.6	704.4	1821.1	1608.0	1399.2
	Sig.	0.000	0.000	0.000	0.000	0.000
Female	Unemployed	5190.1	5213.2	5297.2	5333.0	5391.4
	Private Sector	5212.3	5308.8	5690.9	5716.0	5773.8
	Gov. Emp.	5360.4	5311.2	5803.8	5830.0	5881.8
	Teacher or Lecturer	5829.6	5620.2	6241.5	6243.7	6281.8
	F Value	2886.0	1533.2	4546.3	4188.4	3878.4
	Sig.	0.000	0.000	0.000	0.000	0.000
Total	Unemployed	5178.6	5149.9	5233.9	5254.4	5314.8
	Private Sector	5183.4	5233.9	5621.0	5626.3	5676.5
	Gov. Emp.	5314.9	5268.5	5717.4	5733.3	5765.2
	Teacher or Lecturer	5742.5	5524.2	6102.2	6071.8	6105.6
	F Value	3673.1	2024.0	6301.0	5762.3	5230.7
	Sig.	0.000	0.000	0.000	0.000	0.000

Table 5.15: The Mean of Total Grades of Applicants by the Number of Family Members and the Gender of Candidates in 2005

	Four or less	**Five**	**Six**	**Seven or more**	***F Value***	**Sig.**
Male	5303.4	5301.9	5200.9	5115.5	380.1	0.000
Female	5412.2	5409.0	5278.5	5170.4	1266.3	0.000
Total	5367.9	5367.6	5250.5	5150.6	1533.9	0.000

5.1.4. Family Income

In this section we present the result of the analysis of variance on family income considering the years 2005 to 2009. We consider four levels of family income; *weak, average, good, and* very good. Depending on the income level of family, the candidates can be categorized into four groups where each group corresponds to one level of family income. Table 5.16 summarizes the frequencies of student (corresponding to the four groups of family income) candidates in the WEE for years 2005-2009.

We now discuss the analysis of variance of the mean of total grades of WEE candidates according to family income for each year.

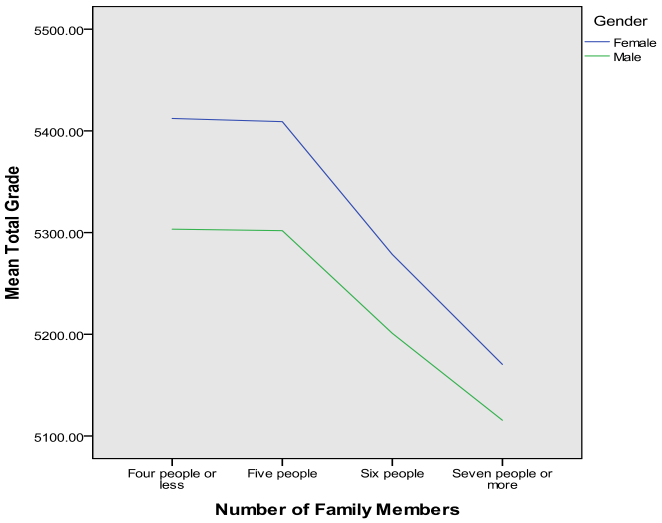

Figure 5.11: The Marginal Mean of Total Grades by Family Size and the Gender of Candidates in 2005

Year 2005

Table 5.17 shows the means of total grades in 2005 in relation to the applicant's family income and their gender. For this year the four different family income categories correspond to annual income of less than 2400$, 2400-3250$, 3250-4800$ and greater than 4800$, respectively. In this year, the means of total grades for applicants in four categories are 5080.3, 5183.9, 5396.3, and 5577.6.

The result shows that the means of total grades for applicants increase with the family income. This pattern is the same for both genders, and the only difference is that the mean of total grades for females is slightly higher than that of males. The ANOVA results clearly shows this fact (see Table 5.17). In other words, the *F values* indicate a positive effect of the family income on the total grades of applicants and consequently on their educational achievement.

Figure 5.12 illustrates the result by displaying the marginal means of total grades with respect to the family income in year 2005.

Year 2006

Table 5.18 shows the means of total grades in 2006 with respect to the applicant's family income and gender. In this year, the four categories (weak, average, good and very good) of family income correspond to the annual income of less 2860$, 2860-5720$, 5720-8580$ and more than 8580$ respectively.

Table 5.16: Frequency Table of Candidates Corresponding to the Family Income in 2005-2009

Year	Gender	Family Income									
		Weak		Average		Good		Very Good		Total	
		No.	%	No.	%	No.	%	No.	%	No.	%
2005	Male	130,795	32.1	131,290	32.2	89,629	22.0	55,805	13.7	407,519	37.4
	Female	204,960	30.0	239,513	35.1	151,952	22.3	86,282	12.6	682,707	62.6
	Total	335,755	30.8	370,803	34.0	241,581	22.2	142,087	13.0	1,090,226	100
2006	Male	107,317	26.6	128,575	31.8	104,156	25.8	64,122	15.9	404,170	37.3
	Female	131,965	19.4	203,467	29.9	225,296	33.1	119,797	17.6	680,525	62.7
	Total	239,282	22.1	332,042	30.6	329,452	30.4	183,919	17.0	1,084,695	100
2007	Male	176,503	46.6	139,603	36.9	40,445	10.7	21,924	5.8	378,475	36.1
	Female	300,773	44.9	268,403	40.1	70,426	10.5	30,566	4.6	670,168	63.9
	Total	477,276	45.5	408,006	38.9	110,871	10.6	52,490	5.0	1,048,643	100
2008	Male	160,265	39.5	163,934	40.4	53,010	13.1	28,628	7.1	405,837	36.8
	Female	264,032	37.9	299,604	43.0	91,334	13.1	41,289	5.9	696,259	63.2
	Total	424,297	38.5	463,538	42.1	144,344	13.1	69,917	6.3	1,102,096	100
2009	Male	129,293	32.5	163,545	33.0	72,037	28.8	33,403	8.4	398,278	38.4
	Female	197,718	31.0	275,225	41.1	115,993	18.1	48,941	7.7	637,877	61.6
	Total	327,011	31.6	438,770	43.1	188,030	18.2	82,344	7.9	1,036,155	100
Total 2005 - 2009	Male	704,173	35.3	726,947	36.5	359,277	18.0	203,882	10.2	1,994,279	37.2
	Female	1,099,448	32.6	1,286,212	38.2	655,001	19.5	326,875	9.7	3,367,536	62.8
	Total	1,803,621	33.6	2,013,159	37.5	1,014,278	18.9	530,757	9.9	5,361,815	100

Table 5.17: The Mean of Total Grades of Applicants by Family Income and the Gender of Candidates in 2005

	Weak	Average	Good	Very good	*F Value*	Sig.
Male	5076.5	5126.9	5336.4	5482.6	1192.4	0.000
Female	5082.6	5215.1	5431.6	5639.0	3964.1	0.000
Total	5080.3	5183.9	5396.3	5577.6	4882.8	0.000

The means of total grades of applicants for these categories are 5154,4, 5157.9, 5280.1 and 5543.8. Figure 5.13 shows the marginal means of total grades with respect to the family income in this year.

Table 5.18: The Mean of Total Grades of Applicants by Family Income and the Gender of Candidates in 2006

	Weak	Average	Good	Very good	*F Value*	Sig.
Male	5075.3	5072.5	5194.4	5423.0	792.8	0.000
Female	5225.1	5206.2	5319.7	5608.4	2101.7	0.000
Total	5157.9	5154.4	5280.1	5543.8	2978.4	0.000

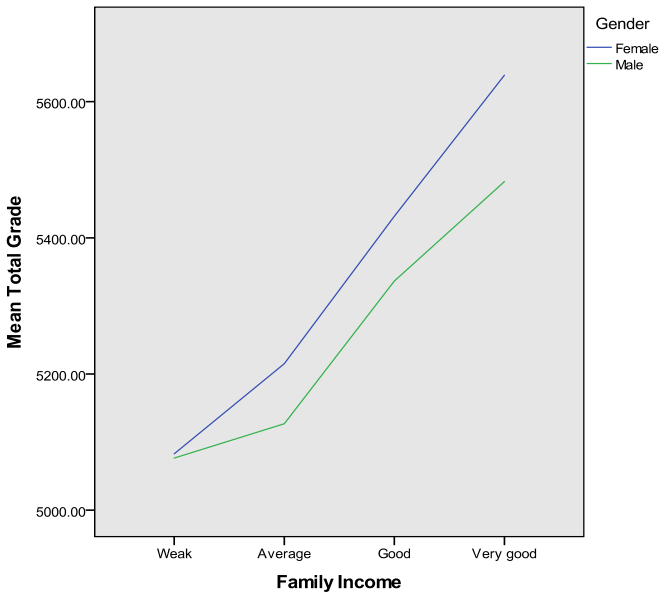

Figure 5.12: The Marginal Mean of Total Grades of Applicants by Family Income and the Gender in 2005

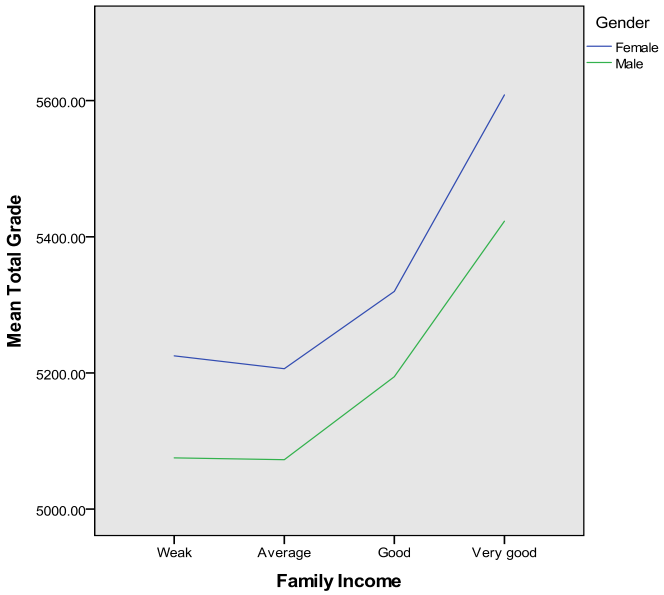

Figure 5.13: The Marginal Mean of Total Grades of Applicants by Family Income and the Gender in 2006

Year 2007

Table 5.19 presents the mean of total grades of applicants regarding their family income and gender in 2007. In this year, families with an annual income less than 3575$, 3575-6435$, 6435-9295$ and more than 9295$ are respectively considered as week, average, good and very good groups. The means of total grades of applicants for these categories are 5398.6, 5677.0, 5996.7, and 6078.5.

It shows that the means of total grades for applicants increase with the family income. Similar to the previous years, again this pattern is the same for both sexes and the only difference is that the means of total grades for females is better than that of males.

Table 5.19: The Mean of Total Grades of Applicants by Family Income and the Gender of Candidates in 2007

	Weak	**Average**	**Good**	**Very good**	***F Value***	**Sig.**
Male	5270.0	5532.1	5851.0	5905.6	2239.7	0.000
Female	5474.0	5752.3	6078.8	6202.5	5046.6	0.000
Total	5398.6	5677.0	5996.7	6078.5	7161.6	0.000

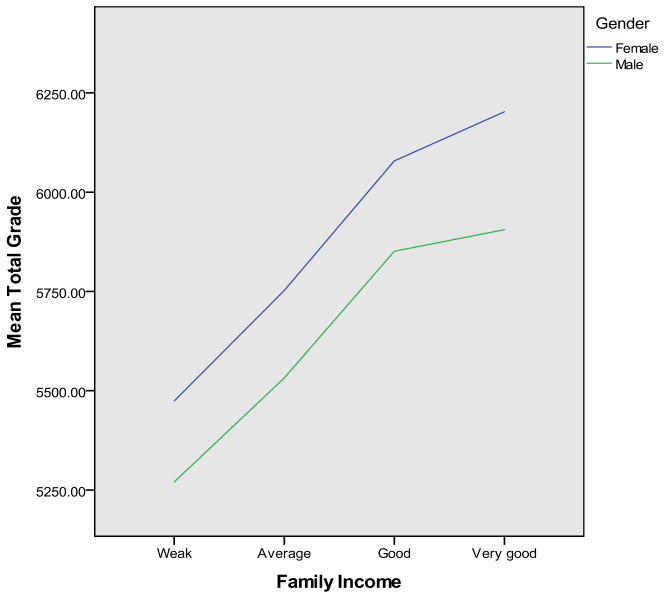

Figure 5.14: The Marginal Mean of Total Grades of Applicants by Family Income and the Gender in 2007

It can been seen from Table 5.19 and Figure 5.14 that the family income has a positive effect on the total grades of candidates.

With an increase in the family income, the total grades of applicants increase. In other words, the result of ANOVA shows that the means of total grades of applicants have gone up with the increase in family income.

Year 2008

For the year 2008, the mean of total grades in relation to the applicants' family income and their gender is shown in Table 5.20. The four family income cate-

gories corresponding to weak, average, good and very good are less than 3570$, 3570-6435$, 6435-9295$ and more than 9295$ annual income respectively.

Again the results show that applicants belonging to good or very good group have comparatively better mean of total grades. Though a similar trend can be observed both for male as well as females, the mean for females is slightly higher than that for the males. In any case, a better means of total grade can be observed with an increase in family income.

The marginal means of total grades with respect to the family income for year 2008 is given in Figure 5.15.

Table 5.20: The Mean of Total Grades of Applicants by Family Income and the Gender of Candidates in 2008

	Weak	**Average**	**Good**	**Very good**	***F Value***	**Sig.**
Male	5231.3	5438.6	5806.9	5984.9	2882.9	0.000
Female	5479.5	5706.4	6058.0	6326.7	5974.3	0.000
Total	5385.8	5611.7	5965.8	6186.8	8673.4	0.000

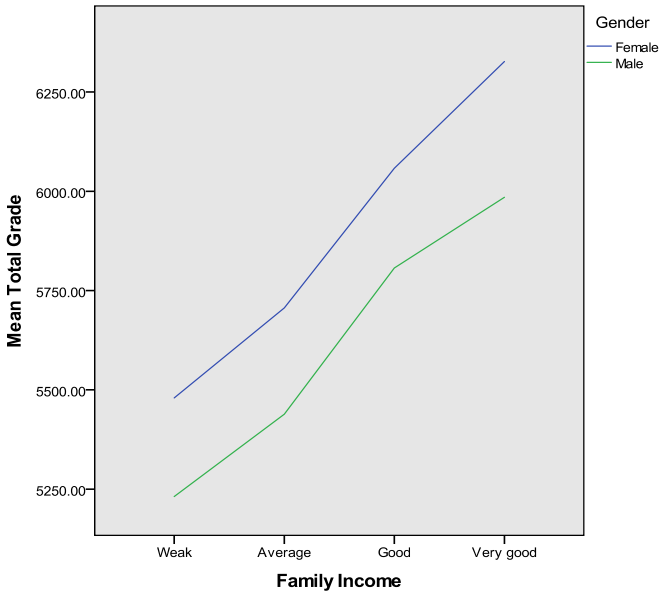

Figure 5.15: The Marginal Mean of Total Grades of Applicants by Family Income and the Gender in 2008

Year 2009

In this year, the groups week, average, good and very good correspond to the families with annual income less 3250$, 3250-5850$, 5850-8450$ and more

than 8450$ respectively. The means of total grades for these four categories are 5439.1, 5579.1, 5921.1, and 6238.0, respectively.

Again it can be seen that in this year, the means of total grades for applicants increase with the family income. From Table 5.21 and Figure 5.16, it can be concluded that candidates' total grades and thus educational performance is positively affected by their family income. This can also be observed from ANOVA results as the means of total grades of applicants grow with the family income.

Table 5.21: The Mean of Total Grades of Applicants by Family Income and the Gender of Candidates in 2009

	Weak	**Average**	**Good**	**Very good**	***F Value***	**Sig.**
Male	5311.3	5392.5	5738.7	6019.8	2580.8	0.000
Female	5522.7	5690.0	6034.5	6387.0	6113.6	0.000
Total	5439.1	5579.1	5921.1	6238.0	8399.7	0.000

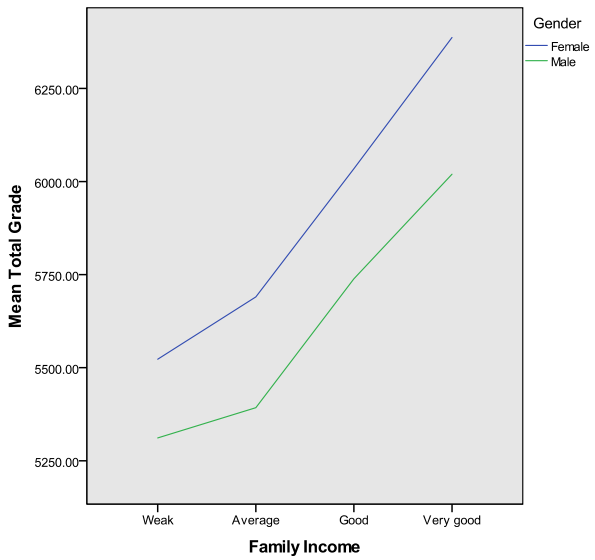

Figure 5.16: The Marginal Mean of Total Grades of Applicants by Family Income and the Gender in 2009

Family Income During 2005-2009

In Table 5.22, we present a collective view of our results for the five years. The result shows that the mean of total grade of applicants approximately is increasing according to the family income level during years 2005 to 2009. The

results indicate a slight improvement in the educational performance of WEE applicants during these five years.

Table 5.22: The Mean of Total Grades of Applicants by Family Income and the Gender of Candidates (2005-2009)

Gender	**Family Income**	**Years**				
		2005	**2006**	**2007**	**2008**	**2009**
Male	Weak	5076.5	5072.5	5270.0	5231.3	5311.3
	Average	5126.9	5075.3	5532.1	5438.6	5392.5
	Good	5336.4	5194.4	5851.0	5806.9	5738.7
	Very good	5482.6	5423.0	5905.6	5984.9	6019.8
	F Value	1192.4	792.8	2239.7	2882.9	2580.8
	Sig.	0.000	0.000	0.000	0.000	0.000
Female	Weak	5082.6	5206.2	5474.0	5479.5	5522.7
	Average	5215.1	5225.1	5752.3	5706.4	5690.0
	Good	5431.6	5319.7	6078.8	6058.0	6034.5
	Very good	5639.0	5608.4	6202.5	6326.7	6387.0
	F Value	3964.1	2101.7	5046.6	5974.3	6113.6
	Sig.	0.000	0.000	0.000	0.000	0.000
Total	Weak	5080.3	5154.4	5398.6	5385.8	5439.1
	Average	5183.9	5157.9	5677.0	5611.7	5579.1
	Good	5396.3	5280.1	5996.7	5965.8	5921.1
	Very good	5577.6	5543.8	6078.5	6186.8	6238.0
	F Value	4882.8	2978.4	7161.6	8673.4	8399.7
	Sig.	0.000	0.000	0.000	0.000	0.000

Conclusion

From the above results, it can be seen that in every year the family income has a positive effect on the total grades of the candidates. In other words, family income affects the means of the total grades of both male and female candidates. We thus conclude that the higher the economic level of a candidate's family, the greater is the chance of the candidate's entering into universities. However, interestingly the mean of total grades for female candidates is slightly higher than that of male candidates. As we already mentioned, probably the high level of socioeconomic status of families enables parents to provide better quality of care and supportive environment for their children.

5.1.5. Age of Participants

We now discuss the *one-way* analysis of variance on the mean of total grades of candidates according to the participant's age for each year. Age of candidates

is categorized into five groups: *younger than 18 years, 18 years, 19 years, 20 to 22 years,* and *older than 22 years*, which is based on the frequencies of the age of applicants.

Years 2005 to 2009

Table 5.23 presents the results of age of participants for the five years 2005 to 2009. The one-way analysis of variance shows an ordering of the mean of total grades of applicants according to their age groups. It shows that the groups "younger than 18 years" and "18 years" have higher total grade than the older age groups. The results show that the young group candidates have better chance of educational achievement than the candidates from the older groups.

The marginal means of the total grades with respect to the age of candidates for the five years are presented in Figure 5.17. This figure shows another view of the result by displaying the marginal means of total grade with respect to the groups of participant's age. It can be seen that the age of participants has a negative effect on the mean of total grades. Consequently it influences the educational performance of applicants. That means that older candidate have a lesser mean of total grades. That is, the mean of total grades decreases with older ages. In other words, young candidates have a better chance to achieve educational attainment than the older applicants.

Conclusion

Table 5.23 provides a collective view of the results for all these years along with *F values.* It can be seen that for the year 2007 and 2008, the age of applicants strongly affects their mean of total grade. Moreover, according to the gender of candidates, the female group has a stronger influence on the mean of total grades than the male group applicants.

In general we can conclude that the increasing age of applicants, as an individual factor, has a negative effect on the students' educational performance.

