İSTANBUL TECHNICAL UNIVERSITY ★ INSTITUTE OF SCIENCE AND TECHNOLOGY

DEFINING CRITICAL FACTORS AFFECTING STUDENT SUCCESS: A DATA MINING APPROACH

M.Sc. Thesis by
İnci BÖLÜKBAŞI

Department : Industrial Engineering

Programme: Engineering Management

MAY 2005

İSTANBUL TECHNICAL UNIVERSITY ★ INSTITUTE OF SCIENCE AND TECHNOLOGY

DEFINING CRITICAL FACTORS AFFECTING STUDENT SUCCESS: A DATA MINING APPROACH

M.Sc. Thesis by
İnci BÖLÜKBAŞI
(507021202)

Date of submission : 30 June 2005

Date of defence examination: 30 May 2005

Supervisor (Chairman): Assis. Prof. Dr. Mehmet Mutlu YENİSEY

Members of the Examining Committee Assis. Prof. Dr. Cafer Erhan BOZDAĞ

Assoc. Prof.Dr. Demet BAYRAKTAR

MAY 2005

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

ÖĞRENCİ BAŞARISINI ETKİLEYEN KRİTİK FAKTÖRLERİN BELİRLENMESİ: BİR VERİ MADENCİLİĞİ YAKLAŞIMI

Yüksek Lisans Tezi
İnci BÖLÜKBAŞI
(507021202)

Tezin Enstitüye Verildiği Tarih: 30 Haziran 2005

Tezin Savunulduğu Tarih: 30 Mayıs 2005

Tez Danışmanı : Yard. Doç.Dr. Mehmet Mutlu YENİSEY

Diğer Jüri Üyeleri Yard. Doç.Dr. Cafer Erhan BOZDAĞ

Doç.Dr. Demet BAYRAKTAR

MAYIS 2005

ACKNOWLEDGEMENTS

I gratefully acknowledge my advisor Assis. Prof. Dr. Mehmet Mutlu Yenisey who has given the idea for the subject and invaluable comments throughout this thesis. Without his consideration, support and guidance, it would be difficult to complete the study.

I would like to thank Arif Onur Dil, Esra Bayram, Nebahat Dönmez, Ersoy Tezel, Barış Fındık, and Murat Tortopoğlu who gave valuable feedback and support during different phases of my preparation.

Finally, I am grateful to my family, especially my parents Hadiye and Metin Bölükbaşı for their endless support and encouragement through out my whole education life.

05.2005 İnci Bölükbaşı

TABLE OF CONTENTS

ABBREVIATIONS

GPA : Grade Point Average
İTÜ : Istanbul Technical University
AI : Artificial Intelligence
RDBMS : Relational Database Management Systems
OLAP : On Line Analytical Processing
IBM : International Business Machines
KDD : Knowledge Discovery in Databases
FBI : Federal Bureau Investigation
DM : Data Mining
CRM : Customer Relationship Management
CART : Classification and Regression Trees
CHAID : Chi-Squared Automatic Interaction Detection
AID : Automatic Interaction Detection
ID3 : Iterative Dichotomizer
SEM : SAS Enterprise Miner
SAT : Scholastic Aptitude Test
EQ-i:YV : Emotional Quotient Inventory: Youth Version
EI : Emotional Intelligence
PCCW : Palmer College of Chiropractic West

LIST OF TABLES

Page No

LIST OF FIGURES

Page No

LIST OF SYMBOLS

p_1, p_2	: Probability of r mutually exclusive events
A_k	: Any given attribute
$H(C/A_k)$	: Entropy of the classification property of Attribute A_k
$P(a_{k,j})$	: Probability of attribute k at value j
$P(c_i/a_{k,j})$	: Probability that the class value is c_i when attribute k is at its j^{th} value
M_k	: total number of values for attribute A_k; j = 1,2,…. M_k
N	: Total numer of different classes; i= 1,2,…N
K	: Total number of attributes; k= 1,2,..K
r	: Number of target levels
B	: Number of branches

DEFINING CRITICAL FACTORS AFFECTING STUDENT SUCCESS: A DATA MINING APPROACH

SUMMARY

In recent years, there has been increasing interest in students' academic achievement. Not surprisingly, an increasing amount of research has been carried out on the factors that influence academic achievement. The subject of this thesis is finding crucial factors that affect success of İTÜ's Management Faculty's students.

Data mining is one of the most popular techniques of finding valuable information through data. A different data mining application takes places in this thesis. The above problem is studied on by applying Decision Tree technique of Data Mining.

The whole study depends up on students' data. Accordingly, a student questionnaire is applied to sophomore, junior and senior students of Management Faculty's undergraduate students. As a result of this questionnaire the input attributes of the Data Mining study is selected. 424 different students' answers are used during the model development. A binary decision tree is generated based on these answers. 48 different attributes are used for the decision tree. 15 attributes are selected as effective on student success. The most important 10 attributes are examined in details.

The result of the pruned Decision Tree modeling has 18 different leaves. After the 18^{th} leaves, the overall correctness ratio of the tree is 0.8208. The most effective factors of success are defines as: Attendance Percentage, Gender, Residence, Most Liked Course Group, Finding Job Duration, Birth Region, Class, Scholarship, Father Education, Studying Hours, General Pleasure Of ITU, Master Degree, Father Alive, Cinema Frequency and Average Monthly Expenditure.

These factors are important according to the model. But it is not possible to say clearly which values of these factors have a positive impact on success. All these variables must be investigated one by one on the decision tree.

ÖĞRENCİ BAŞARISINI ETKİLEYEN KRİTİK FAKTÖRLERİN BELİRLENMESİ : BİR VERİ MADENCİLİĞİ YAKLAŞIMI

ÖZET

Son yıllarda, öğrencilerin akademik başarılarını araştırmaya yönelik artan bir ilgi var. Sonuç olarak, akademik başarıyı etkileyen faktörler üzerinde artan sayıda çalışma sürdürülüyor. Bu çalışmanın konusu, İTÜ İşletme Fakültesi öğrencilerinin başarısını etkileyen önemli faktörleri, bir Veri Madenciliği uygulması yöntemiyle bulmaktır.

Veri Madenciliği kaynak data içinde değerli bilgileri elde etmek için kullanılan en etkili yöntemlerden biridir. Bu tezde farklı bir veri madenciliği çalışması yer alır. Yukarıda açıklanan amaç, Karar Ağaçları tekniği uygulaması ile modellenir.

Tüm çalışmanın kaynak datası öğrenci verileridir. Öğrenci verilerini toplamak amacı ile İşletme Fakültesi 2., 3. ve 4. sınıf lisans öğrencilerinin katıldığı bir anket çalışması yapılır. Bu anketin sonucu olarak Veri Madenciliği çalışmasının girdi değişkenleri seçilir. Model geliştirilirken 424 farklı öğrencinin cevapları kullanılır.

Ankete verilen cevaplara bağlı olarak ikili karar ağacı geliştirilir. Karar ağacı modelinde girdi olarak 48 farklı değişken kullanılır. 15 değişken öğrenci başarısı üzerinde etkili olarak bulunur. En etkili ilk 10 değişken detayları ile ele alınır.

Budanan karar ağacının 18 farklı yaprağı vardır. 18.ci yapraktan sonra ağacın genel doğruluk oranı 0.8208 olarak hesaplanır. Başarı üzerinde en etkili faktörler, etki sırasına göre şunlardır: Okula Devam Yüzdesi, Cinsiyet, İkamet Etme Şekli, En Sevilen Ders Grubu, İş Bulma Süresi, Doğum Bölgesi, Sınıfı, Burs Sahipliği, Baba Eğitim Durumu, Günlük Ders Çalışma Saat Sayısı, İTÜ'de Olmaktan Memnuniyet Derecesi, Master Yapma İsteği, Baba Hayatta Olma Durumu, Sinemaya Gidiş Frekansı ve Aylık Ortalama Para Harcama Miktarı

Kurulan modele göre bu faktörler etkilidirler. Fakat, kesin olarak hangi faktörlerin başarı üzerinde olumlu bir etkisi olduğunu söylemek doğru değildir. Karar ağacı üzerinde her bir değişkenin tek tek incelenmesi ve yorumlanması gerekir. Bazı değişkenler tek başına etkili olurlen, bazı değişkenler bir arada olumlu ya da olumsuz bir etki yaratmaktadır.

1. INTRODUCTION

Throughout the world, college students are becoming increasingly diverse in their abilities, backgrounds, aspirations, and beliefs about college success. All of these factors can lead to differences in academic achievement. However, the most commonly used predictors of college grade point average (GPA) are standardized tests. Like all tests, these tests are imperfect. Hence, they are incomplete predictors of how well a student will do in college coursework; they are necessary but not sufficient markers of progress and prognosis. [1]

In recent years, there has been increasing interest in students' academic achievement, particularly low achievement, in a variety of countries. Recent societal developments, such as moving towards a wider use of high technology and information technology, have increased this interest in terms of ensuring the availability of highly competent employees in the future. Not surprisingly, an increasing amount of research has been carried out on the factors that influence academic achievement, as well as different ways to promote high achievement and optimal skill development in a large number of students. [2]

İstanbul Technical University is one of the leading universities in Turkey. Like every other universities its first duty is to educate and raise generations and make them productive. However, to obtain better results, it is a need to learn about them. Universities should collect data about their students. They should use technology to form databases and use these data to find valuable information about their student. There are too many ways of using data for decision making processes. The first step is to be customer centric and assume all participants as customers that are students, their families and the society as a whole. To be successful, it is a prerequisite to understand all the aspects of your customers beginning from who your customers are. The key is to know deeply who your organization is dealing with. According to this, organizations realized that they have to achieve several objectives. One of the objectives is to get closer to the customer by utilizing the data hidden in scattered

enterprise databases. Examining and analyzing the data can turn raw data into valuable information about customer's needs [3].

For the success of the students, the university is responsible to develop good student relationship management. The key in this issue is to understand who the students are, how they study, what they like, how they live. After these, the university should find out the critical success drivers of students and advise students according to their individual situation. In other words, several performance measures need to be determined and analyzed to see the current situation and make predictions about future.

In this study, the important objective that is targeted is to find out the success drivers of İTÜ Management Faculty students. In the achievement of this target, first of all, the data which are the source of the model have to be provided. Data Collection is fulfilled by applying a survey to sophomore, junior and senior students. Freshmen are not asked to answer the questionnaire, because, the constraint of one year fulfillment in İTÜ is not valid for them. There are various questions in the survey, and one of them is student's recent Cumulative Grade Point Average. It is just possible, if and only if a student passes one year in the university. Accordingly, the survey's target set considers, the students, who have a Cumulative Grade Point Average, in other words, who have been at least one year in İTÜ. After a hard process of questionnaire practice, 438 questionnaires, which are answered, are gathered. The details of the questionnaire will be explained widely in 4^{th} chapter which is Student Questionnaire. All the answers of the questionnaires are entered in an excel file. After the answers are collected in excel file, new variables are formed using others. At last, over 80 numbers of variables are reached.