5.1.6. Region of Residence

As we already mentioned, the region of residence of WEE applicants has influence on their educational performance. In this subsection we discuss the analysis of variance on the mean of total grades of applicants according to the

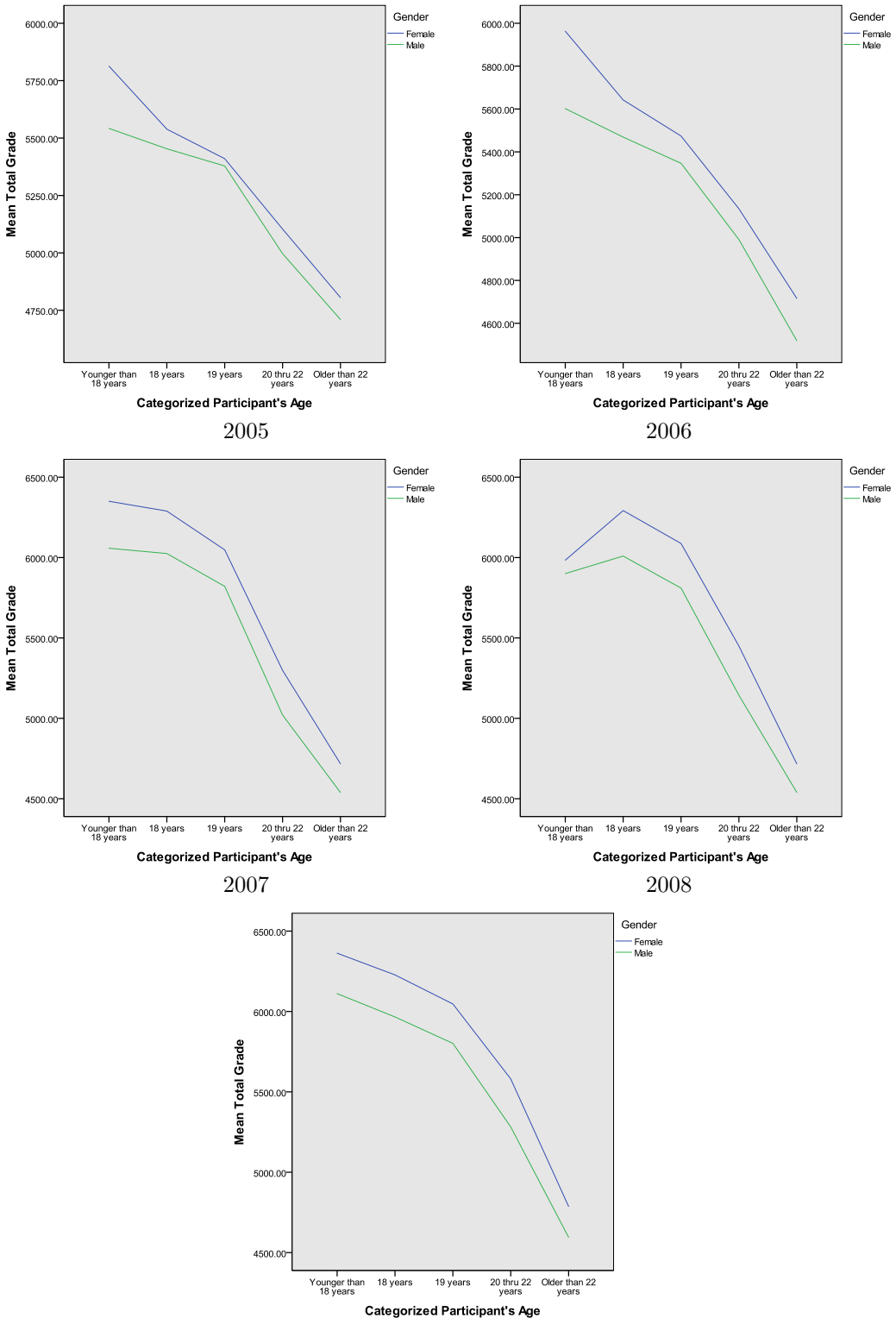

Figure 5.17: The Marginal Mean of Total Grades of Applicants by Age of Participants and the Gender during 2005 to 2009

Table 5.23: The Mean of Total Grades of Applicants by Age of Participants and Gender (2005 - 2009)

Gender	**Age of Participants**	**Years**				
		2005	**2006**	**2007**	**2008**	**2009**
Male	Younger than 18 years	5542.3	5601.8	6058.0	5900.5	5965.3
	18 years	5453.3	5468.7	6025.1	6009.3	5965.9
	19 years	5378.4	5346.4	5821.1	5810.0	5801.9
	20 to 22 years	4996.9	4990.8	5020.3	5143.5	5281.7
	Older than 22 years	4710.0	4519.0	4538.4	4539.1	4594.7
	F Value	5962.0	4332.1	13451.4	13891.2	10524.0
	Sig.	0.000	0.000	0.000	0.000	0.000
Female	Younger than 18 years	5813.3	5962.7	6350.1	5984.0	6362.8
	18 years	5538.8	5642.2	6289.1	6292.21	6227.3
	19 years	5410.2	5474.0	6048.0	6087.4	6047.0
	20 to 22 years	5103.9	5134.6	5298.0	5449.2	5580.1
	Older than 22 years	4805.8	4716.4	4717.2	4716.8	4787.2
	F Value	5088.0	6942.6	26279.9	27390.2	18142.8
	Sig.	0.000	0.000	0.000	0.000	0.000
Total	Younger than 18 years	5670.4	5766.9	6201.8	5941.5	6223.2
	18 years	5506.6	5576.4	6189.6	6184.1	6122.8
	19 years	5398.1	5425.7	5961.7	5981.6	5950.9
	20 to 22 years	5065.5	5083.4	5209.2	5346.5	5477.1
	Older than 22 years	4760.84	4626.3	4642.1	4643.3	47008.2
	F Value	4882.8	2978.4	7161.6	8673.4	8399.7
	Sig.	0.000	0.000	0.000	0.000	0.000

region of residence and gender. Table 5.30 and Table 5.31 present the results of analysis of variance relative to the 30 states/provinces in Iran and the WEE applicants from abroad (i.e., Iranian applicants living outside country). Moreover, the results of homogeneous subset for the province residence of applicants are shown in Figure A.1.

Meanwhile, the ANOVA results show that the means of total grade of applicants from Yazd, Qum, and Khorasan Razavi are increasingly higher than those of candidates from the other provinces. Applicants from Sistan and Baluchestan, Hormozgan, Khuzestan, and Kerman have in decreasing order lower means of total grade in the entire datasets (2005-2009).

Based on the result of analysis of variance, it can be seen that provinces with

better (in order) means of total grades are Yazd, South Khorasan, Isfahan, and Qom. Meanwhile, Sistan and Baluchestan, Hormozgan, Lorestan, Kerman, Kohgiluyeh and Boyer Ahmad, and Ardebil are provinces with unsatisfactory means. Note that, these results can be observed for both genders with slight differences in the sense that female students typically have better means.

In general, the above results show that the province residence of applicants has effects on the educational outcome. As we already mentioned in the previous chapter, the ordering of the provinces as level of socioeconomic status which is coded by NOET is different from the analysis of variance results. To conclude, we believe that based on our results, the government could improve the quality of education in Iran.

5.1.7. Multiple Factorial ANOVA

We now discuss the ANOVA on the whole of the factors as a general linear model[8]. As before we build a separate model for each year from 2005 to 2009.

Year 2005

Table 5.26 presents the results of the analysis of variance using all factors for year 2005. It can be seen that all of the factors do affect the total grade of applicants. The effects of indicated factors are ordered by using the F values. That means, the age of the participant and father's education have a very high effect on the total grade of a candidate, whereas family size and father's occupation have very low effects.

Year 2006

In this year as well as for the consequent years, as we already mentioned, we didn't have any data regarding the mother's occupation and the number of family members. Thus we now discuss only four family background factors which include parental education, father's occupation, and family income. Table 5.27 shows the results of ANOVA for these factors. The most important factors in this year are age and gender, whereas the effects of mother's education and province residence are too low.

[8]For theoretical aspects see Section 3.1.1.

Table 5.24: The Analysis of Variance Results on the Mean of Total Grade in year 2005

Source	Sum of Squares	df	Mean Square	F.	Sig.
Model	2.93E13	54	5.42E11	263797.76	0.000
Categorized Age of Participant	3.84E10	4	9.60E9	4671.83	0.000
Father's Education	8.32E9	3	2.77E9	1349.30	0.000
Gender	2.11E9	1	2.11E9	1027.38	0.000
Family Income	4.75E9	3	1.58E9	771.01	0.000
Province Residence	3.32E10	30	1.11E9	538.87	0.000
Mother's Education	2.92E9	3	9.74E8	473.645	0.000
Mother's Occupation	2.45E9	3	8.15E8	396.66	0.000
Father's Occupation	1.88E9	3	6.28E8	305.667	0.000
No of Family Members	3.65E8	3	1.22E8	59.24	0.000
Error	2.17E12	1054730	2055497.99		
Total	3.14E13	1054784			

Table 5.25: The Analysis of Variance Results on the Mean of Total Grade in year 2006

Source	Sum of Squares	df	Mean Square	F.	Sig.
Model	2.93E13	46	6.38E11	289943.04	0.000
Categorized Age of Participant	5.86E10	4	1.47E10	6664.21	0.000
Gender	6.35E9	1	6.35E9	2888.24	0.000
Family Income	1.58E10	3	5.28E9	2399.90	0.000
Father's Education	1.01E10	2	5.03E9	2287.09	0.000
Father's Occupation	1.35E10	3	4.51E9	2050.90	0.000
Mother's Education	3.69E9	2	1.84E9	838.60	0.000
Province Residence	3.41E10	30	1.14E9	517.23	0.000
Error	2.32E12	1054271	2199794.16		
Total	3.17E13	1054317			

Year 2007

Similar to the previous year, Table 5.28 presents the results of the analysis of variance on the whole of independent factors. The results show that age has a stronger impact of the upon the total grade of the candidates. On the other hand, father's occupation has a lesser effect.

Year 2008

Table 5.29 presents the results of ANOVA for the year 2008. The results are very similar to the results of 2007. Exceptions are the two factors (family

Table 5.26: The Analysis of Variance Results on the Mean of Total Grade in year 2007

Source	Sum of Squares	df	Mean Square	F.	Sig.
Model	3.21E13	48	6.68E11	326161.13	0.000
Categorized Age of Participant	2.20E11	4	5.49E10	26806.74	0.000
Gender	1.54E10	1	1.54E10	7527.84	0.000
Father's Education	7.83E9	3	2.61E9	1273.19	0.000
Mother's Education	7.22E9	3	2.41E9	1174.70	0.000
Province Residence	3.54E10	30	1.18E9	575.77	0.000
Family Income	2.58E9	3	8.62E8	420.435	0.000
Father's Occupation	2.14E9	3	7.14E8	378.19	0.000
Error	2.06E12	1005916	2049664.16		
Total	3.41E13	1005964			

income and province of residence).

Table 5.27: The Analysis of Variance Results on the Mean of Total Grade in year 2008

Source	Sum of Squares	df	Mean Square	F.	Sig.
Model	3.38E13	48	7.05E11	338623.88	0.000
Categorized Age of Participant	2.32E11	4	5.79E10	27832.52	0.000
Gender	1.99E10	1	1.99E10	9567.34	0.000
Father's Education	8.35E9	3	2.78E9	1336.67	0.000
Mother's Education	8.20E9	3	2.73E9	1313.39	0.000
Family Income	5.46E9	3	1.82E9	874.77	0.000
Province Residence	3.48E10	30	1.16E9	558.09	0.000
Father's Occupation	1.02E9	3	3.38E8	162.65	0.000
Error	2.20E12	1058634	2081318.97		
Total	3.603E13	1058682			

Year 2009

Table 5.30 shows the ANOVA results for the year 2009. It can be seen that the order of factors is similar to the previous year ordering though with different F values.

Analysis of Variance During 2005-2009

Finally, Table 5.31 provides a collective view of the ANOVA results for all these years, along with F values.

Table 5.28: The Analysis of Variance Results on the Mean of Total Grade in year 2009

Source	Sum of Squares	df	Mean Square	F.	Sig.
Model	3.22E13	48	6.72E11	324677.16	0.000
Categorized Age of Participant	1.63E11	4	4.07E10	19699.42	0.000
Gender	1.84E10	1	1.84E10	8879.72	0.000
Father's Education	8.19E9	3	2.73E9	1319.36	0.000
Family Income	7.63E9	3	2.54E9	1229.99	0.000
Mother's Education	7.51E9	3	2.50E9	1209.56	0.000
Province Residence	3.49E10	30	1.16E9	561.91	0.000
Father's Occupation	1.37E9	3	4.56E8	220.59	0.000
Error	2.06E12	996868	2068737.30		
Total	3.43E13	996916			

Conclusion

From the results, it can be seen that in every year, all of the independent variables affect the mean of total grades of the WEE candidates. In other words, these results support the theoretical framework we mentioned in the first chapter. That means the constituent variables of the socioeconomic status of the applicants influence the outcomes and affect the students' academic performance.

Table 5.29: The F Valus of ANOVA Results on the Mean of Total Grades as a Dependent Variable During 2005 to 2009

Source \ Years	*F Value*						Sig.
	2005	2006	2007	2008	2009	2005-9	
Father's Education	1349.30	2287.09	1273.19	1336.67	1319.36	6362.42	0.000
Mother's Education	473.65	838.60	1174.70	1313.39	1209.56	4149.71	0.000
Father's Occupation	305.67	2050.90	378.19	162.65	331.81	2245.68	0.000
Mother's Occupation	396.66					1275.71	0.000
Family Income	771.01	2399.90	420.43	874.77	1229.99	1171.92	0.000
No of Family Members	59.24					5863.46	0.000
Gender	1027.378	2888.24	7527.84	9567.34	8879.72	23834.47	0.000
Categorized Age of Participant	4671.83	6664.21	26806.74	28794.30	19699.42	69253.13	0.000
Province Residence	538.87	517.23	575.77	558.09	561.91	2431.65	0.000
Intercept	388586.72	430263.26	555993.18	651389.26	708026.28	2674392.26	0.000
Corrected Model	1428.0	1944.49	4089.37	4365.73	6505.14	12145.47	0.000

Table 5.30: The Mean of Total Grade of Applicants by Province Residence and Gender in 2005-2007

Province of Residence	**2005**			**2006**			**2007**		
	Male	**Female**	**Total**	**Male**	**Female**	**Total**	**Male**	**Female**	**Total**
Abroad	4735.29	5325.63	5049.75	4678.73	5202.79	4955.10	5084.01	5911.6	5579.41
East Azerbaijan	5217.14	5179.66	5194.64	5126.82	5186.40	5162.62	5425.62	5560.99	5508.14
West Azerbaijan	5283.96	5149.50	5209.91	5291.29	5178.95	5228.79	5582.15	5560.91	5570.12
Ardebil	5052.56	5017.84	5032.18	5023.13	5047.61	5037.93	5380.55	5393.47	5388.56
Isfahan	5253.40	5462.88	5385.18	5227.44	5534.49	5422.06	5601.03	5952.05	5829.76
Ilam	5172.63	5268.89	5229.92	5102.97	5286.45	5213.73	5270.95	5502.03	5414.23
Bushehr	4993.14	5070.62	5040.42	4965.04	5124.07	5061.82	5192.57	5469.91	5367.54
Tehran	5233.50	5435.51	5362.64	5212.61	5502.20	5397.81	5603.12	5926.78	5814.12
Chahar Mahal Bakhtiari	5169.33	5305.41	5251.61	5114.25	5299.66	5227.47	5447.82	5662.80	5582.90
South Khorasan	5286.81	5462.04	5390.4	5212.31	5507.68	5386.69	5528.54	5856.58	5726.11
Khorasan Razavi	5406.70	5480.54	5439.09	5358.08	5538.91	5471.68	5659.25	5909.59	5818.32
North Khorasan	5400.91	5335.02	5356.38	5344.43	5314.69	5324.23	5630.05	5629.99	5630.01
Khuzestan	4886.84	5076.08	5009.37	4827.01	5069.74	4985.45	5085.36	5428.61	5316.76
Zanjan	5106.70	5092.54	5098.04	5075.62	5117.25	5101.59	5289.79	5440.52	5384.42
Semnan	5019.05	5328.92	5203.85	5083.63	5506.28	5340.45	5381.38	5922.16	5712.18
Sistan and Baluchestan	4512.03	4626.71	4574.10	4432.26	4636.79	4543.17	4712.40	4939.58	4839.63
Fars	5330.07	5404.03	5376.77	5276.57	5399.07	5354.02	5557.17	5747.22	5681.25
Qazvin	5113.51	5260.64	5208.10	5033.42	5287.39	5196.19	5386.84	5695.19	5589.68
Qom	5422.34	5451.00	5439.09	5378.42	5547.31	5477.29	5755.08	5957.01	5877.06
Kurdistan	5523.76	5294.35	5399.00	5489.00	5370.32	5423.41	5831.29	5735.69	5776.31
Kerman	4838.67	4952.89	4912.55	4815.58	5038.44	4960.06	5157.65	5420.97	5332.32
Kermanshah	5217.22	5331.12	5282.96	5176.19	5352.87	5280.09	5406.21	5679.39	5570.53
Kohgiluyeh and Boyer Ahmad	5217.85	5104.14	5159.86	5132.02	5117.09	5124.26	5376.50	5387.97	5382.75
Golestan	5070.31	5170.10	5131.40	5042.95	5181.55	5128.20	5376.77	5523.31	5469.01
Gilan	5018.29	5039.59	5032.63	5024.15	5119.47	5088.93	5322.45	5472.95	5427.56
Lorestan	5118.17	5053.21	5077.75	5063.98	5089.89	5080.09	5219.51	5378.62	5321.62
Mazandaran	5243.64	5286.85	5270.31	5270.84	5360.72	5326.31	5525.63	5755.68	5670.12
Central	5119.40	5291.70	5227.41	5111.39	5359.22	5270.42	5442.64	5769.99	5655.54
Hormozgan	4730.97	4804.92	4774.74	4614.06	4785.52	4716.60	4837.10	5046.24	4965.08
Hamedan	5222.14	5228.29	5225.90	5157.21	5291.46	5238.49	5304.19	5604.47	5489.14
Yazd	5516.83	5739.74	5644.44	5549.59	5796.19	5693.29	5993.18	6388.55	6225.37
F Value	244.18	503.88	687.53	251.69	532.39	717.07	314.54	703.78	964.27
Significance	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 5.31: The Mean of Total Grade of Applicants by Province Residence and Gender in 2008, 2009, 2005-9

Province of Residence	**2008**			**2009**			**2005-2009**		
	Male	**Female**	**Total**	**Male**	**Female**	**Total**	**Male**	**Female**	**Total**
Abroad	5450.76	5929.13	5710.53	5482.35	5910.23	5721.81	5118.17	5699.60	5439.96
East Azerbaijan	5462.95	5616.40	5554.49	5516.98	5709.40	5628.34	5340.92	5432.58	5395.71
West Azerbaijan	5580.18	5639.92	5613.64	5669.77	5697.80	5685.03	5476.54	5438.84	5455.59
Ardebil	5366.66	5455.58	5421.77	5483.85	5562.05	5531.58	5254.29	5295.80	5279.54
Isfahan	5583.49	5999.11	5851.52	5611.58	6057.26	5892.61	5444.29	5789.95	5664.77
Ilam	5259.60	5496.85	5407.58	5315.48	5557.48	5461.63	5221.27	5420.42	5342.61
Bushehr	5133.76	5473.00	5345.16	5220.73	5570.72	5435.77	5102.11	5350.32	5255.41
Tehran	5614.06	5929.13	5830.63	5616.62	5989.74	5850.83	5443.31	5744.48	5636.04
Chahar Mahal Bakhtiari	5501.51	5717.51	5637.45	5530.65	5776.70	5680.92	5349.83	5554.40	5476.01
South Khorasan	5577.20	5980.71	5815.15	5617.85	6016.46	5850.57	5465.11	5795.36	5660.36
Khorasan Razavi	5637.82	5962.47	5842.36	5702.75	5988.52	5878.14	5549.95	5769.80	5687.91
North Khorasan	5547.11	5720.94	5663.32	5608.96	5758.18	5705.28	5507.42	5549.53	5535.68
Khuzestan	5067.84	5443.48	5317.14	5109.91	5484.07	5838.86	4992.79	5301.32	5195.88
Zanjan	5315.15	5523.26	5444.41	5357.08	5578.78	5491.65	5229.17	5350.47	5304.19
Semnan	5342.64	5829.82	5641.84	5322.35	5761.45	5585.20	5223.57	5664.76	5490.79
Sistan and Baluchestan	4776.68	5019.46	4911.82	4864.61	5089.19	4987.80	4668.67	4878.10	4783.89
Fars	5517.07	5765.18	5677.71	5614.00	5861.03	5772.39	5452.97	5632.46	5568.00
Qazvin	5373.31	5731.75	5604.85	5435.90	5783.75	5655.73	5259.60	5539.81	5440.04
Qom	5693.19	5978.7	5864.15	5747.67	6019.95	5906.82	5594.85	5789.53	5709.98
Kurdistan	5801.20	5785.27	5792.17	5854.73	5826.08	5838.86	5698.33	5607.58	5647.64
Kerman	5144.68	5438.16	5337.87	5235.59	5530.24	5425.66	5040.78	5282.24	5198.29
Kermanshah	5382.15	5681.57	5560.68	5481.24	5770.12	5648.89	5329.22	5560.36	5465.27
Kohgiluyeh and Boyer Ahmad	5315.81	5368.49	5344.68	5392.40	5436.74	5416.40	5287.33	5292.85	5290.28
Golestan	5267.09	5509.39	5417.90	5348.76	5563.19	5478.44	5220.92	5391.81	5326.33
Gilan	5276.00	5495.06	5426.03	5378.89	5560.64	5501.22	5193.53	5326.24	5283.96
Lorestan	5208.61	5361.27	5305.05	5349.47	5493.42	5437.90	5190.86	5274.07	5242.99
Mazandaran	5510.52	5754.01	5662.71	5564.76	5820.15	5720.98	5413.04	5580.42	5516.78
Central	5364.60	5758.47	5616.03	5393.16	5791.67	5641.34	5280.53	5587.27	5475.58
Hormozgan	4877.46	5123.74	5028.30	4887.06	5105.91	5021.78	4792.58	4984.47	4908.97
Hamedan	5305.50	5587.55	5474.18	5371.4	5646.48	5531.81	5273.56	5470.15	5392.01
Yazd	5940.41	6410.87	6214.91	5990.5	6353.23	6197.07	5783.94	6119.92	5978.46
F Value	321.17	724.24	995.21	292.27	623.06	857.88	1334.14	2816.35	3901.55
Significance	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

5.2. Regression Analysis

As already mentioned in Chapter 3, the regression analysis is useful for analyzing continuous dependent and independent variables. Although the data exploited in this study are mostly ordinal independent factors, the regression analysis is employed to investigate the positive/negative effect of independent factors. In this regard, the ordinal variables are coded into numerical values.

In this analysis, the independent variables are coded by NOET as an ordinal factor, such as parental education, parental occupation, family income, family size and the status of county. For instance, the status of county coded by decreasing level. The high, medium and low level of county status are indicated by 1, 2 and 3, respectively. Furthermore, unlike our previous experiments, the age of the applicant is modeled in its original form (i.e., an integer number) rather than the grouped form.

In this section, we consider linear and logistic regression analysis in order to find a prediction model for the academic performance of the applicants. We first consider the linear regression analysis in order to find a linear model to predict the total grade of the applicants. To select the best model we use the stepwise regression method.

5.2.1. Linear Regression

In the following, we present the results of linear regression analysis corresponding to the years from 2005 to 2009.

Year 2005

Table 5.32 shows the results for 2005. It can be seen that the age of candidates is the most important factor affecting the total grades of the applicants. Father's education is the second most important factor. In contrast, the father's occupation has a minimum effect. From Table 5.32 it can be seen that the significance levels of the t test of the mother's education and father's occupation are greater than 0.05. For this reason these factors are not included in the regression model for this year, in contrast to our expectations.

The results of regression analysis for this year have a good coincidence with the results of ANOVA analysis in the same year. The only mismatch is the sign of mother's education in male group which could be due to the low accuracy

Table 5.32: The Results for Linear Regression Model in 2005

	Unstandardized Coefficients		**Standardized Coefficients**		
Variable name	**Estimated Value**	**Std. Error**	**Beta**	**t**	**Sig.**
Age	-80.870	0.628	-0.127	-128.850	0.000
Father's Education	89.944	2.293	0.057	39.232	0.000
Mother's Occupation	143.048	2.129	0.072	67.174	0.000
Status of Province	-206.802	3.339	-0.062	-61.932	0.000
Family Income	72.284	1.601	0.049	45.153	0.000
No. of Family Members	23.194	1.447	0.017	16.033	0.000
Father's Occupation	2.656	1.921	0.002	1.336	0.182
Mother's Education	-2.271	2.447	-0.001	-0.928	0.353
(Constant)	6686.615	16.674	.	401.009	0.000
Female Group					
Father's Education	103.561	2.791	0.066	37.106	0.000
Age	-78.294	0.807	-.119	-96.995	0.000
Mother's Occupation	164.019	2.699	0.081	60.763	0.000
Status of Province	-239.422	3.999	-.075	-59.863	0.000
Family Income	75.500	1.932	0.053	39.083	0.000
No. of Family Members	25.796	1.744	0.020	14.793	0.000
Mother's Education	29.688	2.984	0.018	9.950	0.000
Father's Occupation	7.442	2.342	0.005	3.177	0.001
(Constant)	6572.153	20.843	.	315.314	0.000
Male Group					
Age	-83.615	1.004	-0.135	-83.256	.000
Mother's Occupation	124.610	3.478	0.064	35.831	0.000
Family Income	66.222	2.800	0.043	23.647	0.000
Status of Province	-151.688	5.926	-0.043	-25.597	.000
Father's Education	73.850	3.955	0.046	18.672	0.000
Mother's Education	-50.251	4.213	-0.029	-11.928	0.000
No. of Family Members	15.792	2.541	0.011	6.216	0.000
Father's Occupation	-2.766	3.300	-0.002	-0.838	0.402
(Constant)	6802.096	27.923	.	243.605	0.000

of regression analysis in the study of ordinal variables. For both regression and ANOVA analysis, age and father's education are identified as the most important factors.

As a result of this analysis, the mathematical linear model is shown by the standardized[9] coefficients of the independent factors in 2005,

[9]The regression analysis should be construct by two kind linear models (standardized and unstandardized). The standardized model is constructed without the intercept in the model, whereas the unstandardized regression model is a linear model with a intercept.

$$Total\_Grade = -0.13 \times Age + 0.06 \times Father\_Education + 0.07 \times Mother\_Occupation - 0.06 \times Status\_of\_Province + 0.05 \times Family\_Income + 0.02 \times Family\_Size$$

whereas by the unstandardized coefficients is:

$$Total\_Grade = 6686.62 - 80.87 \times Age + 89.94 \times Father\_Education + 143.05 \times Mother\_Occupation - 206.80 \times Status\_of\_Province + 72.28 \times Family\_Income + 23.19 \times Family\_Size$$

The coefficient of determination of the model calculated by the R square is equal 0.045.

In order to investigate the linear regression model for the male and female applicants, the results are presented by the mathematical models which we presented now. The linear model for the female group in 2005 is:

$$Total\_Grade = 6572.15 + 103.56 \times Father\_Education - 78.29 \times Age + 164.02 \times Mother\_Occupation - 239.42 \times Status\_of\_Province + 75.50 \times Family\_Income + 25.80 \times Family\_Size + 29.69 \times Mother\_Education + 7.44 \times Father\_Occupation$$

The coefficient of determination of this model is $R^2 = 0.055$. For the male group candidates, the linear regression model is:

$$Total\_Grade = 6800.53 - 83.59 \times Age + 124.27 \times Mother\_Occupation + 66.03 \times Family\_Income - 152.06 \times Status\_of\_Province + 72.58 \times Father\_Education - 50.26 \times Mother\_Education + 15.78 \times Family\_Size$$

The coefficient of determination for the male group model is less than the others, with accuracy $R^2 = 0.034$. The above regression analysis in year 2005 shows that the linear model in female group has a maximum accuracy.

As an example, for the case number one from the female group in this year, the mathematical linear model is predicted the mean of total grade of applicant as follows:

$$Total\_Grade = 6572.15 + 103.56 \times Father\_Edu.\,(2) - 78.29 \times Age\,(22) + 164.02 \times Mother\_Occ.\,(1) - 239.42 \times Status\_Pro.\,(2.45) + 75.50 \times Family\_Inc.\,(3) + 25.80 \times Family\_Size\,(4) + 29.69 \times Mother\_Edu.\,(1) + 7.44 \times Father\_Occ.\,(2) = 4982.70$$

Year 2006

Table 5.33 presents the results of the linear regression analysis for the year 2006. Similar to the case in year 2005, for this year the linear model can be constructed by the linear regression model as follows:

$$Total\_Grade = 6851.06 - 96.34 \times Age + 94.68 \times Father\_Education + 114.52 \times Family\_Income - 217.43 \times Status\_of\_Province + 100.25 \times Father\_Occupation + 64.76 \times Mother\_Education$$

The coefficient of determination of this model is $R^2 = 0.057$. Hence, the total grade of applicants can be predicted by independent factors such as age, family background factors, and status of province residence of the WEE candidates. The mathematical linear regression model for the female group of the WEE applicants is presented below:

$$Total\_Grade = 6837.78 - 96.34 \times Age + 112.84 \times Father\_Education - 246.42 \times Status\_of\_Province + 111.00 \times Family\_Income + 100.22 \times Father\_Occupation + 104.26 \times Mother\_Education$$

Here for the female group, the coefficient of determination of the model is $R^2 = 0.068$. Moreover, for the male group of applicants, the accuracy is $R^2 = 0.057$ when considering the following model:

$$Total\_Grade = 6762.74 - 96.34 \times Age + 71.81 \times Father\_Education + 104.57 \times Family\_Income + 104.78 \times Father\_Occupation - 154.544 \times Status\_of\_Province + 26.72 \times Mother\_Education$$

It is clear that the coefficient of determination of female group is better than that of the male group of applicants.

Year 2007

Considering year 2007, the order of the variables according to their importance in the linear regression model is given in Table 5.34. The unstandardized linear regression model for this year is:

$$Total\_Grade = 8825.02 - 161.20 \times Age + 53.52 \times Father\_Education - 251.40 \times Status\_of\_Province + 100.92 \times Family\_Income + 70.80 \times Father\_Occupation + 55.97 \times Mother\_Education$$

For the above regression model, the coefficient of determination is $R^2 = 0.114$. Note that for this year, the regression model has better accuracy than the models for the last years.

We now present the linear regression model constructed for the male and females groups. The model of the female group is as follows:

$$Total\_Grade = 9070.05 - 171.27 \times Age + 87.45 \times Mother\_Education - 286.68 \times Status\_of\_Province + 101.05 \times Family\_Income + 69.84 \times Father\_Occupation + 68.43 \times Father\_Education$$

Table 5.33: The Results for Linear Regression Model in 2006

	Unstandardized Coefficients		**Standardized Coefficients**		
Variable name	**Estimated Value**	**Std. Error**	**Beta**	**t**	**Sig.**
Age	-96.338	0.626	-0.149	-153.879	.000
Father's Education	94.682	1.745	0.069	54.254	0.000
Family Income	114.524	1.455	0.075	78.737	.000
Status of Province	-217.427	3.316	-0.063	-65.562	0.000
Father's Occupation	100.253	1.431	0.067	70.063	0.000
Mother's Education	64.764	2.062	0.040	31.412	0.000
(Constant)	6851.064	16.195	.	423.037	0.000
Female Group					
Age	-96.338	0.801	-0.143	-118.175	.000
Father's Education	112.838	2.150	0.083	52.491	0.000
Status of Province	-246.419	3.988	-0.075	-61.795	0.000
Family Income	110.999	1.783	0.074	62.250	.000
Father's Occupation	100.224	1.724	0.069	58.122	0.000
Mother's Education	104.263	2.596	0.064	40.159	0.000
(Constant)	6837.782	20.308	.	336.700	0.000
Male Group					
Age	-96.605	1.007	-0.154	-95.925	.000
Father's Education	71.808	2.954	0.052	24.306	0.000
Family Income	104.567	2.500	0.066	41.820	.000
Father's Occupation	104.785	2.509	0.066	41.772	0.000
Status of Province	-154.544	5.847	-0.043	-26.433	0.000
Mother's Education	26.721	3.391	0.017	7.880	0.000
(Constant)	6762.740	26.934	.	251.085	0.000

For this model, the coefficient of determination is $R^2 = 0.131$. However, for the male group the accuracy of the model given below is $R^2 = 0.093$ hence lower than the accuracy of model for both combined as well as female group.

$$Total\_Grade = 8370.96 - 146.53 \times Age + 106.13 \times Family\_Income - 177.55 \times Status\_of\_Province + 69.21 \times Father\_Occupation + 40.47 \times Father\_Education + 10.52 \times Mother\_Education$$

Comparing the three models, the order of the importance of the independent factors differs slightly. For instance, in female group model, the mother's education is the second most important factor unlike the other two models where this factor has the least effect.

Table 5.34: The Results for Linear Regression Model in 2007

	Unstandardized Coefficients		**Standardized Coefficients**		
Variable name	**Estimated Value**	**Std. Error**	**Beta**	**t**	**Sig.**
Age	-161.198	0.575	-0.272	-280.488	0.000
Father's Education	53.524	2.383	0.033	22.457	0.000
Status of Province	-251.397	3.413	-0.072	-73.652	0.000
Family Income	100.922	1.998	0.054	50.516	0.000
Father's Occupation	70.796	2.218	0.037	31.924	0.000
Mother's Education	55.973	2.343	0.034	25.371	0.000
(Constant)	8825.023	15.625	.	564.809	0.000
Female Group					
Age	-171.269	0.728	-0.282	-235.239	0.000
Mother's Education	87.449	2.850	0.049	30.685	0.000
Status of Province	-286.681	4.072	-0.084	-70.409	0.000
Family Income	101.054	2.409	0.055	41.942	0.000
Father's Occupation	69.844	2.719	0.036	25.683	0.000
Father's Education	68.431	2.891	0.041	23.666	0.000
(Constant)	9070.046	19.338	.	469.018	0.000
Male Group					
Age	-146.532	0.934	-0.257	-156.879	0.000
Family Income	106.128	3.498	0.056	30.340	0.000
Status of Province	-177.551	6.099	-0.048	-29.113	0.000
Father's Occupation	69.212	1.431	0.036	18.322	0.000
Father's Education	40.472	4.133	0.025	9.793	0.000
Mother's Education	10.521	4.076	0.006	2.581	0.010
(Constant)	8370.962	26.449	.	316.491	0.000

Year 2008

Table 5.35 shows the regression analysis result for the year 2008. For this year, the linear regression model for prediction of the total grade of the applicants (both male and female) is as follows:

$$Total\_Grade = 8626.15 - 149.43 \times Age + 116.26 \times Family\_Income - 249.75 \times Status\_of\_Province + 62.94 \times Father\_Education + 57.21 \times Mother\_Education + 46.90 * Father\_Occupation$$

The coefficient of determination for this model is $R^2 = 0.122$. Now considering the female group, the model is as follows:

$$Total\_Grade = 8880.50 - 159.06 \times Age + 81.69 \times Mother\_Education - 280.47 \times Status\_of\_Province + 114.58 \times Family\_Income + 76.84 \times Father\_Education + 50.65 \times Father\_Occupation$$

For this model, the coefficient of determination is $R^2 = 0.139$. Finally the model for the male group is

$$Total\_Grade = 8164.79 - 135.04 \times Age + 125.55 \times Family\_Income - 186.06 \times Status\_of\_Province + 48.56 \times Father\_Education + 38.19 \times Father\_Occupation + 19.63 \times Mother\_Education$$

with an accuracy of $R^2 = 0.101$. Comparing these models we see that for the female group, mother's education is the second most important factor whereas it is the the second least and the least important factor for the combined and male group respectively.to improve the accuracy of the results

Table 5.35: The Results for Linear Regression Model in 2008

	Unstandardized Coefficients		**Standardized Coefficients**		
Variable name	**Estimated Value**	**Std. Error**	**Beta**	**t**	**Sig.**
Age	-149.435	0.497	-0.283	-300.623	0.000
Family Income	116.280	1.911	0.064	60.863	0.000
Status of Province	-249.755	3.363	-0.070	-74.265	0.000
Father's Education	62.937	2.326	0.038	27.055	0.000
Mother's Education	57.208	2.294	0.032	24.943	0.000
Father's Occupation	46.898	2.209	0.024	21.229	0.000
(Constant)	8626.149	14.383	.	599.744	0.000
Female Group					
Age	-159.056	0.630	-0.295	-252.463	.000
Mother's Education	81.693	2.803	0.046	29.149	0.000
Status of Province	-280.466	4.048	-0.081	-69.290	0.000
Family Income	114.579	2.321	0.064	49.356	.000
Father's Education	76.838	2.848	0.046	26.983	0.000
Father's Occupation	50.646	2.731	0.026	18.544	0.000
(Constant)	8880.501	17.788	.	499.234	0.000
Male Group					
Age	-135.038	0.804	-0.264	-167.867	.000
Family Income	125.553	3.294	0.068	38.117	.000
Status of Province	-186.056	5.898	-0.050	-31.548	0.000
Father's Education	48.562	3.963	0.030	12.253	0.000
Father's Occupation	38.188	3.704	0.019	10.310	0.001
Mother's Education	19.626	3.921	0.011	5.005	0.000
(Constant)	8164.791	24.249	.	336.707	0.000

Year 2009

Table 5.36 shows the result for the year 2009. Similar to the previous years, the linear prediction model for this year can be constructed by the standardized coefficients as follows:

$$Total\_Grade = 8394.22 - 136.75 \times Age + 124.88 \times Family\_Income - 256.61 \times Status\_of\_Province + 62.34 \times Father\_Education + 34.01 \times Mother\_Education + 34.01 \times Father\_Occupation$$

The coefficient of determination for this model is$R^2 = 0.106$. Again distinguishing males and females, the linear model for the female group is

$$Total\_Grade = 8540.46 - 140.67 \times Age + 128.59 \times Family\_Income - 300.94 \times Status\_of\_Province + 77.20 \times Father\_Education + 91.70 \times Mother\_Education + 38.09 \times Father\_Occupation$$

with a coefficient of determination of $R^2 = 0.120$, whereas, for the male group candidates the accuracy is $R^2 = 0.086$ when considering the following model:

$$Total\_Grade = 8085.76 - 129.61 \times Age + 123.53 \times Family\_Income - 178.50 \times Status\_of\_Province + 48.63 \times Father\_Education + 25.58 \times Father\_Occupation + 12.95 \times Mother\_Education$$

Again note that the accuracy rate for the male group model is less than the accuracy of the female group model as well as the accuracy of the combined group. These results indeed further support our previous analysis (ANOVA).

Conclusion

It can be seen from the above linear regression models that almost in every year all individual, environmental and family background factors have effect on the prediction of the total grade of WEE candidates. Meanwhile, the sign of *Beta* values shows that some factors have a positive effect and others have a negative effect on the target variable. For instance, the effect of age of applicants negatively affects the mean of total grade.

Note that from the considered models, the status of province apparently has negative effect. This is due to the fact that a lower integer is used to represent a higher status of province and a vice versa i.e., NOET assigns high (1), mid (2) or low (3) to each county in the provinces based on the socioeconomic status of the county. Thus the 'negative effect' in this case should be oppositely interpreted. This would imply that the status of province indeed has a positive effect on the total grade of applicants.

Furthermore, similar to the ANOVA results in the previous section, gender has a negative effect on the total grade of applicants. As mentioned, the gender

Table 5.36: The Results for Linear Regression Model in 2009

	Unstandardized Coefficients		**Standardized Coefficients**		
Variable name	**Estimated Value**	**Std. Error**	**Beta**	**t**	**Sig.**
Age	-136.754	0.525	-0.253	-260.422	0.000
Family Income	124.883	1.979	0 .073	63.102	0.000
Status of Province	-256.606	3.477	-0.073	-73.794	0.000
Father's Education	62.336	2.380	0.038	26.194	0.000
Mother's Education	60.309	2.320	0 .035	25.991	0.000
Father's Occupation	34.008	2.289	0.018	14.858	0.000
(Constant)	8394.217	15.001	.	559.567	0.000
Female Group					
Age	-140.670	0.672	-0.257	-209.403	.000
Family Income	128.593	2.429	0.076	52.942	.000
Status of Province	-300.943	4.246	-0.088	-70.880	0.000
Father's Education	77.202	2.953	0.047	26.142	0.000
Mother's Education	91.700	2.872	0.053	31.930	0.000
Father's Occupation	38.094	2.869	0.020	13.278	0.000
(Constant)	8540.462	18.766	.	455.110	0.000
Male Group					
Age	-129.607	0.835	-0.247	-155.127	.000
Family Income	123.526	3.344	0.071	36.940	.000
Status of Province	-178.501	5.926	-0.049	-30.121	0.000
Father's Education	48.626	3.952	0.030	12.304	0.000
Father's Occupation	25.581	3.744	0.013	6.832	0.000
Mother's Education	12.955	3.872	0.008	3.346	0.001
(Constant)	8085.760	24.728	.	326.992	0.000

code 1 and 2 correspond to Female and Male respectively. The negative effect of gender shows that the mean of total grade of female candidates is better than that of the male group. Additionally, other family background factors have effects on the total grade of WEE applicants which is supported by analysis of variance results as well.

As a remark on the quality of the obtained linear regression models, the coefficient of determination of all regression models is low, varying between 0.034 and 0.139. Therefore, we exploit data mining techniques to improve the accuracy of our prediction models. The data mining techniques will be explained in Section 5.3.2.

5.2.2. Logistic Regression

To predict the chance of entering into university, in this section we use the logistic regression and apply the stepwise approach. Similar to the previous section, our independent variables are *parental education, parental occupation,*

family income, family size and the *status of province*. In the following we present the results of the logistic regression analysis corresponding to the five different years 2005-2009.

For the mathematical logistic regression models, we assumed that $\pi = p(Acceptance\_at\_University = 1)$ is the chance of entering into university as a binary response. As already mentioned in Subsection 3.1.3, the linear function for the logit of the probability π is the logarithms of the odds for the multiple explanatory variables (k independent factors) follows:

$$logit(\pi) = log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots \ + \beta_k x_k \tag{5.1}$$

Similar to the linear regression, we investigate the gender of applicants as a dummy variable in this analysis. Which in the logistic regression models are generated/constructed by two groups namely male and female of WEE applicants.

Year 2005

The six family background variables namely *father's* and *mother's education, father's* and *mother's occupation, family income and family size* , and also the *status of province, age of participant* and *gender (as a dummy variable)* are used for analysis in 2005.

Table 5.37 shows how three variables, namely age, status of province (as a decreasing ordering coded) and the number of family members, have negative effects on acceptance in universities, whereas other factors have positive effects on the target variable. For instance, if the age of a candidate increased, he/she would have a lesser chance to get accepted by a university. The variables in Table 5.37 are ordered according to their importance in the model. It can be seen that the father's education is the most important factor affecting the acceptance to universities. After that, age has the maximum effect and the father's occupation has the minimum effect in this model. Furthermore, the sign of B coefficient of family size shows that the number of family members has a negative effect.

The multiple logistic regression model for this year is as follows:

$$logit(\pi) =$$

$$0.96 + 0.11 \times Father\_Education - 0.12 \times Age + 0.10 \times Family\_Income + 0.12 \times Mother\_Occupation -$$

$$0.22 \times Status\_of\_Province + 0.04 \times Mother\_Education - 0.01 \times Family\_Size + 0.01 \times Father\_Occupation$$

The accuracy rate of logistic regression model is calculated by classification accuracy, which in the percentage of correct classification are presented for each model. For this model the classification accuracy[10] is 75.4%.

For the female group of WEE applicants, the mathematical logistic regression model as follows:

$$logit(\pi) =$$
$$0.89 + 0.07\times Mother\_Education - 0.12\times Age + 0.10\times Family\_Income - 0.22\times Status\_of\_Province +$$
$$0.13\times Mother\_Occupation + 0.12\times Father\_Education - 0.01\times Family\_Size + 0.01\times Father\_Occupation$$

As an example for the logistic regression mathematical model, the chance of acceptance to university for the case number one from female group is predicted as follows:

$$logit(\pi) =$$
$$0.89 + 0.07\times Mother\_Edu.\,(1) - 0.12\times Age\,(22) + 0.10\times Family\_Inc.\,(3) - 0.22\times Status\_Pro.\,(2.54) +$$
$$0.13\times Mother\_Occ.(1) + 0.12\times Father\_Edu.\,(2) - 0.01\times Family\_Size\,(4) + 0.01\times Father\_Occ.\,(2) = -1.602$$

then $\pi = \frac{e^{-1.60196}}{1+e^{-1.60196}} = 0.167$.

The classification accuracy for this model is equal 75.5% whereas, for the male group model the accuracy is 75.4%. Moreover, the mathematical logistic model for the male group is as follows:

$$logit(\pi) = 1.08 - 0.11\times Age + 0.10\times Father\_Education - 0.21\times Status\_of\_Province +$$
$$0.12\times Mother\_Occupation + 0.08\times Family\_Income - 0.02\times Family\_Size$$

Comparing the male and female groups, we see that mother's education is the most important factor for the female group. Whereas, this factor is not entered for the male groups' logistic model. That means, for the male group of candidates, the mother's education factor is not important in logistic regression model for the year 2005. Moreover, the father's education factor is as second important factor in this group. In other words, the mother's education affects the acceptance chance of applicants at university for female group, whereas, the father's education is the most important factor for the male group of WEE applicants.

[10]Since the acceptance to a university is a binary variable, the probability value obtained by logistic regression model is classified into accepted [0.5,1] and not accepted [0,0.5).

Table 5.37: The Result for the Logistic Regression Model in 2005

Variable name	Estimated Value	S.E.	Wald	df	Sig.	Exp(B)
Father's Education	0.115	0.004	950.414	1	0.000	1.122
Age	-0.118	0.001	8568.895	1	0.000	0.888
Family Income	0.096	0.003	1405.334	1	0.000	1.101
Mother's Occupation	0.120	0.003	1438.513	1	0.000	1.127
Status of Province	-0.224	0.005	1699.445	1	0.000	0.800
Mother's Education	0.038	0.004	92.413	1	0.000	1.039
No. of Family Members	-0.015	0.002	41.362	1	0.000	0.985
Father's Occupation	0.010	0.003	10.727	1	0.001	1.010
(Constant)	0.964	0.031	978.514	1	0.000	2.621
Female Group						
Mother's Education	0.070	0.005	675.365	1	0.000	1.133
Age	-0.121	0.002	5048.379	1	0.000	0.886
Family Income	0.105	0.003	1402.487	1	0.000	1.111
Status of Province	-0.232	0.007	1155.587	1	0.000	0.793
Mother's Occupation	0.133	0.004	1028.764	1	0.000	1.142
No. of Family Members	-0.013	0.003	19.717	1	0.000	0.987
Father's Education	0.125	0.005	675.365	1	0.000	1.133
Father's Occupation	0.013	0.004	10.520	1	0.001	1.013
(Constant)	0.887	0.040	484.379	1	0.000	2.428
Male Group						
Age	-0.115	0.002	3559.058	1	0.000	0.891
Father's Education	0.105	0.006	305.779	1	0.000	1.111
Status of Province	-0.218	0.009	587.799	1	0.000	0.805
Mother's Occupation	0.104	0.005	459.723	1	0.000	1.110
Family Income	0.082	0.004	376.603	1	0.000	1.085
No. of Family Members	-0.018	0.004	21.369	1	0.000	0.982
Father's Occupation	0.007	0.005	2.104	1	0.147	1.007
Mother's Education	-0.007	0.006	1.341	1	0.247	0.993
(Constant)	1.092	0.048	517.662	1	0.000	2.982

Year 2006

For this year, again the dependent variable is acceptance to universities and colleges, while the independent variables or co-variates are age, status of province, father's occupation, father's education, mother's education, family income, and gender is taken as a dummy variable.

The variables in Table 5.38 are ordered according to their importance in the regression model. It can be seen that in this model the age of the candidate is the most important factor affecting the acceptance to universities and after that father's education has the maximum effect. The mother's education of applicants has the minimum effect. The sign of coefficients implies that age

and status of province (as a decreasing coded) have a negative effect on the acceptance to university, whereas parents' education, family income and father's occupation have positive effects.

For this year the mathematical logistic regression model is as follows:

$$logit(\pi) = 1.13 - 0.10 \times Age + 0.09 \times Father\_Education + 0.08 \times Family\_Income - 0.22 \times Status\_of\_Province + 0.08 \times Father\_Occupation + 0.08 \times Mother\_Education$$

The accuracy rate of this binary logistic regression model is 64.3%. For the female group of applicants, the logistic regression model is:

$$logit(\pi) = 1.06 - 0.10 \times Age + 0.09 \times Mother\_Education + 0.13 \times Family\_Income + 0.09 \times Father\_Occupation - 0.19 \times Status\_of\_Province + 0.09 \times Father\_Education$$

For this model the accuracy rate is equal 64.1%. Meanwhile, for the male group of applicants the mathematical logistic model with accuracy 64.8% is higher than the female group models' accuracy. The binary logistic regression model for the male group of WEE applicants is as follows:

$$logit(\pi) = 1.23 - 0.10 \times Age + 0.08 \times Father\_Education - 0.27 \times Status\_of\_Province + 0.09 \times Family\_Income + 0.09 \times Father\_Occupation + 0.07 \times Mother\_Education$$

Similar to the previous year, the mother's education factor has the least effect on the acceptance to university for male group applicants, whereas, the father's education has the least effect on the target variable in female group of candidates.