In the study, SAS data mining software is used to accomplish the objectives of the thesis. Firstly, the excel file is imported in to SAS Enterprise Guide, and the statistical analysis is accomplished. Secondly, to find out the success drivers of the students, a predictive data mining technique, Decision Tree, is applied. According to the students' recent GPA scores, successful and unsuccessful student sets are formed. In the data mining application the aim is to build the tree that is giving the paths to successful and unsuccessful students. The class variable of the study is the student success situation.

2. GLANCE AT DATAMINING

Data mining is treated by many people as more of a philosophy, or a subgroup of mathematics, rather than a practical solution to business problems. It can be seen by the variety of definitions that are used, for example: Data mining is the exploration and analysis of very large data with automatically or semi-automatically procedures for previously unknown, interesting, and comprehensive dependencies.

For the most part, that definition stands up pretty well. The emphasis on large quantities of data is important since the data volumes continue to increase. But there is something to regret; it is the phrase "by automatic or semi-automatic means". Not because it is untrue – without automation it would be impossible to mine the huge quantities of data being generated today – but because there has come to be too much focus on the automatic techniques and not enough on the exploration and analysis. This has misled many people into believing that data mining is a product that can be bought rather than a discipline that must be mastered [4].

Data mining is a discovery process, in that one can uncover information that would typically not found without data mining. Data mining has no fixed presentation of data and allows the user to create inquiries based on the information required.

2.1 Data Warehouse, Data Mart and Data Mining

Data warehousing is the enabling technology for data mining. But its uses go far beyond data mining. Another relevant term is data mart, a variant on the data warehouse.

A data warehouse is an environment not a single technology, comprising a data store and multiple software products often obtained from different vendors. The products include tools for data extraction, loading, storage, access, query and reporting. The data store is a collection of subject oriented, integrated, time variant, nonvolatile data

that is required to support management decision making. Data stores include immense volumes of information detailing every aspect of a particular subject. Data warehouses are typically assembled to support decision oriented management queries. Because frequent and/or sophisticated queries can slow operational systems data warehouses are established and maintained apart from the production systems.

Data mart is a collection of data and tools focused on a specific business unit or problem. Size does not distinguish data marts but they tend to be smaller than data warehouses.

Data mining is a collection of tools and techniques used for inductive rather than deductive analyses. Using sophisticated data mining tools, analysts explore detailed data and business transactions to uncover meaningful insights, relationships, trends, or patterns within the business activity or history. Data mining is used to identify hypothesis; traditional queries are used to test hypothesis. [5]

2.2 History of Data Mining

Recently, data mining has been the subject of many articles in business and software magazines. However, just a few short years ago, few people had even heard of the term data mining. Though data mining is the evolution of a field with a long history, the term itself was only introduced relatively recently, in the 1990s. This section explores the history of data mining.

Data mining roots are traced back along three family lines. The longest of these three lines is classical statistics. Without statistics, there would be no data mining, as statistics are the foundation of most technologies on which data mining is built. Classical statistics embrace concepts such as regression analysis, standard distribution, standard deviation, standard variance, discriminate analysis, cluster analysis, and confidence intervals, all of which are used to study data and data relationships. These are the very building blocks with which more advanced statistical analyses are underpinned. Certainly, within the heart of today's data mining tools and techniques, classical statistical analysis plays a significant role.

Data mining's second longest family line is artificial intelligence, or AI. This discipline, which is built upon heuristics as opposed to statistics, attempts to apply human-thought-like processing to statistical problems. Because this approach

requires vast computer processing power, it was not practical until the early 1980s, when computers began to offer useful power at reasonable prices. AI found a few applications at the very high end scientific/government markets, but the required supercomputers of the era priced AI out of the reach of virtually everyone else. The notable exceptions were certain AI concepts which were adopted by some high-end commercial products, such as query optimization modules for Relational Database Management Systems (RDBMS).

The third family line of data mining is machine learning, which is more accurately described as the union of statistics and AI. While AI was not a commercial success, its techniques were largely co-opted by machine learning. Machine learning which is able to take advantage of the ever-improving price/performance ratios offered by computers of the 80s and 90s, found more applications because the entry price was lower than AI. Machine learning could be considered an evolution of AI, because it blends AI heuristics with advanced statistical analysis. Machine learning attempts to let computer programs learn about the data they study, such that programs make different decisions based on the qualities of the studied data, using statistics for fundamental concepts, and adding more advanced AI heuristics and algorithms to achieve its goals.

Data mining, in many ways, is fundamentally the adaptation of machine learning techniques to business applications. Data mining is best described as the union of historical and recent developments in statistics, AI, and machine learning. These techniques are then used together to study data and find previously-hidden trends or patterns within. Data mining is finding increasing acceptance in science and business areas which need to analyze large amounts of data to discover relations which they could not otherwise find. [6, 7]

Evolution of business data to business information starts with data collection in 1960's. Data collection provides answers to retrospective questions. In other words, it addresses the questions related with past. In 1980's, with the relational databases, data access methods improved dramatically. In 1990's, data warehousing and decision support systems are founded with multidimensional databases, OLAP. Today, data mining is a field which utilizes advanced algorithms, multi processor computers and massive databases. The basic difference between data mining and

previous fields is that data mining aims to answer prospective questions about future. The Table 2.1 describes evolutionary steps of data mining [8].

Table 2.1 Evolution of data mining

Year	Evolutionary Step	Enabling Technology
1960's	Data collection and database creation	computers, tapes, disks
1970's	Relational data model	faster and cheaper computers
1980's	RDBMS, advanced data models	faster and cheaper computers with more storage, On-line analytical processing (OLAP),multidimensional databases, data warehouses
1990's	Data warehousing and data mining	faster and cheaper computers with more storage, advanced computer algorithms

2.3 Data Mining Approach

Organizations are accumulating vast quantities of data in databases, with the recent trend to implement a data warehouse architecture increasing the quality and accessibility of data. This is all being done at great cost, but the information is only valuable if used effectively.