Year 2007

The results for year 2007 are presented in Table 5.39. It can be seen that two factors namely age and status of province (with 1, 2 and 3 corresponding to high, medium and low respectively), negatively affect the acceptance chance. The other factors have a positive effect. For instance, if a candidate comes from a family with high level income, he/she has a better chance to get accepted at a university than others with low level of family income. The variables in Table 5.39 are ordered according to their importance in the logistic regression model. In this model, age is the most important factor affecting the acceptance to

Table 5.38: The Result for the Logistic Regression Model in 2006

Variable name	Estimated Value	S.E.	Wald	df	Sig.	Exp(B)
Age	-0.101	0.001	9911.933	1	0.000	0.904
Father's Education	0.084	0.002	1220.338	1	0.000	1.088
Family Income	0.114	0.002	3065.461	1	0.000	1.120
Status of Province	-0.220	0.005	2226.106	1	0.000	0.802
Father's Occupation	0.087	0.002	1862.112	1	0.000	1.091
Mother's Education	0.081	0.003	815.507	1	0.000	1.084
(Constant)	1.127	0.025	2070.554	1	0.000	3.085
Female Group						
Age	-0.104	0.001	5990.496	1	0.000	0.901
Mother's Education	0.095	0.004	647.828	1	0.000	1.099
Family Income	0.126	0.003	2321.946	1	0.000	1.135
Father's Occupation	0.088	0.003	1186.097	1	0.000	1.091
Status of Province	-0.188	0.006	1033.602	1	0.000	0.828
Father's Education	0.090	0.003	833.824	1	0.000	1.094
(Constant)	1.059	0.032	1075.961	1	0.000	2.8884
Male Group						
Age	-0.097	0.002	3873.234	1	0.000	0.908
Father's Education	0.076	0.004	390.041	1	0.000	1.079
Status of Province	-0.273	0.008	1236.088	1	0.000	0.761
Family Income	0.090	0.003	741.164	1	0.000	1.094
Father's Occupation	0.088	0.003	699.956	1	0.000	1.092
Mother's Education	0.066	0.004	223.596	1	0.000	1.092
(Constant)	11287	0.039	1001.248	1	0.000	3.413

universities. It is followed by the status of province. In contrast, the mother's education has the least effect whereas the father's education has the minimum effect on the target variable.

For the year 2007 the logistic regression model is constructed as follows:

$$logit(\pi) = 3.14 - 0.15 \times Age - 0.34 \times Status\_of\_Province + 0.09 \times Mother\_Education +$$
$$0.06 \times Father\_Occupation + 0.06 \times Family\_Income + 0.02 \times Father\_Education$$

The accuracy rate of this model calculated by the classification accuracy is 59.9%.

In order to investigate the effect of gender of WEE applicants again we constructed logistic regression model for the male and female groups of candidates. For the female group of applicants, the logistic regression model is as follows:

$$logit(\pi) = 3.49 - 0.17 \times Age - 0.35 \times Status\_of\_Province + 0.08 \times Mother\_Education +$$
$$0.06 \times Father\_Occupation + 0.06 \times Family\_Income + 0.02 \times Father\_Education$$

5. Static Descriptional Aspects

For this binary logistic model the accuracy is equal 60.0%, whereas for the male group model it is 60.2%. It is slightly higher than the other group model. The mathematical logistic model for the male group of applicants is constructed as follows:

$$logit(\pi) = 2.61 - 0.13 \times Age + 0.09 \times Mother\_Education - 0.32 \times Status\_of\_Province +$$
$$0.07 \times Family\_Income + 0.06 \times Father\_Occupation + 0.04 \times Father\_Education$$

The above results for this year show that the father's education factor has a minimum effect on the acceptance to universities, whereas the age of applicants has a maximum effect on the target variable in all above logistic regression models.

Table 5.39: The Result for the Logistic Regression Model in 2007

Variable name	**Estimated Value**	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(B)**
Age	-0.153	0.001	25448.005	1	0.000	0.858
Status of Province	-0.344	0.005	5189.254	1	0.000	0.709
Mother's Education	0.090	0.003	748.980	1	0.000	1.094
Father's Occupation	0.061	0.003	384.546	1	0.000	1.063
Family Income	0.062	0.003	488.239	1	0.000	1.064
Father's Education	0.023	0.003	47.877	1	0.000	1.023
(Constant)	3.144	0.024	16927.448	1	0.000	23.197
Female Group						
Age	-0.166	0.001	17349.653	1	0.000	0.847
Status of Province	-0.353	0.006	3516.408	1	0.000	0.703
Mother's Education	0.085	0.004	423.916	1	0.000	1.089
Father's Occupation	0.059	0.004	223.184	1	0.000	1.069
Family Income	0.057	0.004	261.582	1	0.000	1.058
Father's Education	0.016	0.004	14.391	1	0.000	1.016
(Constant)	3.487	0.031	12457.940	1	0.000	32.704
Male Group						
Age	-0.134	0.001	8290.166	1	0.000	0.875
Mother's Education	0.095	0.005	312.594	1	0.000	1.099
Status of Province	-0.322	0.008	1595.032	1	0.000	0.725
Family Income	0.071	0.005	239.735	1	0.000	1.074
Father's Occupation	0.060	0.005	144.911	1	0.000	1.062
Father's Education	0.038	0.005	49.129	1	0.000	1.039
(Constant)	2.606	0.038	4628.787	1	0.000	13.541

Years 2008 and 2009

The results of the regression analysis are shown in Table 5.40 for the year 2008. It can be seen that two factors namely age and status of province, have a negative effect on the acceptance at universities respectively, whereas the other factors have positive effects. The ranking of the independent factors shows that the age of participants is the most important factor affecting the acceptance to universities, after that mother's education has the maximum effect and the father's education of candidate has the minimum effect in this model.

The logistic regression model for the year 2008 is constructed as follows:

$$logit(\pi) = 2.34 - 0.15 \times Age + 0.09 \times Mother\_Education - 0.16 \times Status\_of\_Province + 0.06 \times Father\_Occupation + 0.05 \times Family\_Income + 0.03 \times Father\_Education$$

The classification accuracy for this model is equal to 63.3%.

The mathematical logistic regression model for the female group applicants in year 2008 is as follows:

$$logit(\pi) = 2.76 - 0.17 \times Age + 0.07 \times Mother\_Education + 0.06 \times Father\_Occupation - 0.12 \times Status\_of\_Province + 0.05 \times Family\_Income + 0.03 \times Father\_Education$$

For this model the accuracy rate is 62.9%. For the male group candidates, the binary logistic regression model is constructed as follows:

$$logit(\pi) = 1.80 - 0.12 \times Age + 0.10 \times Mother\_Education - 0.23 \times Status\_of\_Province + 0.07 \times Family\_Income + 0.05 \times Father\_Occupation + 0.04 \times Father\_Education$$

The result for the year 2009 is presented in Table 5.41. A very similar conclusion as 2008 can be made for this year as well.

For this year, the logistic regression model is as follows:

$$logit(\pi) = 1.97 - 0.09 \times Age + 0.07 \times Mother\_Education - 0.24 \times Status\_of\_Province + 0.06 \times Father\_Occupation + 0.03 \times Family\_Income + 0.03 \times Father\_Education$$

Table 5.40: The Result for the Logistic Regression Model in 2008

Variable name	Estimated Value	S.E.	Wald	df	Sig.	Exp(B)
Age	-0.149	0.001	27696.930	1	0.000	0.862
Mother's Education.	0.086	0.003	702.176	1	0.000	1.090
Status of Province	-0.161	0.005	1127.486	1	0.000	0.851
Father's Occupation	0.060	0.003	361.496	1	0.000	1.062
Family Income	0.055	0.003	420.108	1	0.000	1.057
Father's Education	0.033	0.003	100.982	1	0.000	1.034
(Constant)	2.345	0.023	10337.828	1	0.000	10.434
Female Group						
Age	-0.171	0.001	20315.763	1	0.000	0.843
Mother's Education	0.072	0.004	305.796	1	0.000	1.075
Father's Occupation	0.064	0.004	248.958	1	0.000	1.066
Status of Province	-0.120	0.006	402.040	1	0.000	0.887
Family Income	0.049	0.003	203.997	1	0.000	1.050
Father's Education	0.026	0.004	37.272	1	0.000	1.026
(Constant)	2.756	0.030	8394.279	1	0.000	15.743
Male Group						
Age	-0.119	0.001	7905.169	1	0.000	0.888
Mother's Education	0.104	0.005	385.192	1	0.000	1.109
Status of Province	-0.229	0.008	815.867	1	0.000	0.795
Family Income	0.068	0.004	232.907	1	0.000	1.070
Father's Occupation	0.051	0.005	102.727	1	0.000	1.053
Father's Education	0.045	0.005	70.550	1	0.000	1.046
(Constant)	1.798	0.036	2479.279	1	0.000	6.036

The classification accuracy rate for the logistic regression model in this year is 56.8%.

Similar to the previous years, the logistic regression are investigated for male and female groups of WEE applicants. For the female group the mathematical logistic model is:

$$logit(\pi) = 2.25 - 0.11 \times Age - 0.23 \times Status\_of\_Province + 0.07 \times Mother\_Education + 0.06 \times Father\_Occupation + 0.03 \times Family\_Income + 0.02 \times Father\_Education$$

The accuracy rate of this model is equal 57.1%, whereas for the male group logistic models' accuracy is 56.8%. Moreover, the binary logistic regression model for the male group of applicants in year 2009 is as follows:

$$logit(\pi) = 1.58 - 0.08 \times Age + 0.07 \times Mother\_Education - 0.26 \times Status\_of\_Province + 0.05 \times Father\_Occupation + 0.04 \times Family\_Income + 0.04 \times Father\_Education$$

Table 5.41: The Result for the Logistic Regression Model in 2009

Variable name	Estimated Value	S.E.	Wald	df	Sig.	Exp(B)
Age	-0.094	0.001	14482.658	1	0.000	0.910
Mother's Education	0.070	0.003	483.829	1	0.000	1.073
Status of Province	-0.242	0.005	2537.300	1	0.000	0.785
Father's Occupation	0.056	0.003	315.516	1	0.000	1.058
Family Income	0.032	0.003	140.075	1	0.000	1.033
Father's Education	0.027	0.003	67.201	1	0.000	1.030
(Constant)	1.974	0.022	8388.903	1	0.000	7.201
Female Group						
Age	-0.107	0.001	10380.295	1	0.000	0.898
Status of Province	-0.231	0.006	1438.541	1	0.000	0.794
Mother's Education	0.071	0.004	294.956	1	0.000	1.073
Father's Occupation	0.059	0.004	202.231	1	0.000	1.060
Family Income	0.031	0.003	78.611	1	0.000	1.031
Father's Education	0.017	0.004	15.653	1	0.000	1.017
(Constant)	2.254	0.028	6398.355	1	0.000	9.524
Male Group						
Age	-0.077	0.001	4290.448	1	0.000	0.926
Mother's Education	0.067	0.005	169.949	1	0.000	1.069
Status of Province	-0.256	0.008	1065.659	1	0.000	0.774
Father's Occupation	0.050	0.005	103.179	1	0.000	1.052
Family Income	0.037	0.004	70.502	1	0.000	1.038
Father's Education	0.043	0.005	68.377	1	0.000	1.044
(Constant)	1.585	0.034	2213.675	1	0.000	4.880

The logistic regression analysis results for this year is approximately similar to the analysis of year 2007 and 2008.

Conclusion

Table 5.42 presents a collective view of the results for all five years from 2005 to 2009 along with *Beta* coefficients and t value of the linear regression model. This table shows a collective view of the results of the logistic regression analysis for these years. The *B/Beta* coefficients show that the effects of some factors such as age have increased during these years, whereas the influence of family income has decreased.

From the above results, it can be concluded that all factors which represent family background, individual, and environmental situations affect the mean of the total grade and thus affect the acceptance change to a university. As

Table 5.42: The Ordering of the Variables and Coefficients of the Linear and Logistic Regression Models during 2005 to 2009

Variable name	**Standardized Linear Regression for Five Years 2005 to 2009**												
	2005		**2006**		**2007**		**2008**		**2009**		**Total 2005-9**		**Sig.**
	Beta	**t**	**Beta**	**t**	**Beta**	**t**	**Beta**	**t**	**Beta**	**t**	**Beta**	**t**	
Age	-0.127	-128.850	-0.149	-153.879	-0.272	-268.825	-0.283	-300.623	-0.253	-260.422	-0.212	-488.498	0.000
Status of Province	-0.062	-61.932	-0.063	-65.562	-0.074	-76.372	-0.070	-74.265	-0.073	-73.794	-0.057	-130.862	0.000
Father's Education	0.057	39.232	0.069	54.254	0.037	25.843	0.038	27.055	0.038	26.194	0.060	98.027	0.000
Mother's Education	-0.001	-0.928	0.040	31.412	0.034	25.371	0.032	24.943	0 .035	25.991	0.052	87.566	0.000
Father's Occupation	0.002	1.366	0.067	70.063	0.037	31.935	0.024	21.229	0.018	14.858	0.029	62.679	0.000
Mother's Occupation	0.072	67.174									0.034	78.051	0.000
Family Income	0.049	45.153	0.075	78.737	0.060	56.625	0.064	60.863	0 .073	63.102	0.027	59.112	0.000
No. of Family Members	0.017	16.033									-0.054	-122.934	0.000
Variable name	**Standardized Linear Regression for Five Years 2005 to 2009**												
	2005		**2006**		**2007**		**2008**		**2009**		**Total 2005-9**		**Sig.**
	B	**Wald**	***B***	**Wald**	***B***	**Wald**	***B***	**Wald**	***B***	**Wald**	***B***	**Wald**	
Age	-0.118	8568.895	-0.101	9911.933	-0.153	25448.005	-0.149	27696.930	-0.094	14482.658	-0.116	7644.3896	0.000
Status of Province	-0.224	1699.445	-0.220	2226.106	-0.344	5189.254	-0.161	1127.486	-0.242	2537.300	-0.187	7732.698	0.000
Father's Education	0.115	950.414	0.084	1220.338	0.023	48.887	0.033	100.982	0.027	67.201	0.065	2463.860	0.000
Mother's Education	0.038	92.413	0.081	815.507	0.090	748.980	0.086	702.176	0.070	483.829	0.096	4779.740	0.000
Father's Occupation	0.010	10.727	0.087	1862.112	0.061	384.546	0.060	361.496	0.056	315.516	0.067	3312.211	0.000
Mother's Occupation	0.120	1438.513									0.175	34771.585	0.000
Family Income	0.096	1405.334	0.114	3085.461	0.062	488.239	0.055	420.108	0.032	140.075	0.013	157.956	0.000
No. of Family Members	-0.015	41.362									-0.100	12004.435	0.000
(Constant)	0.964	978.514	1.127	2070.554	3.144	16927.448	2.345	10337.828	1.974	8388.903	1.719	26487.221	0.000

a result, the educational performance of WEE participants are indeed influenced by these factors. Moreover, the results of this analysis are in accordance with the previous analysis. As we already mentioned, in each year five/six factors namely *father* and *mother education, father* and *mother occupation* and *family income* have a positive effect, whereas three factors *age of candidates, the number of family members* and *status of province*[11] negatively affect the acceptance at universities.

5.3. Data Mining Techniques

As mentioned in this study before, we have two factors as target variables, *the total grade* and *acceptance at university*. In order to find a classification model we evaluate five algorithms namely Artificial Neural Networks (ANN), Classification and Regression Tree (CART), Chi-square Automatic Detection (CHAID), Quick Unbiased Efficient Statistical Tree (QUEST), and Classification algorithm C5.0. Whereas, in order to a prediction models are investigated by three first algorithms; ANN, CART, and CHAID. In this section, we present the results of these evaluations followed by a comparison between the results of various models.

5.3.1. Classification Models

In this section we will consider individual and family background as well as environmental factors as independent and acceptance at university as a target variable. Note that the acceptance at university is a binary attribute with the values *accepted* or *not accepted*. Consequently, we used the classification method in order to classify the applicants in two groups based on these values. In the following, the results of various algorithms for the five year datasets are presented.

5.3.1.1. Artificial Neural Networks (ANN)

The neural networks algorithms are used for two main purposes: forecasting and classification. In other words, ANN can be used not only for classification but also for forecasting.

We now present the results of classification methods for our datasets. We take as the target variable the acceptance of candidates at university, whereas individual factors namely *age* and *gender*, family background factors such as *parental education* and *occupation*, *family income*, and *status province* as environmental factors are taken as independent attributes. In the following, we

[11] As already mention the status of province are coded in decreasing levels.

first show the results of the ANN algorithm for years 2005 to 2009 collectively. Later, we describe the results for each individually.

Figure 5.18(2005-9) shows the result of the neural networks model on the total data from 2005 to 2009. It shows that the parental education of WEE applicants has a positive effect on the acceptance at university behind the age of applicants. The analysis of this model shows that the classification is a 64.43% correct classification.

The results of the classification model for the years 2005 to 2009 individually are shown in Figure 5.18 and Table 5.43.

Year 2005

For this year, the independent variables we considered are family background factors (father's and mother's education, father's and mother's occupation, family income, and the number of family members), individual factors (age and gender) and an environmental factor (the status of province of applicants).

In order to find the importance of independent factors in the classification model, Figure 5.18(2005) and the first column of Table 5.43 show the variables based on the final classification model, which are age, mother's occupation, father's education, mother's education, status of province, the number of family members, family income, father's occupation, and gender. The variables are presented in order of decreasing importance. The analysis of this model shows an accuracy rate of about 75.92%.

Year 2006

As we noted before, we have only four family background factors for the years 2006-2009. Unlike in the year 2005, we do not have the number of family members and mother's occupation for our analysis.

Figure 5.18(2006) and the second column of Table 5.43 present the independent variables in order of decreasing importance based on the final model. These include the age of participants, mother's education, status of province, family income, father's occupation, father's education, and gender. For this analysis, we observed an accuracy rate of 64.4% for the model.

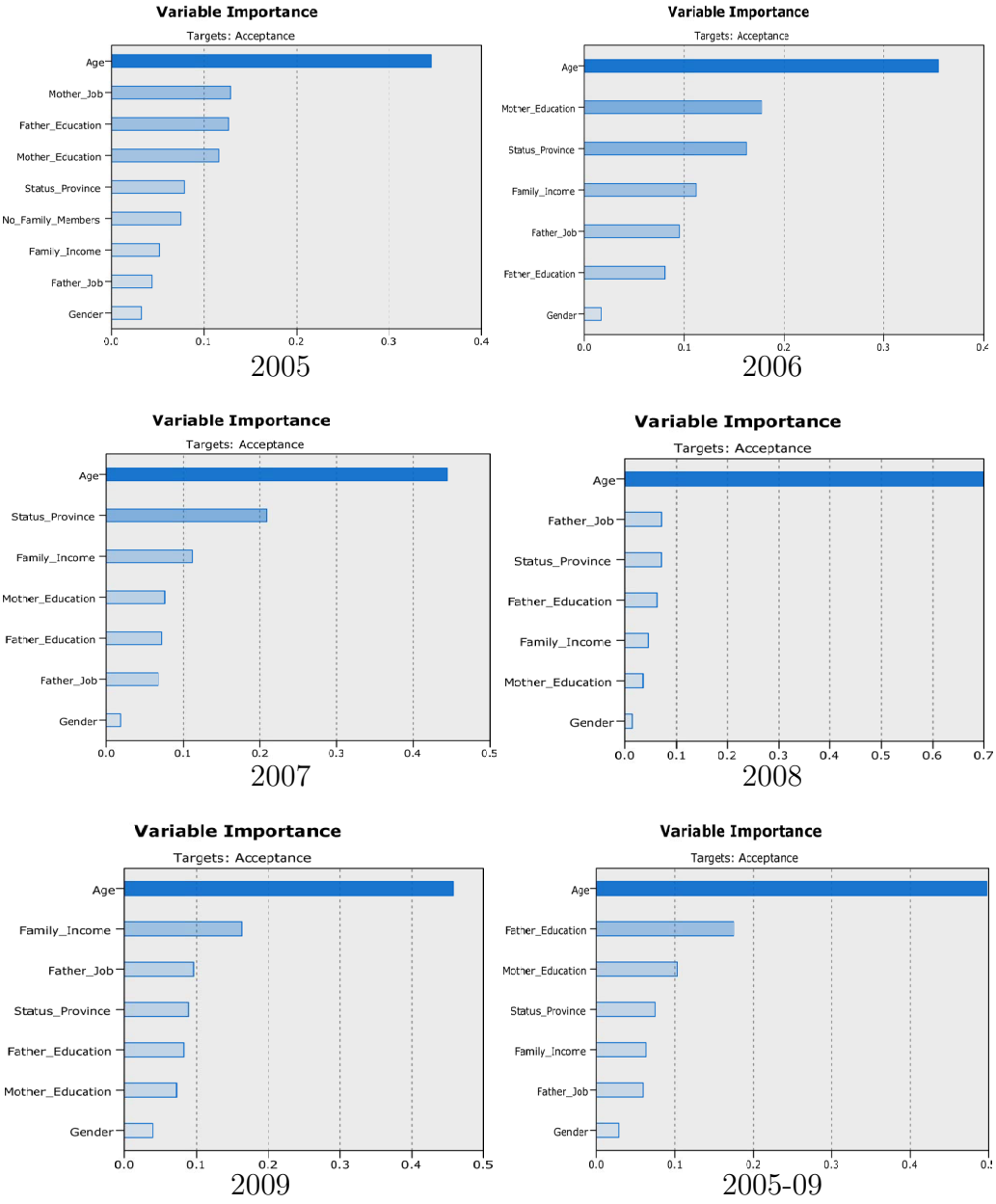

Figure 5.18: The Neural Networks Results for Classification Models in 2005-2009

Year 2007

From Figure 5.18(2007) and in the third column of Table 5.43, it can be seen that the classification model shows several factors in increasing order of importance. Note that the ordering of factors now is the age of participants, status of province, family's income, mother's education, father's education, father's occupation, and gender. The accuracy rate of this analysis of the classification model is 60.75%.

Table 5.43: The Results of Neural Networks Model for Classification in 2005-2009

Year					
2005		**2006**		**2007**	
Variable	**Imp. R.***	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**
Age	0.346	Age	0.355	Age	0.444
Mother Occupation	0.128	Mother Education	0.178	Status Province	0.208
Father Education	0.126	Status Province	0.162	Family Income	0.112
Mother Education	0.116	Family Income	0.112	Mother Education	0.076
Status Province	0.079	Father Occupation	0.095	Father Education	0.072
No. of Family	0.075	Father Education	0.081	Father Occupation	0.068
Family Income	0.052	Gender	0.017	Gender	0.019
Father Occupation	0.044	-	-	-	-
Gender	0.032	-	-	-	-
Accuracy Rate	75.92	Accuracy Rate	64.4	Accuracy Rate	60.75

2008		**2009**		**2005-09**	
Variable	**Imp. R.**	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**
Age	0.7	Age	0.459	Age	0.498
Father Occupation	0.071	Family Income	0.163	Father Education	0.175
Status Province	0.071	Father Occupation	0.096	Mother Education	0.103
Father Education	0.063	Status Province	0.089	Status Province	0.074
Family Income	0.046	Father Education	0.082	Family Income	0.063
Mother Education	0.063	Mother Education	0.072	Father Occupation	0.058
Gender	0.014	Gender	0.039	Gender	0.029
Accuracy Rate	63.5	Accuracy Rate	56.59	Accuracy Rate	64.43

*__Imp. R.__ denoted as importance rate.

Year 2008

In Figure 5.18(2008) and in the fourth column of Table 5.43 the results of the neural networks model are presented for this year. It shows the order of importance of the independent variables based on the final classification model on the target variable. The observed order of importance is: age of participant, father's occupation, status of province, father's education, family's income, mother's education, and gender. The model for this year has 63.5% accuracy.

Year 2009

For the year 2009, Figure 5.18(2009) and the fifth column of Table 5.43 present the results for the importance order of variables based on the final classification model. The variables are age of applicant, family's income, father's occupation,

status of province, father's education, mother's education, and gender. The accuracy rate of the model is 56.59%.

Conclusion

The above results show that in all five years (2005-2009), all the independent variables (individual, environmental, and family background factors) have important effects on the classification model. Furthermore, the constructed model has an accuracy of more than 60% for the classification of the target variable as an educational outcome.

5.3.1.2. Classification and Regression Tree (CART)

We first present the results of this algorithm for the dataset (2005 to 2009), and later describe the results for each year. Again note that the number of family background factors is different for each year. In year 2005 unlike the other years, we have two additional factors namely *mother's occupation* and *family size*.

Figure 5.19(2005-9) shows that only five factors have an effect on the classification model. In other words, it shows that father's occupation and gender of applicants are not important in this model. The accuracy rate of the model for a correct classification is 63.43%. Figure 5.19 and Table 5.44 present a comparison of results of CART models based on the total dataset and each of five years data from 2005 to 2009. These results show that age of applicants is important for the classification model in all years. Further, after province, mother's education, and father's education are important factors for the acceptance at university. The accuracy rate of the models shows that the model for the year 2005 with 75.9% accuracy/correct classification has a higher rate than the other models in this analysis.

5.3.1.3. Chi-square Automatic Detection (CHAID)

In this section, we consider CHAID algorithm for identification of a classification model for our study. The results of this algorithm are based on the total dataset 2005-2009 and each year of the dataset from 2005 to 2009.

Figure 5.20(2005-9) presents the CHAID model for the total dataset (2005 to 2009). This model shows that all independent variables have effects on

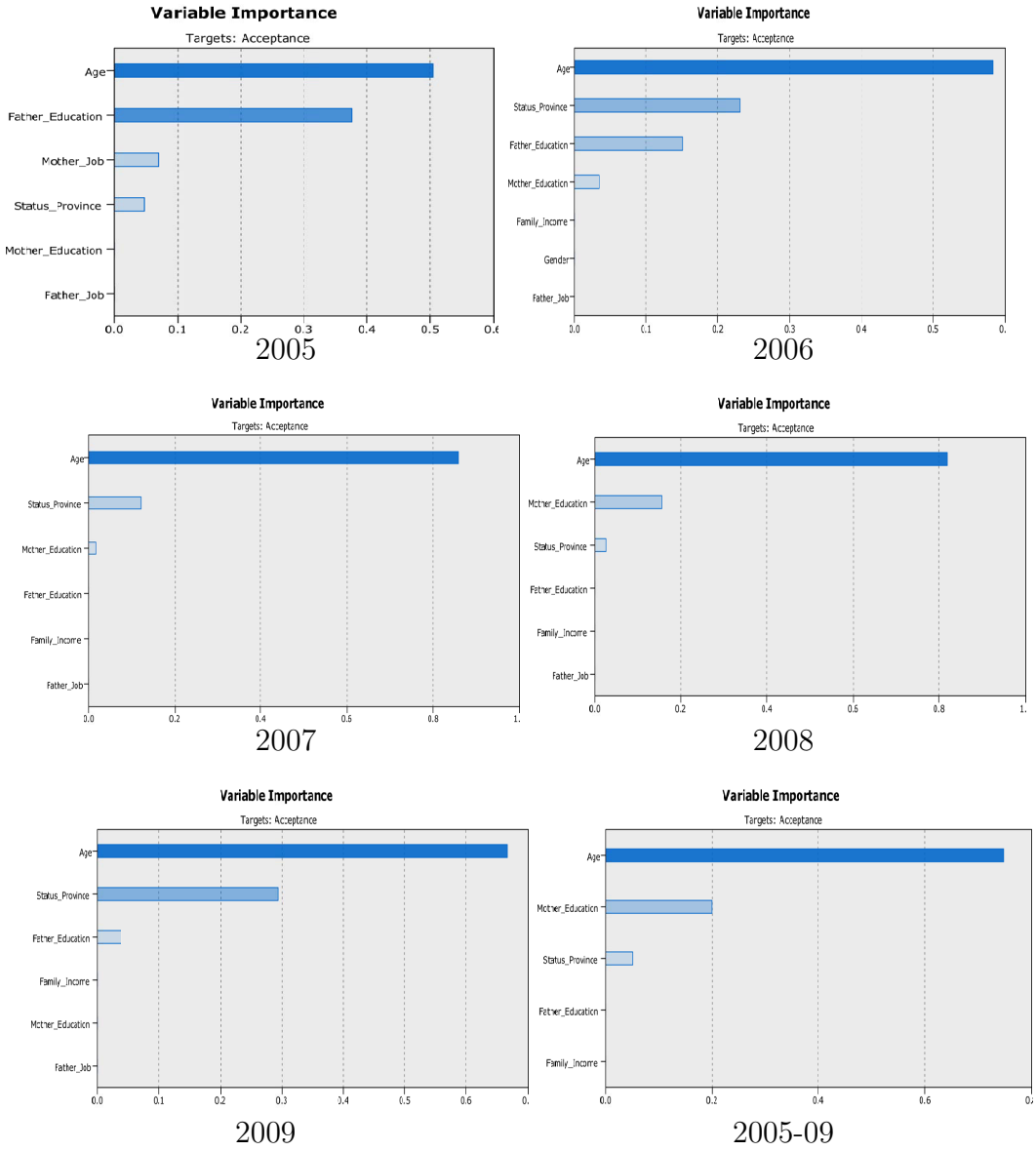

Figure 5.19: The Classification and Regression Tree Results for Classification Models in 2005-2009

acceptance at university. The rate of accuracy of the model is 63.22% for the correct classification.

Table 5.45 and Figure 5.20 show the result of the CHAID algorithm based on the five years dataset from 2005 to 2009. These results present the order of importance of variables in the classification model. It is clear that the age of WEE applicants has a large effect on acceptance at university. Additionally, parental education and status of province have effects on the classification model for acceptance at university as an educational outcome.

Table 5.44: The Classification and Regression Tree Results for Classification Models in 2005-2009

Year					
2005		**2006**		**2007**	
Variable	**Imp. R.**	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**
Age	5.04E-1	Age	5.84E-1	Age	8.61E-1
Father Education	3.77E-1	Status Province	2.31E-1	Status Province	10.22E-1
Mother Occupation	7.04E-2	Father Education	1.51E-1	Mother Education	1.73E-2
Status Province	4.83E-2	Mother Education	0.34E-1	Father Education	3.43E-5
Mother Education	1.63E-4	Family Income	0.0	Family Income	3.43E-5
Father Occupation	1.63E-4	Gender	0.0	Father Occupation	3.43E-5
-	-	Father Occupation	0.0	-	-
Accuracy Rate	75.9	Accuracy Rate	64.71	Accuracy Rate	64.1

2008		**2009**		**2005-09**	
Variable	**Imp. R.**	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**
Age	8.19E-1	Age	6.66E-1	Age	7.49E-1
Mother Education	1.55E-1	Status Province	2.95E-1	Mother Education	1.99E-1
Status Province	2.57E-2	Father Education	3.89E-2	Status Province	5.18E-2
Father Education	7.44E-5	Family Income	1.07E-4	Father Education	2.18E-5
Family Income	7.44E-5	Mother Education	1.07E-4	Family Income	2.18E-5
Father Occupation	7.44E-5	Father Occupation	1.07E-4	-	-
Accuracy Rate	64.01	Accuracy Rate	59.56	Accuracy Rate	63.43

The accuracy rates of models shows that the year 2005 is the best model with 75.92% correct classification rate, whereas the year 2009 with 58.5% rate has the lowest accuracy.

5.3.1.4. Classification Algorithm (C5.0)

We now discuss the application of C5.0 algorithm for finding a classification model.

Figure 5.21 and Table 5.46 present the results of the C5.0 classification algorithm based on the five years datasets 2005-2009. It can be seen that in year 2005 the status of province is of utmost importance. The results of this algorithm show that all independent factors have an effect on the acceptance at university as a target variable.

Accuracy rates of this analysis on the five years datasets are 61.09% to 76.07% for the year 2005 to 2009. It shows that the year 2005 model has a maximum accuracy rate, whereas the rate of year 2009 model has a minimum

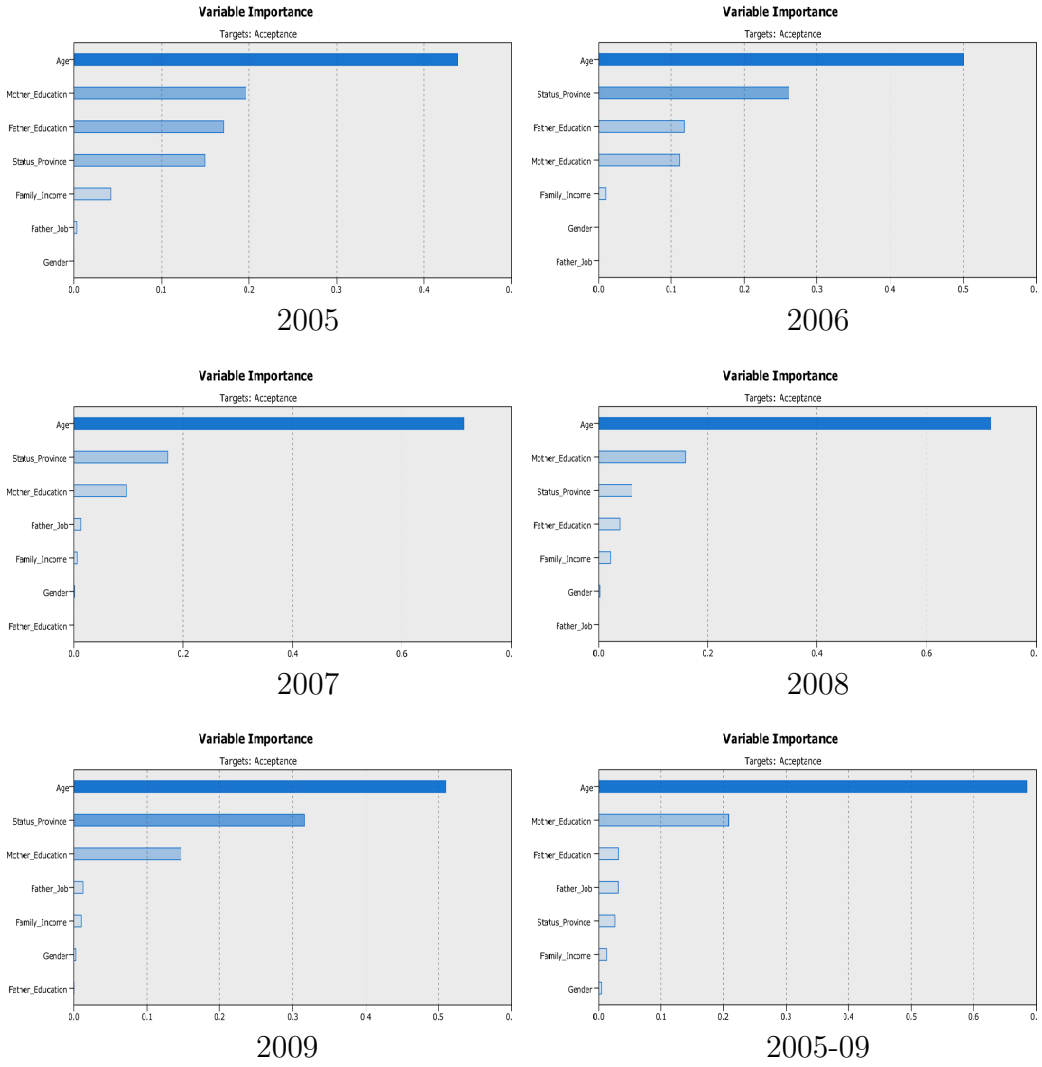

Figure 5.20: The Chi-square Automatic Detection Results for Classification Models in 2005-2009

accuracy. Figure 5.21(2005-9) shows a C5.0 model based on the total datasets 2005-2009 with ordering of importance variables.

5.3.1.5. Quick Unbiased Efficient Statistical Tree (QUEST)

The QUEST algorithm is used for classification purposes in this section. Again, the acceptance at university is considered as a target variable whereas family background, individual and environmental factors are independent variables. As already mentioned, in the year 2005 dtataset we analyzed nine independent

Table 5.45: The Chi-square Automatic Detection Results for Classification Models in 2005-2009

Year					
2005		**2006**		**2007**	
Variable	**Imp. R.**	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**
Age	4.38E-1	Age	5.0E-1	Age	7.12E-1
Mother Education	1.96E-1	Status Province	2.61E-1	Status Province	1.71E-1
Father Education	1.71E-1	Father Education	1.17E-1	Mother Education	9.64E-2
Status Province	1.94E-1	Mother Education	1.12E-1	Father Occupation	1.33E-2
Family Income	4.21E-2	Family Income	0.10E-1	Family Income	5.56E-3
Father Occupation	4.13E-3	Gender	0.0	Gender	1.15E-3
Gender	1.23E-4	Father Occupation	0.0	Father Education	1.71E-5
Accuracy Rate	75.92	Accuracy Rate	64.71	Accuracy Rate	61.42
2008		**2009**		**2005-09**	
Variable	**Imp. R.**	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**
Age	7.18E-1	Age	5.10E-1	Age	6.85E-1
Mother Education	1.58E-1	Status Province	3.17E-1	Mother Education	2.08E-1
Status Province	0.61E-1	Mother Education	1.47E-1	Father Education	0.32E-1
Father Education	0.39E-1	Father Occupation	1.29E-2	Father Occupation	0.31E-1
Family Income	0.21E-1	Family Income	1.06E-2	Status Province	0.27E-1
Gender	0.02E-1	Gender	2.41E-3	Family Income	0.13E-1
Father Occupation	0.0	Father Education	2.91E-4	Gender	0.04E-1
Accuracy Rate	64.37	Accuracy Rate	58.5	Accuracy Rate	63.22

factors whereas in the other year datasets we have seven factors. In this subsection we will present the results of the QUEST algorithm based on the complete five years datasets and later for each year dataset separately.

Similar to the previous algorithms, Figure 5.22(2005-9) presents the ordering of variable importance in the classification model based on the total datasets 2005-2009. It shows that age of applicants, status of province, and parental education are important for the acceptance at university as an educational outcome.

The results of the QUEST algorithm are presented in Figure 5.22 and Table 5.47. It can be seen that the age of applicants is the most important factor in all five years' models. The second and third important factors are different in five years. For instance in 2005, the ordering of importance factors are father's education, family income, and mother's education, whereas in year 2009, status

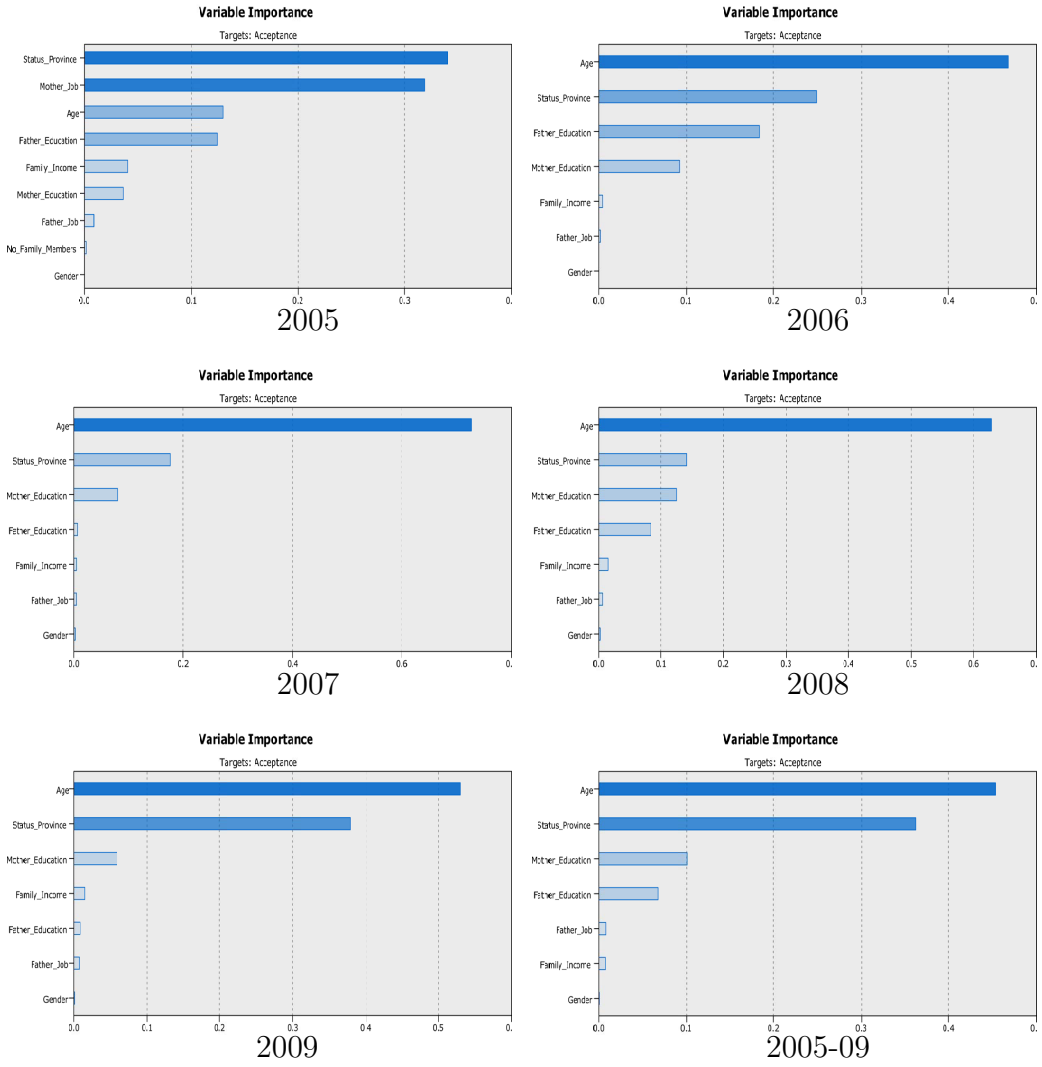

Figure 5.21: The Classification Algorithm (C5.0) Results for Classification Models in 2005-2009

of province, mother's education, and father's education compromises the order of importance.

Conclusion

The above mentioned results show that all independent factors have effects on the acceptance at university as a target variable. This supports the theoretical model mentioned in Chapter 1.

5.3.2. Prediction Models

In order to investigate the prediction purpose, we used three different algorithms as prediction models. The total grade is a continuous variable as the

Table 5.46: The Classification Algorithm (C5.0) Results for Classification Models in 2005-2009

Year					
2005		**2006**		**2007**	
Variable	**Imp. R.**	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**
Status Province	3.40E-1	Age	4.70E-1	Age	7.26E-1
Mother Occupation	3.19E-1	Status Province	2.60E-1	Status Province	1.76E-1
Age	1.30E-1	Father Education	1.74E-1	Mother Education	0.80E-1
Father Education	1.24E-1	Mother Education	0.92E-1	Father Education	0.06E-1
Family Income	0.41E-1	Family Income	0.03E-1	Family Income	0.05E-1
Mother Education	0.36E-1	Father Occupation	0.01E-1	Father Occupation	0.04E-1
Father Occupation	0.09E-1	Gender	0.0	Gender	0.02E-1
No. of Family	0.02E-1	-	-	-	-
Gender	0.0	-	-	-	-
Accuracy Rate	76.07	Accuracy Rate	65.53	Accuracy Rate	62.87
2008		**2009**		**2005-09**	
Variable	**Imp. R.**	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**
Age	6.30E-1	Age	5.30E-1	Age	4.50E-1
Status Province	1.40E-1	Status Province	3.80E-1	Status Province	3.62E-1
Mother Education	1.25E-1	Mother Education	5.88E-2	Mother Education	1.0E-1
Father Education	0.83E-1	Family Income	1.49E-2	Father Education	6.75E-2
Family Income	0.15E-1	Father Education	8.90E-3	Father Occupation	8.35E-3
Father Occupation	0.06E-1	Father Occupation	7.05E-3	Family Income	7.85E-3
Gender	0.07E-1	Gender	9.60E-4	Gender	4.75E-4
Accuracy Rate	65.79	Accuracy Rate	61.09	Accuracy Rate	65.87

educational outcome of the WEE applicants is predicted by the prediction method. Similar to the previous section, the independent attributes are divided into three categories which are individual (age, gender), environmental (status of province), and family background factors (parental education, parental occupation, family income, number of family member).

According to the year of the dataset, the number of independent factors is different. For instance, we analyzed six family background factors for the year 2005, while for the other years we used only four factors. In the following we will present the results of the prediction models based on five years datasets and the total datasets from 2005 to 2009.

5.3.2.1. Artificial Neural Networks (ANN)

As we noted the neural networks analysis is used for prediction purposes as well as classification analysis. In this subsection, we present the results of

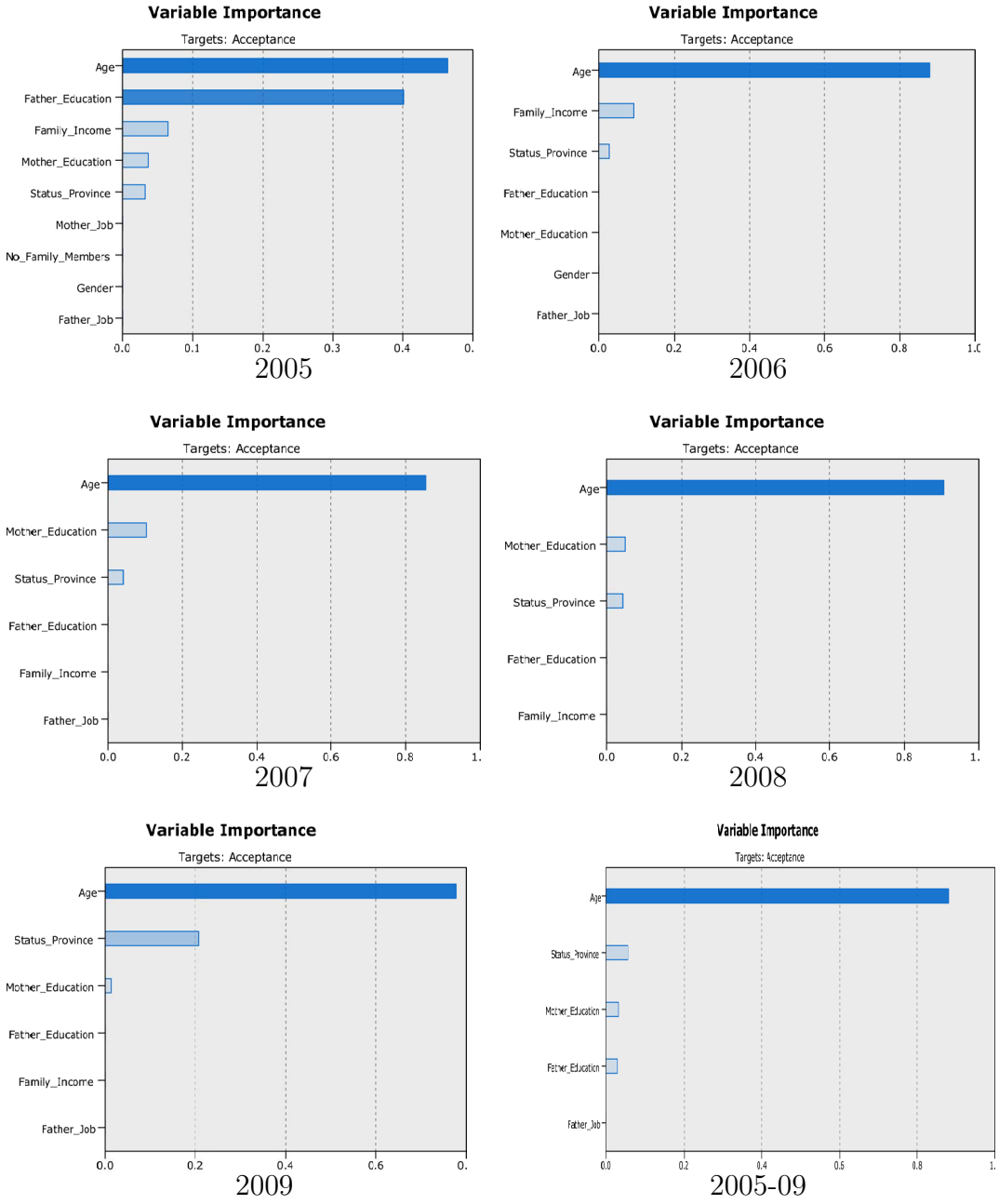

Figure 5.22: The Quick Unbiased Efficient Statistical Tree Results for Classification Models in 2005-2009

the prediction analysis of the total grades of WEE applicants as a continuous target variable. In the following we will show the results of the prediction model over all five years datasets (i.e. 2005-2009) and later for each year's dataset separately.