Users have been using query tools, OLAP servers, Business Intelligence tools, Enterprise Information Systems and a wide range of other packaged software to analyze their data. However, the more numerate analysts have recognized that there are hidden patterns, relationships and rules in their data which cannot be found by using these traditional methods.

The answer is to use specialist 'data mining' software which harnesses advanced mathematical algorithms to examine large volumes of detailed data. Data mining is the process of extracting valid, previously unknown and ultimately comprehensible information from large databases and using it to make critical business decisions. The software is able to sift large volumes of data to find nuggets of information which yield gold in the form of competitive advantage.

Data mining can be carried out on any data file, from a spreadsheet to a data warehouse. Transaction processing systems can be mined to generate benefits which can help to justify implementing data warehouse architecture.

Data mining is very different from querying, where the user knows what is in the database and knows what information to ask for. Data mining with a query tool is like mining coal with a spoon. [9]

Data mining is about discovering new things about business from the data that have been collected. Using standard statistical techniques to explore database does not discover new outcomes. In reality what is being done is making a hypothesis about the business issue that is addressed and then attempting to prove or disprove the hypothesis by looking for data to support or contradict the hypothesis.

Data mining uses an alternative approach beginning with the premise that no patterns of data are known by the user. In this case, user does not have to develop a hypothesis, and just simply ask, what is new, interesting and valuable in the data. In this case, data mining algorithm tells about all the previously unknown and ultimately comprehensible information that users had. Data mining therefore provides answers without users having to ask specific questions.

The difference between the two approaches is summarized in Figure 2.1. [3]

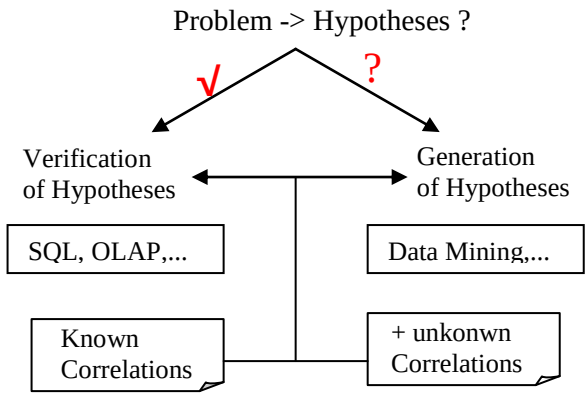


Figure 2.1 Standard and data mining approaches on information detection

Data mining has two different goals: verification of a user's hypothesis, and discovery, the finding of new patterns in data. Discovery (sometimes referred to as knowledge discovery in databases, or KDD) includes prediction, (regression and classification) and description (summarization, visualization, and detection of changes and deviations). Some KDD tools are generic; others are domain specific. Domain specific tools represent an important trend, moving knowledge discovering technology directly into the hands of business users. Among the elements that make this possible are putting the problem in the business user's terms, providing support

for specific key business analyses, representing results in a form geared to the business problem being solved, and providing support for an exploratory process.[10]

2.4 Application Areas

The goal of data mining is to produce new knowledge that decision makers can act upon. It does this by using sophisticated techniques such as artificial intelligence to build a model of the real world based on data collected from a variety of sources. Data mining is used to understand customer behavior, predict buying patterns, eliminate inappropriate surgical procedures, detect fraud, and other innovative applications. It has helped organizations to provide more meaningful services to customers, increase revenue, reduce expenses.

Customer churn and loyalty are important issues for most of the companies. Data mining can predict which customers are likely to leave the company and go to a competitor. Using this information, marketing strategies can be developed for customer retention as a part of Customer Relationship Management.

Fraud detection is widely used in credit card services and telecommunications. Data mining uses historical data to build models of fraudulent behavior and use techniques to identify which transactions are most likely to be fraudulent in future.

Credit scoring is mainly concerned whether to extend credit to an applicant or not. The aim is to anticipate and reduce defaults and serious delinquencies. Other concerns in credit risk management are the maintenance of existent credit lines and determining the best action to be taken on delinquent accounts. [11]

Below are some examples of the many uses of data mining:

Pharmaceutical

- Discovering new uses for existing drugs

Healthcare - Insurance

- Utilization analysis of hospitals
- Identifying behavior pattern of risky customers
- Improving preventive care

- Predicting which customers will buy new policies

Banking

- Understanding customer behavior
- Improving cross selling efforts
- Identifying loyal customers
- Predicting for credit card customer attrition
- Detecting patterns of fraudulent credit card usage

Telecommunications

- Identifying products and services that maximize life time value of customers
- Establishing marketing campaigns to improve market share
- Churn analysis [11]

Internet Applications

- Identifying the user interest patterns
- Determining most probable customers who will perform e-commerce [12]

Governmental

Many of the nation's commercial establishments have long ransacked commercial databases hoping to determine which consumers are likeliest to buy a particular luxury car or life insurance policy. Now FBI hopes to sift though the same databases to identify threats before they occur. Such databases often correlate an individual, and his various IDs, with his attributes such as smoking behavior, clothing size, any arrest records, household income, magazine subscriptions, height, weight, chronic health conditions, and many other characteristics. People concerned with privacy worry about that FBI's use of such databases could create false alarms.

Terrorists are consumers too, because terrorists buy products, rent apartments, and use credit cards. The FBI hopes that DM analysis of the data will reveal patterns that could help prevent future attacks. [13]

2.5 Data Mining Styles

Traditionally the task of identifying and utilizing information hidden in data has been achieved through some form of statistical methods. Typically users formulate a guess about a possible relationship in the data and evaluate this hypothesis via a statistical test. This a largely time-intensive, user-driven, top-down approach to data analysis. With data mining, the interrogation of the data is done by the data mining algorithm rather than by the user. [14]

The essence of data mining is to analyze data to generate descriptive or predictive models to solve problems. The goal of a descriptive model is to discover patterns in the data and to understand the relationship between attributes represented by the data. The goal of a predictive model is to predict future outcomes based on past records with known answers. There are two styles of data mining used to generate models: supervised and unsupervised. Berry and Linoff [4] refer to them as directed and directed.

2.5.1 Supervised (Directed) Modeling

Supervised or directed modeling is a top-down approach, used when it is known what is being looked for. This often takes the form of predictive modeling, where we know exactly what we want to predict. It is goal oriented. The task is to explain the value of some particular field. The user selects the target field and directs the computer to tell how to estimate, classify, or predict it. Supervised learning approach is very common in the business world and predictive modeling is an example of this. For instance, if the churn rate for the next month is to be predicted, then purpose of the mining process is known. In other words, for supervised learning, independent variables are fed into the model and the dependent variable is predicted. The prediction is compared with the actual dependent value to assess the validity of the model.

2.5.2 Unsupervised (Undirected) Modeling

Unsupervised or undirected modeling is a bottom-up approach, that lets the data speak for itself. It finds patterns in the data and leaves it up to the user to determine whether or not these patterns are important. There is no target field. The user asks the computer to identify patterns in the data that may be significant. Unsupervised

learning finds patterns and profiles in the data, and leaves the interpretation to the user. Clustering technique is an example of unsupervised learning.

In other words, unsupervised modeling is used to recognize relationship in the data, and supervised modeling is used to explain those relationships once they have been found. [4, 14]

2.6 Data Mining Techniques

A variety of techniques have been developed over the years to explore and extract information from large data sets. When the name data mining was coined, many of these techniques were simply grouped together under this general heading [3]. Data mining tools use machine learning or statistical modeling techniques such as decision trees, neural networks, clustering and regression analysis. Many different algorithms are used for implementing these techniques.

It is known that technical understanding is actually one small component of mastering data mining. Without at least a high-level understanding of the most important data mining algorithms, you will not be able to understand when one technique is called for and when another would be more suitable. It is also needed to understand what is going on inside a model in order to understand how best to prepare the model set used to build it and how to use various model parameters to improve results.

Fortunately, the level of understanding needed to make good use of data mining algorithms does not require detailed study of machine learning or statistics. Just as it is possible to take wonderful photographs without understanding exactly how exposure to light causes the photosensitive chemicals in the emulsion to change color, it is quite possible to make good use of data mining tools without understanding all the details of how they work. By the same token, just as some understanding of the photographic process is essential to making good choices of film speed, aperture, and length of exposure, so a basic understanding of the principle data mining algorithms is essential for anyone wishing to master the art of data mining. [4]

2.6.1 Supervised (Predictive) Modeling Techniques

Predictive data mining is applied to a range of techniques that find relationships between a specific variable, called the target variable and the other variables in the data.

If the data element under investigation is discrete meaning that it has a small number of fixed values, the task is called classification. On the other hand if the data element is continuous, meaning that it can take a large number of values and exhibits a common unit of measure, the task is called regression.

Classification is a learning method frequently adopted in the fields of data mining, statistics, machine learning, genetic algorithm and neural networks [15]. The classification problem is a two-step process, where the first is to build a classification model by analyzing the training sample set described by attributes and the second is to use this model to classify the future sample for which the class label is not known. For example, classification model can be used that is learned from the existing customers' data to predict what services a new customer would like. [16]

Classification predicts class or group membership. With classification the predicted output, the class, is categorical. A categorical variable has only a few possible values, such as yes-no, high-middle-low, etc. Regression predicts a specific value. Regression is used in cases where the predicted output can take on many possible values, and the output is therefore continuous.

Techniques that dominate the commercially available classification and regression tools today are, decision trees, neural networks, naïve bayes, K–nearest neighbor.

Decision Trees

Decision trees are a way of representing a series of rules that lead to a class or value. The graphic output is similar in structure to a tree. A decision tree consists of the decision node, branches, and leaves. The first component is the decision node, which specifies a test to be carried out. Each branch of a node leads either to another decision node or stopping point, called a leaf node. By navigating the decision tree, one can assign a value or class to a case by deciding which branch to take, starting at the top node and moving to each subsequent node until a leaf node is reached. Each node uses the data from the case to choose the appropriate branch.

Neural Networks

Neural networks are more complicated than other techniques. Based on an early model of human brain function, it mimics the brain's ability to learn from its mistakes. It is often referred to as a "black box" technology and involves very careful data cleansing, selection, preparation, and preprocessing. A neural network starts with an input layer, where each node corresponds to a predictor variable. These input nodes are connected to a number of nodes in a hidden layer. Each input node is connected to every node in the hidden layer. The nodes in the hidden layer may be connected to nodes in another hidden layer, or to an output layer. The output layer consists of one or more response variables. Neural networks are often used in financial markets for prediction and forecasting.

Naïve Bayes

Naïve bayes analyzes the relationship between each independent variable (the predictor) and the dependent variable (the predictee) to derive a conditional probability for each relationship. When a new case is analyzed, a prediction is made by combining the effects of the independent variables on the dependent variable. Naïve bays require only one pass through the data to generate a classification model. This makes it a very efficient data mining technique. However it does not handle continuous data, limiting inputs only to categorical data.