The results of the neural networks prediction model based on the total datasets from 2005 to 2009 are presented in Figure 5.23(2005-9). This figure shows that age of applicants, status of province, parental education have

Table 5.47: The Results of Quick Unbiased Efficient Statistical Tree Model for Classification on 2005-2009

Year					
2005		**2006**		**2007**	
Variable	**Imp. R.**	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**
Age	4.64E-1	Age	8.78E-1	Age	8.55E-1
Father Education	4.01E-1	Family Income	0.92E-1	Mother Education	1.04E-1
Family Income	6.41E-2	Status Province	0.29E-1	Status Province	4.09E-2
Mother Education	3.68E-2	Father Education	0.0	Father Education	6.68E-5
Status Province	3.27E-2	Mother Education	0.0	Family Income	6.68E-5
Mother Occupation	1.95E-4	Gender	0.0	Father Occupation	6.68E-5
No. of Family	1.95E-4	Father Occupation	0.0	-	-
Gender	1.95E-4	-	-	-	-
Father Occupation	1.95E-4	-	-	-	-
Accuracy Rate	76.02	Accuracy Rate	64.35	Accuracy Rate	61.28
2008		**2009**		**2005-09**	
Variable	**Imp. R.**	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**
Age	9.07E-1	Age	7.79E-1	Age	8.82E-1
Mother Education	4.92E-2	Status Province	2.08E-1	Status Province	5.69E-2
Status Province	4.34E-2	Mother Education	1.19E-2	Mother Education	3.18E-2
Father Education	7.69E-5	Father Education	1.17E-4	Father Education	2.92E-2
Family Income	7.69E-5	Family Income	1.17E-4	Father Occupation	2.65E-5
-	-	Father Occupation	1.17E-4	-	-
Accuracy Rate	63.77	Accuracy Rate	58.55	Accuracy Rate	62.82

effects on the prediction model. The accuracy for this model is 0.338 as a linear correlation for prediction of the total grade of applicants.

The neural networks results of prediction based on the five years datasets are represented in Figure 5.23 and Table 5.48. Similar to the previous algorithms in classification models, these results show that the age of applicants is the most important factor in all prediction models. The order of the other independent factors in these models for the five years is different. For example, in the year 2005 status of province, father's education, mother's occupation, family income, mother's education, father's occupation, and gender is the order of importance of attributes with a minimum accuracy in the five years models.

By contrast, in the year 2008 with maximum accuracy, father's education, status of province, mother's education, gender, family income, and father's occupation are the decreasing order of independent factors.

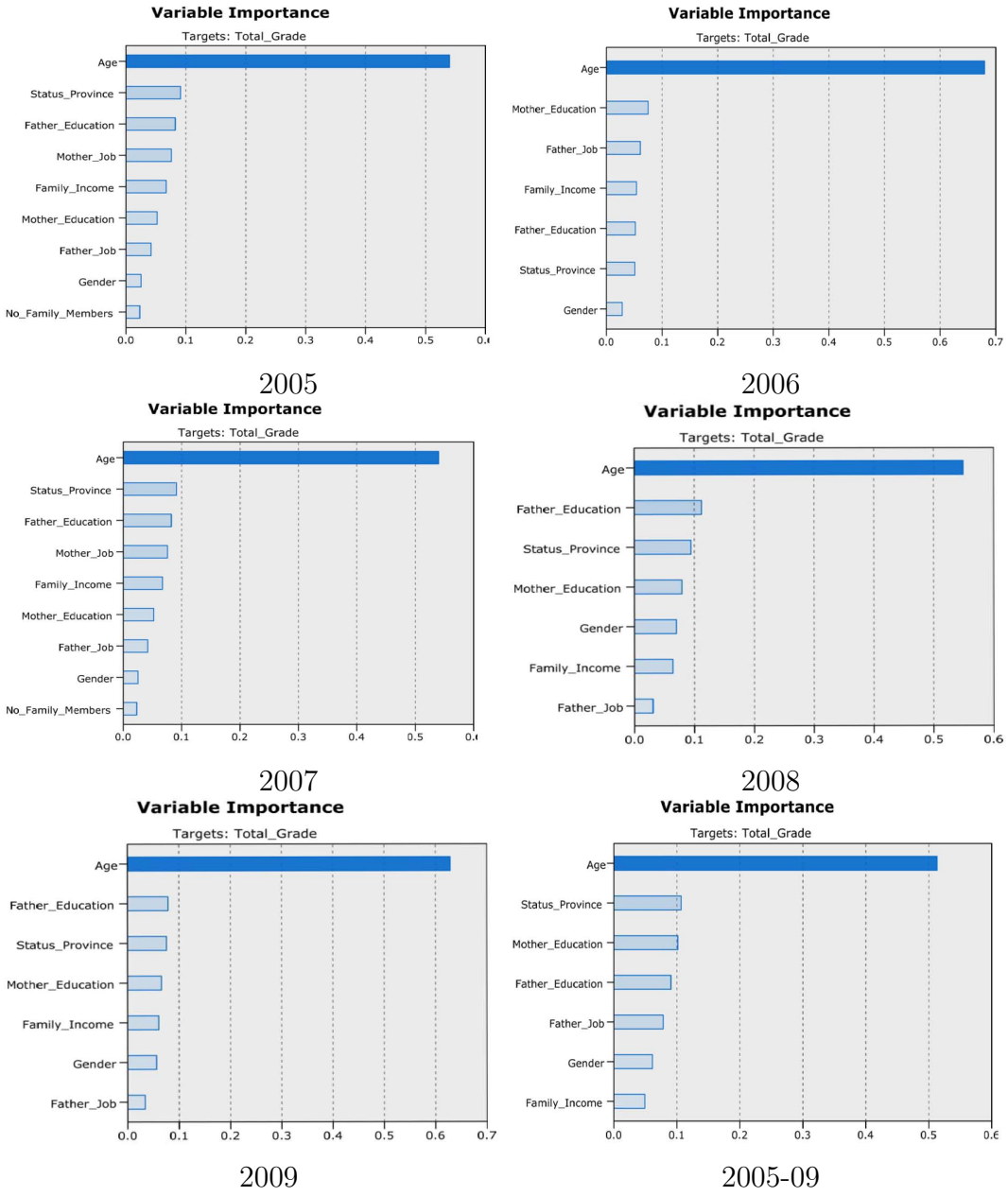

Figure 5.23: The Neural Networks Results for Prediction Models in 2005-2009

Conclusion

From the above given results for the prediction analysis, it can be seen that the target variable is predicted by the all independent variables in this study which are family background, individual, and environmental factors as an indicator of socioeconomic status of WEE candidates. In other words, the results show that the socioeconomic status affects the total grades and consequently the educational outcome of applicants.

Table 5.48: The Neural Networks Results for Prediction Models in 2005-2009

Year					
2005		**2006**		**2007**	
Variable	**Imp. R.**	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**
Age	0.54	Age	0.68	Age	0.618
Status Province	0.091	Mother Education	0.075	Status Province	0.08
Father Education	0.082	Father Occupation	0.061	Mother Education	0.76
Mother Occupation	0.076	Family Income	0.054	Father Occupation	0.064
Family Income	0.067	Father Education	0.052	Father Education	0.062
Mother Education	0.052	Status Province	0.051	Gender	0.057
Father Occupation	0.042	Gender	0.028	Family Income	0.043
Gender	0.025	-	-	-	-
Linear Correlation	0.253	Linear Correlation	0.272	Linear Correlation	0.408
2008		**2009**		**2005-09**	
Variable	**Imp. R.**	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**
Age	0.549	Age	0.629	Age	0.513
Father Education	0.112	Father Education	0.079	Status Province	0.107
Status Province	0.094	Status Province	0.76	Mother Education	0.101
Mother Education	0.079	Mother Education	0.66	Father Education	0.09
Gender	0.07	Family Income	0.61	Father Occupation	0.078
Family Income	0.064	Gender	0.056	Gender	0.061
Father Occupation	0.031	Father Occupation	0.34	Family Income	0.049
Linear Correlation	0.416	Linear Correlation	0.377	Linear Correlation	0.338

5.3.2.2. Classification and Regression Tree (CART)

Like the neural network algorithm, the classification and regression tree algorithm could be used for both prediction and classification. We used the CART algorithm for continues/range target variable. The results of this algorithm based on the total dataset 2005-2009 and on each of the five years datasets are presented in Figure 5.24 and Table 5.49.

These results show that the model of the year 2005 has minimum accuracy whereas the year 2008 model has a maximum linear correlation. The order of important factors in 2005 is number of family members, father's occupation, status of province, father's education, age, family income, mother's education, and gender. By contrast, the order of independent factors in year 2008 for the prediction model is age, gender, father's occupation, status of province, father's education, family income, and mother's education.

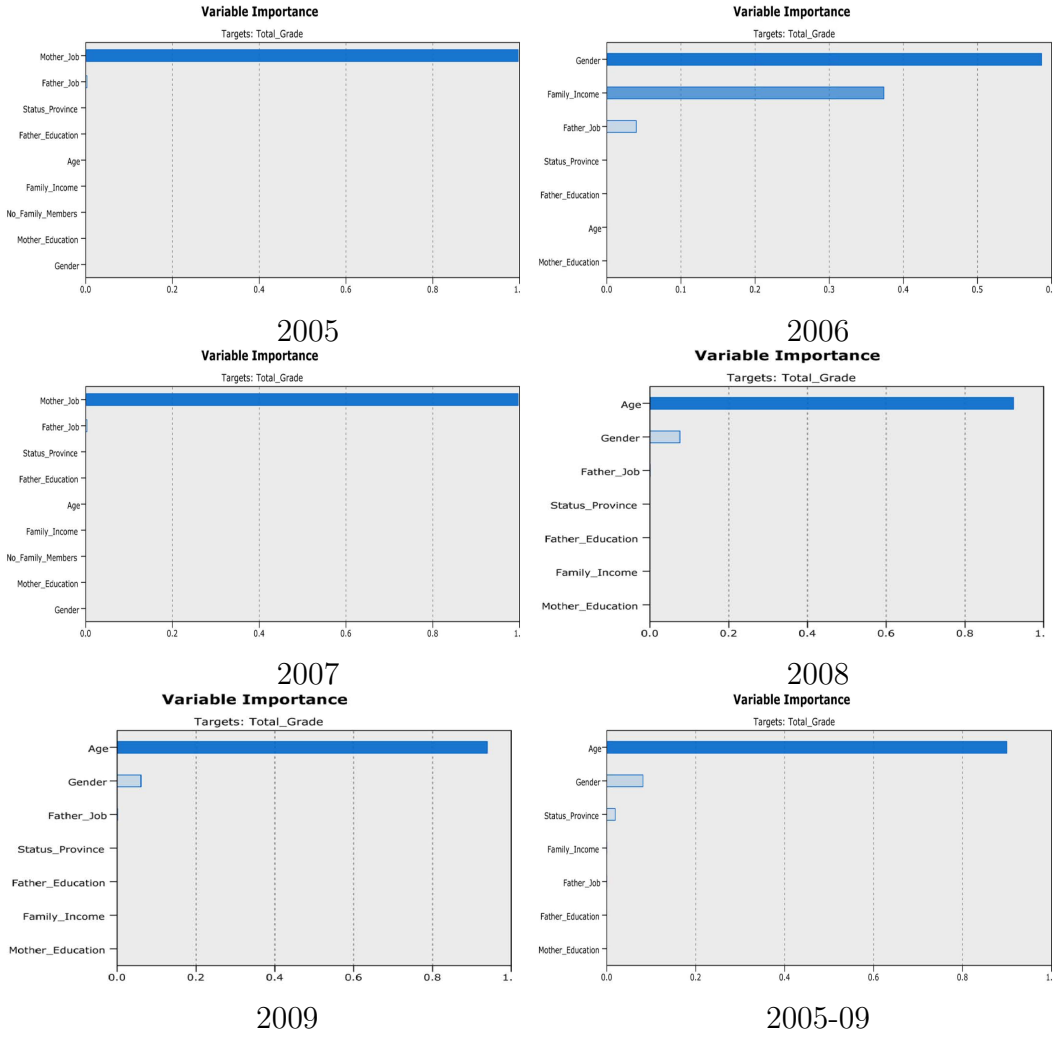

Figure 5.24: The Classification and Regression Tree Results for Prediction Models in 2005-2009

Figure 5.24(2005-9) shows the order of importance variables by CART algorithm based on the total datasets 2005 to 2009. It shows that the age of the applicant is the most important factor, after that gender, status of province, family income, father's occupation, father's and mother's education are then the important factors in decreasing order in this prediction model.

5.3.2.3. Chi-square Automatic Detection (CHAID)

We use the CHAID algorithm for predicting our continuous target variable which is the total grades of WEE applicants. We first present the results for all five years which is followed by the result of each individual year (c. f. Figure 5.25(2005-9)).

Table 5.49: The Classification and Regression Tree Results for Prediction Models in 2005-2009

Year					
2005		**2006**		**2007**	
Variable	**Imp. R.**	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**
No. of Family	0.5	Gender	0.587	Age	0.96
Father Occupation	0.5	Family Income	0.374	Gender	0.038
Status Province	0.0	Father Occupation	0.04	Family Income	0.001
Father Education	0.0	Status Province	0.0	Father Occupation	0.001
Age	0.0	Father Education	0.0	Status Province	0.0
Family Income	0.0	Age	0.0	Father Education	0.0
Mother Education	0.0	Mother Education	0.0	Mother Education	0.0
Gender	0.0	-	-	-	-
Linear Correlation	0.241	Linear Correlation	0.263	Linear Correlation	0.404
2008		**2009**		**2005-09**	
Variable	**Imp. R.**	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**
Age	0.923	Age	0.939	Age	0.899
Gender	0.076	Gender	0.06	Gender	0.81
Father Occupation	0.001	Father Occupation	0.001	Status Province	0.019
Status Province	0.0	Status Province	0.0	Family Income	0.0
Father Education	0.0	Father Education	0.0	Father Occupation	0.0
Family Income	0.0	Family Income	0.0	Father Education	0.0
Mother Education	0.0	Mother Education	0.0	Mother Education	0.0
Linear Correlation	0.407	Linear Correlation	0.37	Linear Correlation	0.326

As can be seen from Figure 5.25(2005-9), gender, unlike in the other models, is of utmost importance followed by status of province, father's education , age, family income, mother's education and father's occupation. The accuracy for this prediction model is 0.33 linear correlation between target variable and predicted variable.

Figure 5.25 and Table 5.50 show the results of the CHAID algorithm based on five year datasets (i. e. 2005 to 2009). These results show that similar to the previous prediction algorithms, the model of the year 2005 has a minimum accuracy whereas the model of the year 2008 has a maximum linear correlation as accuracy for the continuous target variable.

Table 5.50: The Results of CHAID Models for Prediction in 2005-2009

Year					
2005		**2006**		**2007**	
Variable	**Imp. R.**	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**
Age	1.0	Status Province	0.551	Age	1
Mother Education	0.0	Gender	0.315	Status Province	0.0
Father Education	0.0	Family Income	0.134	Father Education	0.0
Status Province	0.0	Father Education	0.0	Family Income	0.0
Family Income	0.0	Age	0.0	Mother Education	0.0
Father Occupation	0.0	Mother Education	0.0	Father Occupation	0.0
Gender	0.0	Father Occupation	0.0	Gender	0.0
Linear Correlation	0.247	Linear Correlation	0.256	Linear Correlation	0.401
2008		**2009**		**2005-09**	
Variable	**Imp. R.**	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**
Age	0.936	Age	0.956	Gender	1
Gender	0.064	Gender	0.044	Status Province	0.0
Status Province	0.0	Status Province	0.0	Father Education	0.0
Father Education	0.0	Father Education	0.0	Age	0.0
Family Income	0.0	Family Income	0.0	Family Income	0.0
Mother Education	0.0	Mother Education	0.0	Mother Education	0.0
.	.	Father Occupation	0.0	Father Occupation	0.0
Linear Correlation	0.408	Linear Correlation	0.364	Linear Correlation	0.33

Conclusion

From the results presented in the previous sections we can simply conclude that all the independent factors have effects on the total grade which we consider as the target variable in this work. Hence, it is totally supported by the theoretical model we mentioned in Chapter 1.

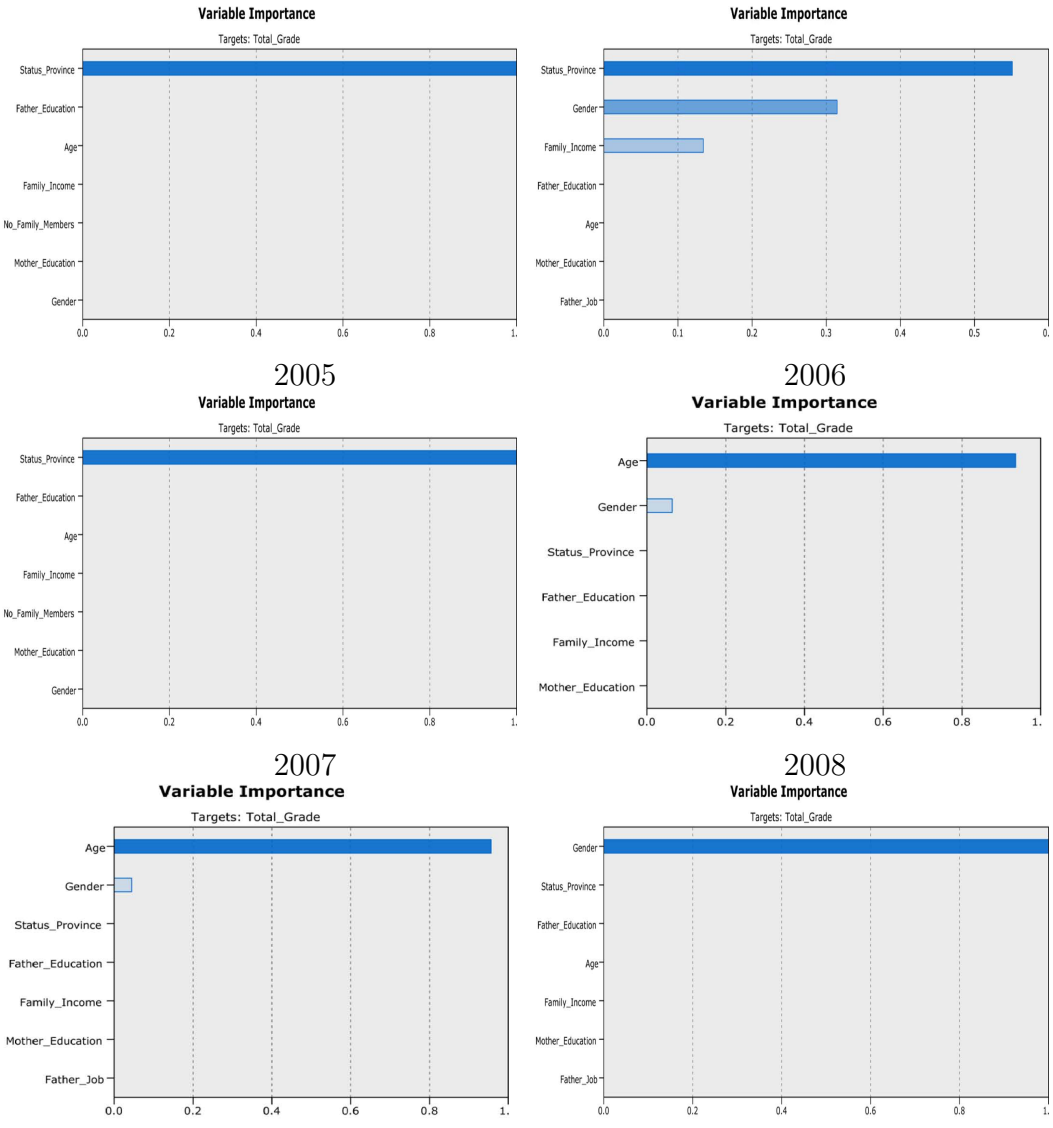

Figure 5.25: The CHAID Models for Prediction in 2005-2009

6. Examining Dynamic Aspects in the Observations

For an installed data mining system being used in daily operation, a user is concerned about the system's future performance as the extracted knowledge is based on past behavior of the analyzed objects. If future behavior is very similar to past behavior, using the initial data mining system can be justified. However, if the behavior changes over time, the continued use of the initial system might lead to non-acceptable results. This leads to a new research area, termed as *dynamic data mining*, that is concerned about any combination of traditional data mining with dynamic aspects.

Dynamic data mining is increasingly attracting attention from different research communities. Meanwhile there is an increasing interest in the related techniques among the user of the installed data mining systems since most of these installations will need to be updated in the future.

There are three basic approaches when a data miner wants to continue applying a data mining system. In the first approach, the user neglects changes in the environment and keeps on applying the initial system without updating. In the second approach, every certain period - which depends on the particular application - a new system is developed by using all the available data. In the third approach, based on the initial system and "new data", an update of the classifier is performed.

The aim of this chapter is to examine the dynamic aspects that might be available in the observations. If such aspects exist then a data miner can decide whether to use the second or the third approach mentioned above. To the best of our knowledge, when dealing with the Iranian educational data, there is no comprehensive study that takes into account such dynamic aspects.

Having the data over five years makes it possible for us to examine this phenomenon. In other words, it is possible to test the stability of a certain model by using the data in succeeding years. Such tests are the subject of

this chapter. Instability of a model indicates the existence of dynamics in observations and makes the use of dynamic data mining methods necessary.

In the following section we focus on classification models. The approach is, however, applicable for prediction models as well.

6.1. Classification Models

The algorithms we use for classification in this section are Neural Network, CART, CHAID, C5.0 and QUEST. Our target variable is here the candidate's WEE-acceptance. Beginning from 2005, we use the data of the first year for training and the data of the succeeding years for testing[1] of the model.

The results which are presented in Table 6.1 show that with the exception of two cases the accuracy rates calculated for the test data are significantly different from those calculated for the training data. Therefore a model generated by the data of one year generally cannot be used for classification of the target variables in the succeeding years. In other words, either a new model should be constructed or the old model should be adopted by using the new data.

We continue our examination and use the data of the two years 2005-2006, 2006-2007 and 2007-2008 as training datasets. The data of the succeeding years in each case have been used for testing of the three constructed classification models. The accuracy rates are presented in Table 6.2. The results show that in most of the cases the difference between the accuracy of the test and training data is significant. An exception is the case of the training datasets 2006-2007. This means that constructed classification models using two successive datasets cannot be used for the classification of the next year.

[1] As already mentioned in data mining method, for to evaluate the model which is created by training dataset, that analysis on the different part of dataset namely testing dataset.