K-Nearest Neighbor

K-nearest neighbor is a technique that classifies each record in a data set based on a combination of the classes of the K records most similar to is in a historical data set. It has no distinct training (model building) phase because the training data is actually the model.

2.6.2 Unsupervised (Descriptive) Modeling Technique

Descriptive data mining is applied to a range of techniques which find patterns inside the data without any prior knowledge of what patterns exist. There are two common methods for unsupervised modeling, association and clustering.

Association

Association is used to determine which things go together. A typical application that can be built using an association function is market basket analysis. It finds affinity

groupings that discover what items are usually purchased with others predicting the frequency with which certain items are purchased at the same time.

Clustering

Clustering is task of segmenting a diverse group into a number of similar subgroups or clusters. In clustering, there are no predefined classes and no examples. The records are grouped together on the basis of self-similarity. It is up to the miner to determine what meaning, if any, to attach to the resulting clusters. Clustering is often used to prepare data for another step in analysis. Some of the common algorithms used to perform clustering include Kohonen feature maps and K-means [4, 10, and 17].

3. A GENERIC DATAMINING METHOD

A data mining application should be applied according to a flow. The flow is based on data mining logic. The generic data mining method comprises of seven steps.

- Defining the business issue in a precise statement
- Defining the data model and the data requirements
- Sourcing data from all available repositories and preparing the data (the data could be relational or in flat files, stored in a data warehouse, computed, created on-site or bought from another party. They should be selected and filtered from redundant information).
- Evaluating the data quality
- Choosing the mining function and defining the mining run
- Interpreting the results and detecting new information
- Deploying the results and the new knowledge into business

These steps are illustrated in Figure 3.1

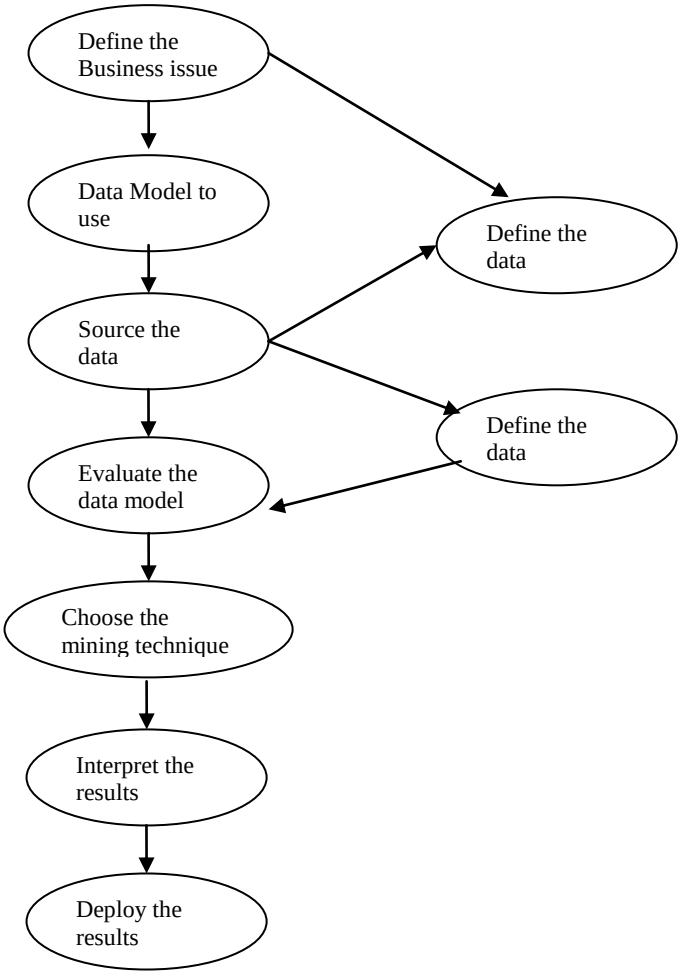


Figure 3.1The generic method

3.1 Defining the Business Issue

Often, organizations approach data mining from the perspective that there must be some value in the data have been collected, so data mining is just used to discover what is there. Using the mining analogy, this is rather like choosing a spot at random and starting to dig for gold.

Data mining is about choosing the right tools for the job and then using them skillfully to discover the information in the data. There are a number of tools that can be used, and sometimes combinations of the tools are used, to make real discoveries and extract the value from the data.

The first step in data mining method is therefore to identify the business issue that is desired to address and then determine how the business issue can be translated into a question, or set of questions, that data mining can solve.

By business issue it is meant that there is an identified problem that needs an answer, where it is suspected, or known that the answer is buried somewhere in the data, but it is not certain where it is. A business issue should fulfill the requirements of having:

- A clear description of the problem to be addressed
- An understanding of the data that might be relevant
- A vision for how you are going use the mining results in your business

3.2 Defining a Data Model To Use

The second step in the generic data mining method is to define the data to be used. A business can collect and store vast amounts of data. Usually the data is being stored to support various applications, the best way to do this is to use some form of data warehouse. Although not all data warehouse architectures are the same, one way they can be used efficiently to support applications is shown in Figure 3.2.

In this case each end user application is supported by its own data mart which is updated at regular intervals or when specific data changes, to reflect the needs of the application.

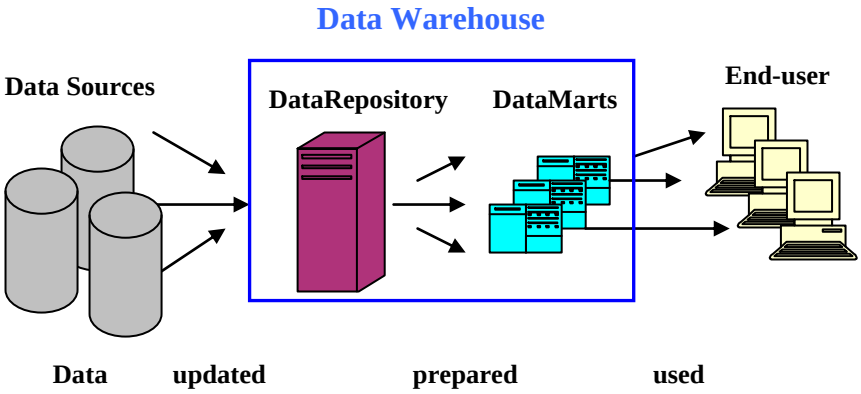


Figure 3.2 Data warehouse architecture

In this structure each data mart has its own specific data and holds knowledge about how the data was derived, the data format used, what aggregations have been performed, what data cleansing has been done and so on. Where the data is being used routinely to support a specific business application, the data and metadata together form what is called a data model that supports the application. [18]

3.3 Sourcing and Preprocessing the Data

The third step in the generic data mining method is the sourcing and preprocessing of the data that populates the data model. Having a defined data model provides the necessary structure, in terms of the variables that are going to be mined, but the data is still have to be provided.

Data sourcing and preprocessing comprises the stages of identifying, collecting, filtering and aggregating (raw) data into a format required by the data models and the selected mining functions. Sourcing and preparing the data are the most time consuming parts of any data mining project. Where the data is derived from a data warehouse, many of these stages will already have been performed.

The data sources can be different by origin and content as shown in Figure 3.3.

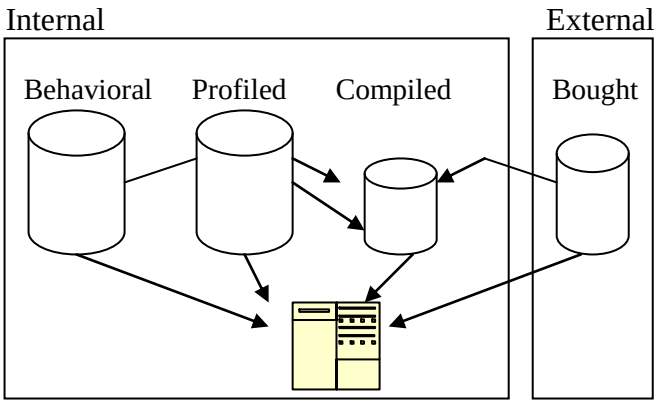


Figure 3.3 Data sources by origin and content

Every business uses standard internal data sources. Many of them are similar from their point of content. Therefore, a customer database or a product database could be found in nearly any data scenario.

Data mining, in common with many other analysis tools, usually requires the data to be in one consolidated table or file. If the variables required are distributed across a number of sources then this consolidation must be performed such that a consistent set of data records is produced. If the data is stored in relational database tables then the creation of a new tables or a database view is relatively straight forward, although where complex aggregations are required this can have a significant impact on database resources.

If the data has not been derived from a data warehouse then the data preprocessing functions of cleansing, aggregated, transforming, and filtering must be undertaken. Even when the data is derived from a data warehouse, there may still be some additional transformations of the data that need to be performed before mining can proceed. Structurally the data preprocessing

can be displayed as in Figure 3.4.

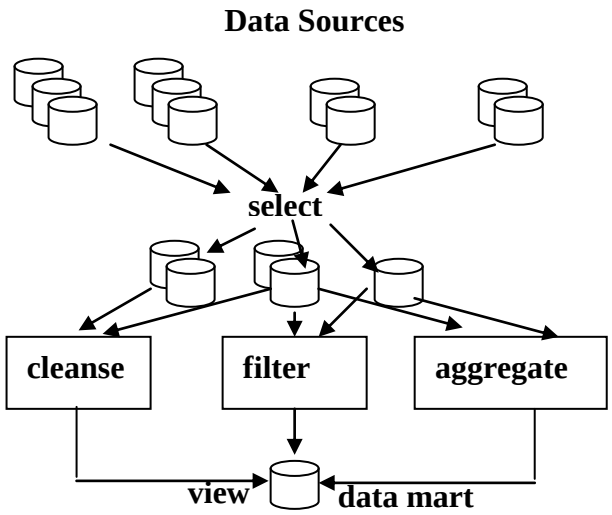


Figure 3.4 Data preprocessing

Data mining tools usually provide limited capability to cleanse the data, because this is a specialized process and there are a number of products that can be used to do this efficiently. Aggregation and filtering can be performed in a number of different ways depending on the precise structure of data sources.

3.4 Evaluating the data model

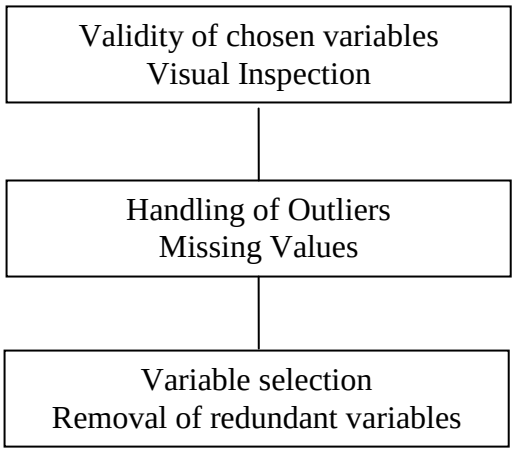


Figure 3.5 Steps of evaluation

Having populated the data model with data, it is still have to be ensured, that the data used to populate model fulfills the requirement of completeness, exactness and relevance. To do this the fourth step in the generic data mining method, which is to perform an initial evaluation, takes place. Its steps are described in Figure 3.5.