Table 6.1: The Results of the Classification Models Based on One Year Datasets

Year 2005									
Neural Net		**CART**		**CHAID**		**C5.0**		**QUEST**	
Variable	**Imp. R.**	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**
Age	4.08E-1	Age	4.18E-1	Father Ed.	5.42E-1	Father Ed.	4.06E-1	Age	5.08E-1
Stat. Prov.	2.10E-1	Father Ed.	2.41E-1	Stat. Prov.	2.46E-1	Stat. Prov.	3.13E-1	Father Ed.	4.06E-1
Mother Ed.	1.18E-1	Stat. Prov.	1.93E-1	Mother Ed.	1.07E-1	Mother Ed.	1.48E-1	Mother Ed.	4.72E-2
Father Ed.	1.06E-1	Mother Ed.	7.67E-2	Family Inc.	0.92E-1	Age	1.00E-1	Stat. Prov.	3.82E-2
Father Occ.	0.68E-1	Family Inc.	7.15E-2	Father Occ.	0.13E-1	Family Inc.	0.33E-1	Family Inc.	3.50E-4
Family Inc.	0.68E-1	Father Occ.	2.05E-4	Gender	0.0	Gender	0.0	Father Occ.	3.50E-4
Gender	0.22E-1	-	-	Age	0.0	-	-	-	-
A. R.* 2005	75.89%	75.99%		75.94%		76.07%		75.95%	
A. R. 2006	64.52%	64.31%		64.31%		64.56%		64.49%	
A. R. 2007	55.19%	54.67%		54.67%		55.04%		55.17%	
A. R. 2008	63.31%	62.95%		62.95%		63.22%		63.2%	
A. R. 2009	52.57%	51.94%		51.94%		52.37%		52.31%	
Year 2006									
Age	4.54E-1	Age	5.86E-1	Stat. Prov.	4.65E-1	Age	4.56E-1	Age	5.08E-1
Stat. Prov.	1.43E-1	Stat. Prov.	2.55E-1	Father Ed.	3.81E-1	Stat. Prov.	2.65E-1	Father Ed.	4.06E-1
Family Inc.	1.16E-1	Father Ed.	1.59E-1	Family Inc.	0.89E-1	Father Ed.	1.76E-1	Mother Ed.	4.72E-2
Father Ed.	0.97E-1	Mother Ed.	0.0	Mother Ed.	0.65E-1	Mother Ed.	0.94E-1	Stat. Prov.	3.82E-2
Mother Ed.	0.93E-1	Father Occ.	0.0	Gender	0.0	Family Inc.	0.06E-1	Family Inc.	3.57E-4
Father Occ.	0.81E-1	-	-	Father Occ.	0.0	Father Occ.	0.03E-1	Father Occ.	3.57E-4
Gender	0.15E-1	-	-	-	-	Gender	0.0	-	-
A. R. 2006	64.28%	64.64%		64.45%		65.24%		75.95%	
A. R. 2007	54.67%	56.59%		55.33%		57.12%		55.17%	
A. R. 2008	62.95%	63.72%		62.96%		63.46%		63.2%	
A. R. 2009	51.94%	53.82%		52.46%		53.99%		52.31%	
Year 2007									
Age	5.75E-1	Age	8.50E-1	Age	7.27E-1	Age	7.07E-1	Age	8.54E-1
Stat. Prov.	1.07E-1	Stat. Prov.	1.24E-1	Stat. Prov.	1.51E-1	Stat. Prov.	1.71E-1	Stat. Prov.	1.45E-1
Father Occ.	1.02E-1	Mother Ed.	4.88E-2	Mother Ed.	0.96E-1	Mother Ed.	0.76E-1	Gender	7.12E-4
Father Ed.	0.072E-1	Father Ed.	7.2E-5	Father Occ.	0.15E-1	Father Ed.	0.33E-1	Mother Ed.	6.987E-5
Mother Ed.	0.58E-1	Family Inc.	7.2E-5	Father Ed.	0.05E-1	Family Inc.	0.07E-1	-	-
Family Inc.	0.53E-1	Father Occ.	7.2E-5	Family Inc.	0.04E-1	Father Occ.	0.04E-1	-	-
Gender	0.33E-1	-	-	Gender	0.01E-1	Gender	0.02E-1	-	-
A. R. 2007	59.45%	60.98%		61.34%		62.22%		60.74%	
A. R. 2008	57.43%	60.72%		62.61%		62.56%		61.43%	
A. R. 2009	56.89%	57.39%		57.58%		58.34%		57.04%	
Year 2008									
Age	4.47E-1	Age	7.97E-1	Age	7.14E-1	Age	6.14E-1	Age	8.78E-1
Stat. Prov.	1.45E-1	Mother Ed.	1.76E-1	Mother Ed.	1.54E-1	Stat. Prov.	1.55E-1	Mother Ed.	1.03E-1
Family Inc.	1.18E-1	Stat. Prov.	2.65E-2	Stat. Prov.	1.09E-1	Mother Ed.	1.12E-1	Stat. Prov.	1.92E-2
Mother Ed.	0.96E-1	Father Ed.	7.86E-5	Family Inc.	0.23E-1	Father Ed.	0.89E-1	Father Ed.	8.45E-5
Father Occ.	0.87E-1	Family Inc.	7.86E-5	Father Ed.	0.0	Family Inc.	0.16E-1	Family Inc.	8.45E-5
Father Ed.	0.81E-1	Father Occ.	7.86E-5	Father Occ.	0.0	Father Occ.	0.13E-1	Father Occ.	8.45E-5
Gender	0.26E-1	-	-	Gender	0.0	Gender	0.001	-	-
A. R. 2008	63.08%	63.88%		64.37%		65.42%		63.83%	
A. R. 2009	53.78%	55.92%		56.26%		56.77%		54.52%	

*A. R.indicated as accuracy rate.

Table 6.2: The Results of the Classification Models Based on Two Years Datasets

Year 2005, 2006									
Neural Net		**CART**		**CHAID**		**C5.0**		**QUEST**	
Variable	**Imp. R.**	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**
Father Ed.	2.56E-1	Age	5.18E-1	Age	5.62E-1	Age	4.43E-1	Age	5.43E-1
Father Occ.	2.23E-1	Father Ed.	5.41E-1	Father Ed.	2.57E-1	Stat. Prov.	2.32E-1	Stat. Prov.	2.16E-1
Age	1.73E-1	Mother Ed.	9.34E-2	Mother Ed.	0.88E-1	Father Ed.	1.95E-1	Father Ed.	1.95E-1
Mother Ed.	1.23E-1	Stat. Prov.	4.76E-2	Family Inc.	0.62E-1	Mother Ed.	0.99E-1	Mother Ed.	4.65E-2
Family Inc.	1.10E-1	Father Occ.	3.44E-5	Stat. Prov.	0.16E-1	Family Inc.	0.18E-1	Family Inc.	3.75E-5
Stat. Prov.	0.83E-1	-	-	Gender	0.0	Father Occ.	0.12E-1	Father Occ.	3.75E-5
Gender	0.32E-1	-	-	Father Occ.	0.0	Gender	0.0	-	-
A. R. 05, 06	70.18%	70.32%		70.32%		70.82%		70.3%	
A. R. 2007	54.74%	55.75%		55.34%		55.82%		55.92%	
A. R. 2008	62.98%	63.55%		63.38%		63.39%		63.58%	
A. R. 2009	52.0%	52.94%		52.77%		52.97%		52.98%	
Year 2006, 2007									
Age	4E-1.452	Age	0.509	Age	6.45E-1	Age	5.09E-1	Age	7.63E-1
Mother Ed.	1.45E-1	Mother Ed.	2.29E-1	Mother Ed.	2.43E-1	Stat. Prov.	2.59E-1	Mother Ed.	1.98E-1
Family Inc.	1.11E-1	Stat. Prov.	1.44E-1	Father Ed.	0.62E-1	Mother Ed.	1.40E-1	Stat. Prov.	2.11E-2
Father Occ.	0.91E-1	Father Ed.	2.05E-5	Stat. Prov.	0.32E-1	Father Ed.	0.69E-1	Father Ed.	1.87E-2
Stat. Prov.	0.89E-1	Family Inc.	2.05E-5	Father Occ.	0.15E-1	Father Occ.	0.13E-1	Father Occ.	3.14E-5
Father Ed.	0.72E-1	Father Occ.	2.05E-5	Gender	0.03E-1	Family Inc.	0.07E-1	-	-
Gender	0.40E-1	-	-	Family Inc.	0.0	Gender	0.02E-1	-	-
A. R. 06, 07	60.95.%	61.98%		62.27%		64.15%		62.25%	
A. R. 2008	63.46%	61.93%		62.68%		63.12%		63.12%	
A. R. 2009	54.19%	57.63%		56.73%		57.68%		56.66%	
Year 2007, 2008									
Age	0.562	Age	8.01E-1	Age	0.79	Age	0.661	Age	7.57E-1
Mother Ed.	0.131	Stat. Prov.	1.04E-1	Mother Ed.	0.134	Stat. Prov.	0.175	Mother Ed.	9.2E-2
Family Inc.	0.101	Mother Ed.	9.54E-2	Father Ed.	0.031	Mother Ed.	0.096	Father Ed.	6.55E-2
Father Occ.	0.07	Father Ed.	1.65E-5	Stat. Prov.	0.024	Father Ed.	0.054	Stat. Prov.	6.04E-2
Father Ed.	0.06	Family Inc.	1.65E-5	Father Occ.	0.015	Family Inc.	0.007	Father Occ.	2.54E-2
Stat. Prov.	0.055	Father Occ.	1.65E-5	Family Inc.	0.003	Father Occ.	0.005	Family Inc.	1.59E-5
Gender	0.021	-	-	Gender	0.002	Gender	0.002	Gender	1.59E-5
A. R. 07, 08	59.7%	62.29%		61.99%		64.02%		61.92%	
A. R. 2009	53.0%	57.06%		56.7%		57.68%		57.2%	

Table 6.3: The Results of the Classification Models Based on Three Years Datasets

Year 2005-2007									
Neural Net		**CART**		**CHAID**		**C5.0**		**QUEST**	
Variable	**Imp. R.**	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**
Age	4.14E-1	Age	5.81E-1	Age	6.39E-1	Age	4.11E-1	Age	7.21E-1
Mother Ed.	2.07E-1	Stat. Prov.	2.88E-1	Mother Edu.	1.66E-1	Stat. Prov.	4.04E-1	Stat. Prov.	1.92E-1
Father Occ.	1.66E-1	Father Ed.	1.31E-1	Father Ed.	1.16E-1	Father Ed.	9.70E-2	Father Ed.	4.88E-2
Stat. Prov.	0.82E-1	Mother Ed.	1.67E-5	Stat. Prov.	0.57E-1	Mother Ed.	7.18E-2	Mother Ed.	3.80E-2
Father Ed.	0.63E-1	Father Occ.	1.67E-5	Father Occ.	0.16E-1	Father Occ.	1.06E-2	Father Occ.	2.04E-5
Family Inc.	0.52E-1	-	-	Family Inc.	0.04E-1	Family Inc.	1.16E-3	-	-
Gender	0.15E-1	-	-	Gender	0.02E-1	Gender	3.54E-4	-	-
A. R. 05-07	61.35%	62.1%		61.78%		63.66%		61.7%	
A. R. 2008	63.49%	62.46%		63.85%		62.86%		63.75%	
A. R. 2009	53.03%	55.38%		53.72%		56.49%		53.48%	
Year 2006-2008									
Age	5.19E-1	Age	7.77E-1	Age	6.96E-1	Age	5.66E-1	Age	6.47E-1
Stat. Prov.	1.20E-1	Mother Ed.	1.92E-1	Mother Ed.	2.01E-1	Stat. Prov.	2.17E-1	Mother Ed.	2.42E-1
Mother Ed.	1.06E-1	Stat. Prov.	3.09E-2	Father Ed.	0.54E-1	Mother Ed.	1.20E-1	Father Ed.	9.60E-2
Father Ed.	1.00E-1	Father Ed.	1.43E-5	Father Occ.	0.33E-1	Father Ed.	0.80E-1	Stat. Prov.	1.57E-2
Father Occ.	0.65E-1	-	-	Family Inc.	0.09E-1	Family Inc.	0.08E-1	Family Inc.	1.57E-5
Family Inc.	0.61E-1	-	-	Stat. Prov.	0.05E-1	Father Occ.	0.07E-1	Father Occ.	1.57E-5
Gender	0.29E-1	-	-	Gender	0.01E-1	Gender	0.02E-1	-	-
A. R. 06-08	62.06%	62.65%		62.5%		64.57%		62.62%	
A. R. 2009	54.43%	55.66%		55.64%		57.26%		56.51%	

Table 6.4: The Results of the Classification Models Based on Four Years Datasets

Year 2005-2008									
Neural Net		**CART**		**CHAID**		**C5.0**		**QUEST**	
Variable	**Imp. R.**	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**	**Variable**	**Imp. R.**
Age	4.58E-1	Age	7.61E-1	Father Ed.	4.87E-1	Age	4.73E-1	Age	8.30
Family Inc.	1.46E-1	Father Ed.	2.08E-1	Mother Ed.	3.82E-1	Stat. Prov.	3.29E-1	Father Ed.	0.87E-1
Father Ed.	1.15R-1	Stat. Prov.	3.16E-2	Family Inc.	4.51E-2	Father Ed.	9.61E-2	Mother Ed.	0.75E-1
Father Occ.	0.96E-1	Mother Ed.	1.49E-5	Stat. Prov.	4.38E-2	Mother Ed.	7.78E-2	Stat. Prov.	0.08E-1
Stat. Prov.	0.84E-1	Father Occ.	1.49E-5	Father Occ.	4.07E-2	Father Occ.	1.27E-2	-	-
Mother Ed.	0.77E-1	-	-	Gender	7.02E-4	Family Inc.	1.07E-2	-	-
Gender	0.23E-1	-	-	-	-	Gender	3.54E-4	-	-
A. R. 05-08	64.64%	65.29%		64.8%		67.39%		65.05%	
A. R. 2009	51.94%	54.03%		52.78%		56.11%		53.86%	

The other results that are presented in the Tables 6.3 and 6.4 support the above findings in general.

6.2. Conclusion

Put together the results indicate that for our observations a model which is generated using the data in a certain year cannot be used always for classification in the succeeding years. In other words, application of a dynamic data mining approach is a better alternative.

7. Conclusion and Further Research

The goal of this thesis was to conduct a focused and in-depth comprehensive study of the impact of socioeconomic status of the Iranian Wide Entrance Examination (WEE) applicants' families on the educational achievement of their children. To reach this goal we used various statistical methods, such as variance and regression analysis, as well as data mining techniques, including various kinds of decision trees and artificial neural networks.

The following sections include our results in summary

7.1. Comprehensive Testing of Applicability of Data Mining Methods

Since middle of the last century, the scientists from the field "Machine Learning" (ML) have developed different methods that can be used for classification, forecasting and clustering. Decision Trees, Artificial Neural Networks and Self-organizing map are examples of such approaches. This means that besides the statistical approaches like Discriminant Analysis, Variance Analysis, Regression Analysis and Clustering Analysis, researchers can today use the above mentioned ML-Based-procedures as well. ML-Based methods are often called Data Mining approaches.

In the last years Data Mining procedures were applied in different areas. In some cases their performance has been better than the classical statistical approaches, see Michie et al. [64]. This fact has led us to use in our study not only the classical statistical approaches but also Data Mining methods. In our case study, we have shown that the Data Mining algorithms can be applied successfully as an alternative to the statistical approaches. Our results show that in many situations the performance of Data Mining approaches are even better than the statistical methods. Specially, we have observed that if the understandability of the results is important, decision tree is the best choice

for this purpose.

7.2. Construction of Quantitative Classification and Forecasting Models for Performance Prediction

The classification and forecasting models which we have constructed and described in the fifth chapter can be applied by interested users e.g. WEE-applicants, their parents, other researchers, higher education planners etc. Among other uses of such prediction results, the WEE-applicants can be well prepared for the WEE.

7.3. Impact of the Results of Static Models

As mentioned before, the main goal of our research was to conduct of a comprehensive study of the impact of socioeconomic status of the Iranian Wide Entrance Examination (WEE) applicants' families on the educational achievement of their children. We have described in detail the results we obtained by using the static models in the fifth chapter. The results have very interesting application aspects. Two examples:

We discovered that the province-residence of the applicants has an effect on their educational outcome. However, the ordering of the provinces as level of socioeconomic status which is coded by the NOET is different from the results we obtained by using analysis of variance. This means that the NOTE should examine his coding procedure.

The second example is about the effect of applicants' gender and the education level of their parents. We learned that the WEE-performance of the girls is affected by the education level of the mother. On other hand, the WEE-performance of the boys is depended significantly on the father's education.

7.4. Examining of Dynamic Behavior of Observations

To the best of our knowledge, when dealing with the Iranian educational data, there is no comprehensive study that takes into account the dynamic behavior of observations.

Having the data over five years made it possible for us to examine this concept. In other words, it was possible to test the stability of a certain model

by using the data in succeeding years. Instability of a model indicates the existence of dynamics in observations.

The results of the sixth chapter indicate that for our observations a model which is generated using the data in a certain year cannot always be used in the succeeding years for classification. In other words, application of a Dynamic Data Mining approach is a better alternative.

7.5. Preparation of a Relative big Dataset for Alternative Studies

The two WEE-questionnaires that are applied for collection of our used data are answered by the applicants. They are well aware that wrong information could lead to their disqualification. Due to this fact the quality of the data used is generally good. In understanding data preparation phases described in the fourth chapter, however, we have observed that certain data cleaning operations are necessary. Eliminating of duplicates and outliers are examples. A dataset with about six million cleaned observations that we have generated at the end of the data preparation process is not only used in our study, but, can be used in other similar projects in future.

7.6. Future Research

In this section, we present some of our ideas for further research:

- The data we used are collected in the period 2005 to 2009. Meanwhile, the new data from 2010 to 2014 are available as well. Given the fact that the WEE in Iran will continue in to the future, the WEE-data remains as a very interesting panel data source that can be used, specially, for the exploration of structural change of the target variables we used in our study.

- Besides the WEE-Data, many universities in Iran collect information about their students. The combination of such data with WEE-Data makes it possible to examine the impact of socioeconomic status of WEE-applicants' families on the performance of their children during their study period at a university. A lot of other studies can be conducted by using the combined dataset, as well.

- In the sixth chapter, we have observed that in some cases our observations have a dynamic effect. Using Dynamic Data Mining allows construction

of classification and forecasting models by considering such dynamic aspects. This could be the subject of a new research project.

- Last but not least, the methodology we used in our study can be applied to the educational data of the other countries.

Bibliography

[1] Agresti, A. (2007). *An Introduction to Categorical Data Analysis. Second edition.* Wiley Series in Probability and Statistics.

[2] Albert, C. (2000). Higher education demand in spain: The influence of labour market signals and family background. *Higher Education*, 40(2):147–162.

[3] Allison, P. D. (2001). *Logistic Regression Using the SAS System : Theory and Application.* Wiley-SAS.

[4] Belley, P., Frenette, M., and Lochner, L. (2008). Educational attainment by parental income: A canada-us comparison. Technical report, mimeo, Dept of Economics, University of Western Ontario.

[5] Berry, M. J. and Linoff, G. S. (2004). *Data mining techniques: for marketing, sales, and customer relationship management.* Wiley. com.

[6] Bhardwaj, B. K. and Pal, S. (2011). Mining educational data to analyze students' performance. *International Journal of Advanced Computer Science and Applications*, 2(6):63–69.

[7] Blau, D. M. (1999). The effect of income on child development. *Review of Economics and Statistics*, 81(2):261–276.

[8] Blau, P. M. and Duncan, O. D. (1967). *The American Occupational Structure.* New York: Wiley and Sons.

[9] Bourdieu, P. (Eds, . O. (1977). *Cultural Reproduction and Social Reproduction In Power and Ideology in Education.* Oxford University Press.

[10] Bourdieu, P. (1974). *Cultural reproduction and social reproduction In Knowledge, Education, and Cultural Change.* Tavistock Publications.

[11] Bramer, M. (2007). *Principles of data mining.* Springer.

[12] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. T. (1984). *Classification and Regression Trees.* Wadsworth, Belmont, California,.

[13] Buchmann, C. (2002). Measuring family background in international studies of education: Conceptual issues and methodological challenges. *Methodological advances in cross-national surveys of educational achievement*, pages 150–197.

[14] Cameron, S. V. and Heckman., J. J. (2001). The dynamics of educational attainment for black hispanic, and white males. *Journal of Political Economy*, 109(3):455–499.

[15] Chambers, E. A. and Schreiber, J. B. (2004). Girls academic achievement: Varying associations of extracurricular activities. *Gender and Education*, 16:327–346.

[16] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). Step by step data mining guide, crisp. Technical report, CRISP, DM Consortium.

[17] Cunha, F. and Heckman, J. J. (2006). Investing in our young people. Technical report, University of Chicago, Department of Economics.

[18] Dasu, T. and Johnson, T. (2003). *Exploratory data mining and data cleaning*, volume 479. Wiley. com.

[19] de Dios Jiménez, J. and Salas-Velasco, M. (2000). Modeling educational choices. a binomial logit model applied to the demand for higher education. *Higher Education*, 40(3):293–311.

[20] Dobson, A. (1983). *Introduction to Statistical Modelling.* Chapman and Hall.

[21] Eitle, T. M. (2005). Do gender and race matter? explaining the relationship between sports participation and achievement. *Sociological Spectrum*, 25:177–195.

[22] Erdogan, S. Z. and Timor, M. (2005). A data mining application in a student database. *Journal of Aeronautics and Space Technologies*, 2(2):53–57.

[23] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996a). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37.

[24] Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. & Uthurusamy, R. (1996b). *Advances in knowledge discovery and data mining.* MIT Press.

[25] Featherman, D. L. and Hauser, R. M. (1978). *Opportunity and Change.* New Yourk: Academic Press.

[26] Fu, X. and Wang, L. (2005). *Data mining with computational intelligence.* Springer-Verlag Berlin Heidelberg.

[27] Gharun, M. (2002). Comparison of the effects of socio-economic factors on the demand for admission to higher education (in persian). *Journal of Educational Science, Research and Planning in Higher Education*, 40:91–110.

[28] Giudici, P. (2003). *Applied Data Mining: Statistical Method for Business and Industry.* John Wiley & Sons.

[29] Giudici, P. and Figini, S. (2009). *Applied Data Mining for Business and Industry.* Wiley Online Library.

[30] Gladieux, L. E. and Swail, W. S. (1998). Financial aid is not enough: Improving the odds of college success. *The College Board Review*, 185:16–21, 30–31.

[31] Gota, A. A. (2012). *Effects of parenting styles, academic self-efficacy, and achievement motivation on the academic achievement of university students in Ethiopia.* PhD thesis, Faculty of Computing, Health, and Science, Edith Cowan University.

[32] Gottfredson, D. C. (1981). Black-white differences in the educational attainment process: what have we learned? *American Sociological Review*, pages 542–557.

[33] Han, J., Kamber, M., and Pei, J. (2006). *Data mining: concepts and techniques.* Morgan kaufmann.

[34] Heckman, J. J. and Masterov, D. V. (2007). The productivity argument for investing in young children. *Applied Economic Perspectives and Policy*, 29(3):446–493.

[35] Iacovou, M. (2001). Family composition and childrens educational outcomes. Technical report, Institute for social and economic research, University of Essex.

[36] Jamali, E. (2010). The effects of socioeconomic status on educational performance of higher education applicants in iran (in persian). *Quarterly Journal of Higher Education in Iran,*, 10:25–54.

[37] Jamali, E. (2012). The trend of socioeconomic status of nation wide examination applicants on their educational performances in 2001 to 2009 (in persian). *Quarterly Journal of Higher Education in Iran,*, 16:25–56.

[38] Jeynes, W. H. (2002). Examine the effects of the academic achievement of adolescents: the challenge of controlling for family income. *Journal of Family and Economic*, 23(2):189–210.

[39] Jimenez, J. d. D. and Salas-Velasco, M. (2000). Modeling educational choices. a binomial logit model applied to the demand for higher education. *Journal of Higher Education*, 40(1):293–311.

[40] Johnson, T. and Dasu, T. (2003). Data quality and data cleaning: An overview. In *ACM SIGMOD international conference on Management of data*.

[41] Kalmijn, M., K. G. (1996). Race, cultural capital, and schooling an analysis of trends in the united states. *Sociology of education*, 69:22–34.

[42] Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29(2):119–127.

[43] Khaeir, M. (1997). Investigate the relationship between some indicators of social class and academic achievement in a group of high school students in the first year of the new system (in persian). *Humanities and Social Sciences, Shiraz University*, 24:77–114.

[44] KhÃŽi, L. T. . (1986). Toward a general theory of education. *Comparative Education Review*, 30(1):12–29.

[45] Khodaie, E. (2009). A study on the relationship between parental socioeconomic capital and the probability of students' acceptance in the 2006

nationwide examination of iran (in persian). *Quarterly Journal of Higher Education in Iran*, 4:65–84.

[46] Khodaie, E. (2010). Effective factors on passing in national entrance exam in postgraduate level (in persian). *Quarterly journal of Research and Planning in Higher Education,*, 15(4):19–34.

[47] Kleinbaum, D. G. and Klein., M. (2010). *Analysis of Matched Data Using Logistic Regression.* Springer New York.

[48] Kodde, D. A. and Ritzen, J. M. (1988). Direct and indirect effects of parental education level on the demand for higher education. *Journal of Human Resources*, 23(3):356–371.

[49] Kumar, V. and Chadha, A. (2011). An empirical study of the applications of data mining techniques in higher education. internat. *Journal of Advanced Computer Science and Applications*, 2(3):80–84.

[50] Kutner, M., Nachtsheim, C., Neter, J., and Li, W. (2004). *Applied Linear Statistical Models*. 5th edition, McGraw-Hill.

[51] Lanquillon, C. (1997). Dynamic neural classication. Master's thesis, University of Braunschweig.

[52] Larose, D. T. (2005). *Discovering knowledge in data: an introduction to data mining*. Wiley. com.

[53] Lefebvre, P. and Merrigan, P. (2010). The impact of family background, cognitive and non-cognitive ability in childhood on post-secondary education participation. Technical report, Human Resources and Social Development Canada.

[54] Loh, W. and Shih, Y. (1997). Split selection methods for classification trees. *Statistica sinica*, 7(4):815–840.

[55] Luan, J. (2002a). *Data mining and its applications in higher education.*, chapter 2, pages 17–36. Wiley Periodicals, Inc.

[56] Luan, J. (2002b). Data mining and knowledge management in higher education-potential applications. In *In Workshop associate of institutional research international conference, Toronto*.

[57] Luan, J. (2004). Data mining applications in higher education. Technical report, SPSS Executive, 7.

[58] Lynch, P. D. (1975). School and family predictors of achievement and dropout in elementary schools of a developing country. Technical report, University Park, Pennsylvania.

[59] Macmillan, I. R. (1998). *Growing up scared, the effects of violent victimization in adolescence on adult socio-economic attainment.* PhD thesis, Department of Sociology, University of Toronto.

[60] Majoribanks, K. (1996). Family learning environments and students outcomes. *Journal of Comparative Family Studies*, 27:373–394.

[61] Marini, M. M. (1978). The transition to adulthood: Sex differences in educational attainment and age at marriage. *American Sociological Review*, pages 483–507.

[62] McGaw, B. and Lievesley, D. (2003). Literacy skills for the world of tomorrow, chapter 6, family background and literacy performance. Technical report, OECD Programme for International Student Assessment (PISA) - UNESCO.

[63] Meshkani, A. and Nazemi, A. (2009). *Introduction to Data Mining (in Persian).* Azad University Press.

[64] Michie, Donald, D. J. S. and Taylor., C. C. (1994). *Machine learning, neural and statistical classification.*

[65] Mohammadi, A. (1995). The impact of social stratification on educational opportunities (in persian). Master's thesis, University of Tehran.

[66] Montgomery, D. C. (2001). *Design and Analysis of Experiments (5th ed.).* New York: Wiley.

[67] Montgomery, D. C., P. E. A. . V. G. G. (2012). *Introduction to linear regression analysis.* John Wiley & Sons.

[68] Nazemi, A. (2012). *Credit Risk Management with Data Mining Methodology.* Shaker.

[69] Noghani, M. (2007). The effect of cultural capital on academic achievement in high school students inequalities in access to higher education (in persian). *Academic Journal of Human Sciences - Research*, 91:71–101.

[70] Parack, S., Zahid, Z., and Merchant, F. (2012). Application of data mining in educational databases for predicting academic trends and patterns. In *International Conference onTechnology Enhanced Education (ICTEE), IEEE*.

[71] Pedrosa, R. H., Dachs, J. N. W., Maia, R. P., Andrade, C. Y., and Carvalho, B. S. (2006). Educational and socioeconomic background of undergraduates and academic performance: consequences for affirmative action programs at a brazilian research university. In *IMHE/OECD General Conference, Paris,*.

[72] Psacharopoulos, G. and Sanyal, B. (1981a). Student expectations and labour market performance: The case of the philippines. *Higher Education*, 10(4):449–472.

[73] Psacharopoulos, G. and Sanyal, B. C. (1981b). *Higher Education and Employment: The IIEP Experience in Five Less Developed Countries.* IIEP Publications, International Institute for Educational Planning,.

[74] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers.

[75] Roshan, A. R. and Salehi, M. J. (2003). Determination the social-economic status of students at the public universities in tehran (in persian). Technical report, Institute for Research and Planning in Higher Education.

[76] Rozada, M. G. and Menendez, A. (2001). Higher education subsidies in argentinam. Technical report, Center for International Higher Education.

[77] Sacker, A., Schoon, I., and Bartley, M. (2002). Social inequality in educational achievement and psychological adjustment throughout childhood: magnitude and mechanisms. *Social Science and Medicine*, 55:863–880.

[78] Sarandy, P. (1997). An investigation of the effects of some factors influencing the sstudying quality of tabriz university students (in persian). Technical report, Tabriz University.

[79] Shahi, M. R. A. (2009). *Developing Poweful and Comprehensible Models for Corporate Default Prediction. PhD Thesis, University of Southampton.* PhD thesis, School of Management, University of Southampton.

[80] Shi, Y. (2006). *Dynamic data mining on multi-dimensional data.* PhD thesis, State University of New York at Buffalo.

[81] Singh, C. and Gopal, A. (2010). Performance analysis of faculty using data mining techniques. *International Journal of Computer Science and Application*, pages 140–144.

[82] Stefansky, W. (1971). Rejecting outliers by maximum normed residual. *The Annals of Mathematical Statistics*, 42 (1):35–45.

[83] Stefansky, W. (1972). Rejecting outliers in factorial designs. *Technometrics*, 14(2):469–479.

[84] Steinbach, M. (2000). An introduction to cluster analysis for data mining. a cluster survey. Technical report.

[85] Sun, Y. (1999). The contextual effects of community social capital on academic performance. *Social Science Research*, 28(4):403–426.

[86] Sun, Y. (2001). Family environment and adolescents' well-being before and after parents' marital disruption: A longitudinal analysis. *Journal of Marriage and Family*, 63(3):697–713.

[87] Sun, Y., Hobbs, D., Elder, W., and Li, Y. (1994). Multi-level analyses of television viewing among high school students: A contrast between nonmetropolitan rural and other communities. *Journal ofResearch in Rural Education*, 10(2):97–107.

[88] Susnea, E. (2010). Using artificial neural networks in e-lerning systems. *Scientific Bulletin-University Politehnica of Bucharest*, 72 (4):91–100.

[89] Taber, C. R. (2001). The rising college premium in the eighties return to college or return to unobserved ability. *The Review of Economic Studies*, 68(3):665–691.

[90] Tan, P.-N., Steinbach, M., and Kumar, V. (2006). *Introduction to Data Mining.* Pearson Addison-Wesley.

[91] Western, J., McMillan, J., Durrington, D., et al. (1998). Differential access to higher education: The measurement of socioeconomic status, rurality and isolation. Technical report, Department of Employment, Education, Training and Youth Affairs, University of Queensland.

A. Appendix

A.1. List of our Publications

Journal Articles:

- Mirashrafi, S. B., Bol, G., and Nakhaeizadeh, G. (2013a) The Effect of Family Background, Socioeconomic Status and Individual Factors on University Admission in Iran. *International Journal for Cross-Disciplinary Subjects in Education (IJCDSE), 4(2): 1130-1136.*

- Mirashrafi, S. B., Bol, G., and Nakhaeizadeh, G. (2013b) The effect of individual factors, family background and socioeconomic status on university admission in Iran in 2007. *Literacy Information and Computer Education Journal (LICEJ), 4(3): 1042–1048.*

- Mirashrafi, S. B., Bol, G., and Nakhaeizadeh, G. (2013c) The Effect of Family Background and Socioeconomic Status on Academic Performance of Higher Education Applicants. *International Journal of Technology and Inclusive Education (IJTIE), 2(1): 130-138.*

- Mirashrafi, S. B., Khodaie, E., and Jamali E. (2016) Family Background and Socioeconomic Status Effects on Educational Performances by Data Mining Methods: A Case Study in Iran. *International Journal for Cross-Disciplinary Subjects in Education (IJCDSE), 7(1): 2735-2741.*

Conference Papers:

- Mirashrafi, S. B., Bol, G., Khodaie, E., and Nakhaeizadeh, G. (2012a) The impact of family background and socioeconomic status on university admission in Iran. *In ATINER'S Conference Paper Series, No: EDU2012- 0141, Athens, Greece.*

- Mirashrafi, S. B., Bol, G., and Nakhaeizadeh, G. (2012b) The impact of family background and socioeconomic status on university admission in

Iran in 2009. *In Canada International Conference on Education (CICE-2012), Guelph, Ontario, Canada.*

- Mirashrafi, S. B., Bol, G., and Nakhaeizadeh, G. (2012c) The impact of family background and socioeconomic status on university admission in Iran. *In Ireland International Conference on Education (IICE-2012), Dublin, Ireland.*

- Mirashrafi, S. B., Bol, G., and Nakhaeizadeh, G. (2012d) The impact of family background and socioeconomic status on university admission in Iran in 2008. *In London International Conference on Education (LICE-2012), London, UK.*

- Mirashrafi, S. B., Khodaie, E., and Jamali E. (2014a) The Effects of Family Background and Socioeconomic Status on Educational Performances in Iran. *In Ireland International Conference on Education (IICE-2014), Dublin, Ireland.*

- Mirashrafi, S. B., Khodaie, E., and Jamali E. (2014b) The Effects of Socioeconomic Status of Iranian Wide Entrance Examination applicants on Their Educational Performance during 2001-2009 by Data Mining Techniques. *6th International Conference on Education and New Learning Technologies (EDULEARN14), Barcelona, Spain.*

- Mirashrafi, S. B., Khodaie, E., and Jamali E. (2014c) The Effects of Family Background and Socioeconomic Status on Educational Performances in Iran by Data Mining Techniques. *The 12th Iranian Statistical Conference (ISC12), Kermanshah, Iran.*

- Mirashrafi, S. B., Jamali E., and Mehdi A. (2014d) Data Mining in Educational Databases: a Case Study. *International Conference on Business, Information, and Cultural Creative Industry (ICBIC2014), Taipei, Taiwan.*

A.2. Summary of literature

Measurement of Family Background Factors on Educational Outcomes

Tables A.1 to A.6 show a summary of the international studies according to the measurement of family background factors and their effects on educational outcomes. The majority of these studies show that parental education, family income, parental occupation as an indicator of socioeconomic status have effects on educational performance.

A.3. ANOVA Results for the Province Residence of WEE Applicants

Figure A.1 provides a result by the multiple/post-hoc group comparisons in ANOVA according to the mean of total grades of applicants by their province of residence for the five years (2005-2009).

Province of Residence	N	Subset																
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Sistan and Baluchestan	150746	4783.8933																	
Hormozgan	118722		4908.9689																
Khuzestan	349394			5196.6845															
Kerman	208509			5198.2927															
Lorestan	196017				5242.9880														
Bushehr	78200				5255.4118														
Ardebil	118184					5279.5374													
Gilan	159354					5283.9639	5283.9639												
Kohgiluyeh And Boyer Ahmad	101227					5290.2772	5290.2772												
Zanjan	81956						5304.1897												
Golestan	117442							5326.3304											
Ilam	84139							5342.6142											
Hamedan	144950								5392.0136										
Azarbaijan East	282527								5395.7140										
Abroad	2058									5439.9640									
Qazvin	77182									5440.0441									
Azerbaijan Western	193139									5455.5940	5455.5940								
Kermanshah	212740										5465.2735								
Central	91402										5475.5780	5475.5780							
Chahar Mahal Bakhtiari	92958										5476.0121	5476.0121							
Semnan	48450											5490.7917							
Mazandaran	245722												5516.7798						
North Khorasan	55749												5535.6752						
Fars	421438													5567.9966					
Tehran	976370														5636.0400				
Kurdistan	117492														5647.6389				
South Khorasan	42171															5647.6389			
Isfahan	373088															5660.3564			
Khorasan Razavi	374815															5664.7662			
Qom	80153																5687.9173		
Yazd	74302																	5709.9795	
Sig.		1.000	1.000	.821	.244	.346	.072	.127	.728	.168	.080	.179	.076	1.000	.276	.130	1.000	1.000	5978.4641
																			1.000

Figure A.1: Subsets of Province According to the Mean of Total Grade of Applicants in Total 2005-2009

Table A.1: International Studies of the Relationship Between Family Socioeconomic Status and Educational Outcomes

Author(s) of Study	Year	Country	Measures of Family Socioeconomic Status	Outcome	Results
Hansen & Haller	1973	Costa Rica	Occupational status consumption status (index of parental education, house construction, and household possessions)	Attainment	Indirect (through aspirations) on attainment
Comber & Keeves	1973	19 countries	Home background (index)	Science achievement	Positive effect
Kerckhoff	1974	Great Britain	Father's education, Father's occupational status	Attainment	Positive effects on attainment
Rosier	1974	Australia	Home circumstance	Science achievement	Positive effects on attainment
Shukla	1974	India	Father's occupational status, Father's education, Mother's education, Use of dictionary, Nnumber of books in the home, Family size	Science achievement in the home	Positive effect of father's occ. and books
Pollock	1974	Scotland	Father's occupational, Number of books in the home, Family size	Science achievement	Positive effects of father's occ.
Heyneman	1976	Uganda	Parents' occupational, Parents' education, Household possessions	National exam performance	Positive effect
Currie	1977	Uganda	Father's education, Father's occupational status	Attainment	Positive effects on attainment
Lanzas & Kingston	1981	Zaire	Education of relative with greatest influence on student's life (e.g. mother, father, uncle, grandparent)	English achievement	Modest positive effect for student's living with parents, no effect for those living with relatives
Cooksey	1981	Cameroon	Mother's and Father's education, Mother's and Father's occupation, Home amenities (running water, electricity, toilet, refrigerator, cooker)	National exam performance	Positive effect on performance
Niles	1981	Sri Lanka	Family SES (index of father's occupational, Father's education, Mother's education, family incom)	achievement	Positive effect

Table A.2: International Studies of the Relationship Between Family Socioeconomic Status and Educational Outcomes - Continue 1

Author(s) of Study	Year	Country	Measures of Family Socioeconomic Status	Outcome	Results
Cochrane & Jamison	1982	Thailand	Father's education, Mother's education, Landownership	Enrollment, Attainment	Education vars positive on enrollment Indirect (through aspirations) on attainment
Simkus & Andorka	1982	Hungary	Father's occupation	Attainment	Positive effects on attainment
Heyneman & Loxley	1983	29 countries	Father's occupation, Father's education, Mother's education, Books in home, Dictionary or other measure of consumption in home, Family incom	Science achievement	Positive effect, but smaller than school effects, especially in poorer countries
Behrman & Wolfe	1984	Nicaragua	Father's education, Mother's education, Number of siblings, Mother present	Attainment	Positive effects on attainment, stronger effect of mother's ed. than father's ed. on all children
Mukweso, Papagiannis & Milton	1984	Zaire	Father's education, Father's occupational status, Index of consumption goods	Attainment	Positive effects on attainment
Whyte & Parrish	1984	China	Father's education, Father's occupational status	Attainment	Positive effects on attainment
Heyneman, Jamison & Montenegro	1984	Philippines	Parents' education	Science math Filipino achievement	Positive effect net of textbooks
Robinson & Garnier	1985	France	Father's education, Father's class	Attainment	Positive effects on attainment
Smith & Cheung	1986	Philippines	Father's education, Father's occupational status	Attainment	Positive effects on attainment
Lockheed, Vail & Fuller	1986	Thailand	Father's occupation, Mother's education, Home language	Math achievement	Positive effect on math pretest, negligible effect on math post-test, net of pretest
Jamison & Lockheed	1987	Nepal	Father's education, Father's literacy, Father's modernity, Caste, Household landholdings	Enrollment	Positive effects on attainment

Table A.3: International Studies of the Relationship Between Family Socioeconomic Status and Educational Outcomes - Continue 2

Author(s) of Study	Year	Country	Measures of Family Socioeconomic Status	Outcome	Results
King & Lillard	1987	Malaysia	Father's education, Mother's education,	Attainment	Positive effects on attainment, larger effects of mother's ed. on daughter's attainment
Kodde and Ritzen	1988	Netherlands	Parental education, Family Income	Demand for higher education	Direct/indirect effects
Lockheed, Fuller & Nyirongo	1989	Thailand	Mother's education, Father's occupation	Math achieve-ment	Positive effects
Riddell	1989	Zimbabwe	Father's occupation, Father's education, Electricity in the home	English and math achieve-ment	Positive effect
Jimenez & Lockheed	1989	Thailand	Father's occupation, Mother's education, Home language	Math achieve-ment	Father's occupation positive on achievement gains for males in single-sex school, mother's ed. for females
Holloway, Fuller, Hess et al.	1990	Japan United States	Father's occupation, Father's education, Mother's education	Educational achieve-ment	Positive effect
Lee & Lockheed	1990	Nigeria	Father's occupation (professional versus non-professional)	Math achieve-ment	Positive effects, net of school type
Katsillis & Rubinson	1990	Greece	Family SES (index of Father's education, Father's occupation, Mother's education, Family income) Father's class status	Achievement (GPA)	Positive effects of family SES on achievement, no effect of father's class status
Pong & Post	1991	Hong Kong	Father's occupational status, Mother's education	Attainment	Positive effects on attainment
Lin & Bian	1991	China	Father's education, Father's occupational status	Attainment	Positive effects on attainment
Paterson	1991	Scotland	Father's education, Mother's education, Household composition	Attainment	Positive effects on attainment

Table A.4: International Studies of the Relationship Between Family Socioeconomic Status and Educational Outcomes - Continue 3

Author(s) of Study	Year	Country	Measures of Family Socioeconomic Status	Outcome	Results
Shavit & Pierce	1991	Israel	Mother's education, Father's education, Father's occupational status	Attainment	Positive effects on attainment
Lockheed & Longford	1991	Thailand	Father's occupation, Mother's education, Home use of four-function calculator, Home language	Math achievement	Positive effects
Zuzovsky & Aitkin	1991	Israel	Family SES (index of Father's occupation, Mother's education, household composition)	Science achievement	Positive effect but varies by school
Gamoran	1991	United States	Parent's education	Math achievement	Positive effect
Stevenson & Baker	1992	Japan	Father's education, Mother's education, Family income	University enrollment	Positive effects on University enrollment
Hout, Raftery & Bell	1993	United States	Father's education, Father's occupational status, Mother's education	Attainment	Positive effects on attainment
Blossfeld	1993	Germany	Father's education, Father's occupational status	Attainment	Positive effects on attainment
De Graaf & Ganzeboom	1993	Netherlands	Father's education, Father's occupational status	Attainment	Positive effects on attainment
Jonsson	1993	Sweden	Father's education, Father's occupational status	Attainment	Positive effects on attainment
Kerckhoff & Trott	1993	England Wales	Father's education, Father's occupational status	Attainment	Positive effects on attainment
Cobalti & Schizzerotto	1993	Italy	Father's education, Father's occupational status	Attainment	Positive effects on attainment
Buchmann, Charles & Sacchi	1993	Switzerland	Father's education, Father's occupational status	Attainment	Positive effects on attainment

Table A.5: International Studies of the Relationship Between Family Socioeconomic Status and Educational Outcomes - Continue 4

Author(s) of Study	Year	Country	Measures of Family Socioeconomic Status	Outcome	Results
Tsai & Chiu	1993	Taiwan	Father's education, Father's occupational status, Mother's education	Attainment	Positive effects on attainment
Treiman & Yamaguchi	1993	Japan	Father's education, Father's occupational status	Attainment	Positive effects on attainment
Mateju	1993	Czechoslovakia	Father's education, Father's occupational status	Attainment	Positive effects on attainment
Szelenyi & Aschaffen-burg	1993	Hungary	Father's education, Father's occupational status	Attainment	Positive effects on attainment
Heyns & Bialecki	1993	Poland	Father's education, Father's occupational status	Attainment	Positive effects on attainment
Shavit	1993	Israel	Father's education, Father's occupational status	Attainment	Positive effects on attainment
Lillard & Willis	1994	Malaysia	Father's education, Father's earning, Mothers education	Attainment	Positive effects on attainment, Mother's ed. stronger for daughters
Fuller, Singer & Keiley	1995	Botswana	Mother's education, Mother's employment status, senior male's employment status, household quality and possessions	Enrollment	Mother education significantly related to dropout, no effects of other variables
Gerber & Hout	1995	Soviet Russia	Parents' education, occupational status of main income earner in household	Attainment	Positive effects on attainment
Baker, Riordan & Schaub	1995	Belgium New Zealand Thailand Japan	Father's occupation, Mother's education, Home language	Math achieve-ment	Used as a control to model effect of mixed versus single sex schools
Pong	1996	Malaysia	Household head's earned income, Mother's education	Enrollment	Positive effects on enrollment

Table A.6: International Studies of the Relationship Between Family Socioeconomic Status and Educational Outcomes - Continue 5

Author(s) of Study	Year	Country	Measures of Family Socioeconomic Status	Outcome	Results
Tansel	1997	Cote D'Ivoire Ghana	Father's education, Mother's education, Total household expenditure	Enrollment, Attainment	Positive effects of father's and mother's education, Mother's ed. stronger for daughters in Ghana
Sarandy	1997	Iran, Tabriz	Parental occupation, Parental education	Academic success	Positive effects
Khaeir	1997	Iran, Shiraz	Father occupation, Mother education	Education achievement	Positive effects
Zhou, Moen & Tuma	1998	China	Father's education, Father's occupational status	Entry to levels of schooling	Positive effects on entry all levels
Wong	1998	Czechoslovakia	Father's education, household possessions	Attainment	Positive effects on attainment
Buchmann	2000	Kenya	Parents' education, household financial status	Enrollment	Positive effects on enrollment
Albert	2000	Spain	Education level of parents, especially of the mother, economic status of the father	Entering into university	Increases the chance of students
Jimenez & Velasko	2000	Spain	income, occupation, education level of the parents,	Demand for higher education	have opportunities of studying at better schools/universities
Gharun	2002	Iran	Parental education, Number of family members/Family size	Demand of higher education	Positive/Negative effects
Roshan & Salehi	2003	Iran, Tehran	Parental education, parental occupation, family income	Educational achievement	Positive/negative effects, without replaced by each other
Lefebvre & Merrigang	2010	Canada	Parental occupation, family income	Education achievement	Positive effects
Khodaei	2010	Iran	Parental education	Educational achievement	Positive effects on the children's success
Jamali	2012	Iran	Parental education, father occupation, family income	Educational performance	Positive effects