The first step visual inspection comprises browsing the input data with visualizing tools. This may lead to the detection of implausible distributions. For example, a wrong join of tables during the data preparation step could result in variables containing values actually belonging to different fields.

The second step deals with the identification of inconsistencies and the resolution of errors. Unusual distributions found within the first step could be induced by a badly done data collection. Many outlying and missing values produce biased results. For example, the interesting information of a correlation between variables indicating the behavior of a customer group and variables describing new offerings could be completely suppressed if too many missing values are accepted. The mitigation of outliers and the transformation of missing values in meaningful information however improve the data quality enormously.

The last step is the final selection of features/variables for the mining run. Variables could be superfluous by presenting the same or very similar information as others,

but increasing the run time. Dependent or highly correlated variables could be found with statistical tests like bivariate statistics, linear and polynomial regression. Dependent variables should be reduced by selecting one variable for all others or by composing a new variable for all correlated ones by factor or component analysis.

Not all variables remaining after the statistical check are nominated as input; only variables with a clear interpretation and variables that make sense for the end user should be selected. A proven data model simplifies this step. The selection of variables in that stage can indeed only be undertaken with practical experience in the respective business or research area.

3.5 Choosing the data mining technique

Besides the steps defining business issues, data modeling, and preparation, data mining also comprises the crucial step of the selection of the best suited mining technique for a given business issue. This is the fifth step in the generic data mining method. This step not only includes defining the appropriate technique or mix of techniques to use, but also the way in which the techniques should be applied. Several types of techniques (or algorithms) are available.

The selection of the method to use will often be obvious, for example, market basket analysis in retail will use the association technique which was originally developed for this purpose. However, the association technique could be applied to a diverse range of applications, for example, to discover the relationships between faults occurring in production runs and the sources from which the components were derived. The challenge is usually not which technique to use but the way in which the technique should applied. Because all of the data mining techniques require some form of parameter selection, this then requires experience of how the techniques work and what the various parameters do.

3.6 Interpreting the results

Interpreting the results is the sixth step in the generic mining method. The results from performing any type of data mining can provide a wealth of information that can sometimes be difficult to interpret. The interpretation phase requires the input from a business expert who can translate the mining results back into the business

context. It is important that the results are presented in such a way that they are relatively easy to interpret.

To assist in the interpretation process, it is necessary to have a range of tools that enable you to visualize the results and to provide the necessary statistical information that you need to make the interpretation.

While interpreting the results also the validity result of the data mining model is checked. While evaluating the model, there are different methods to be used. The common, model testing method is Simple Validation. In this method, typically the whole data set is divided in to test data and train data. Between 5 and 33 percent of the whole set is selected as test data. The rest of the data is used to train the model. After the model is trained, the same model is applied on the test data. For example, for a classification model, the ratio of the number of events that are wrongly classified and the total number of events give the error percentage of the model. According to this percentage, it is decided if the model should be improved or already a valid one. [19]

The total number of events in this study is 438. This is a very restricted data set to form a train and a test data set. Actually, in classification models it is generally recommended to perform the model on the train set then use the test data set. To divide the data set and try to test it, will not give a good result in the model. So, in this study all the data that are provided as input to the model will be used just to develop the model.

3.7 Deploying the results

The seventh and final step in the generic data mining method is perhaps the most important of all. It deals with the question of how to deploy the results of the data mining into business. If data mining is seen as an analytical tool, the full potential of what data mining has to offer is failed to realize.

When data mining is performed both discovering new things about customers and determining how to classify or how to predict particular characteristics or attributes are achieved. In all these cases data mining creates mathematical representations of the data that are called models. These models are very important, because they not

only provide with a deeper insight of business but can themselves be deployed in or used by other business processes, for example, CRM systems.

3.8 Efforts required

It is difficult to be prescriptive about the amount of effort that will be required to perform a typical data mining project. If starting from a position of having no data warehouse and with data in disparate files and databases then the type of effort profile required is shown in Figure 3.6.

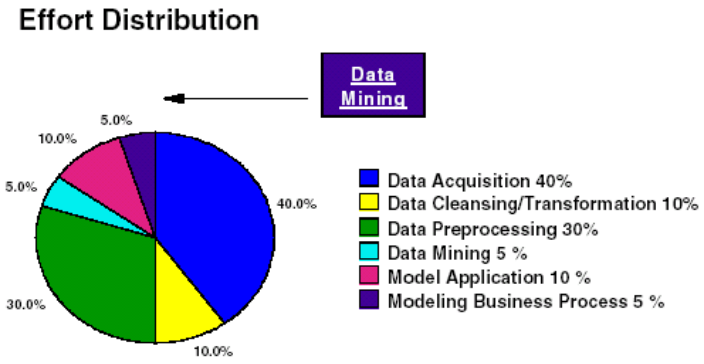


Figure 3.6 Effort Distribution

It is seen that the vast majority of the effort is spent in the data preparation activities. If the data is in a usable form, then the task of mining business becomes much easier [18].

4. STUDENT QUESTIONNARIE

4.1 Objective

The aim of the study is to build a model that defines the critical success factors of İTÜ Management Faculty students. It is obvious that, the source of the study is the students, themselves. The model must be built up by using the data about students. So, many different categories of data are needed. The only solution is to apply a comprehensive questionnaire and collect all the necessary information about students. The data preparation step of this data mining study is done by collecting the students' questionnaires' answers. The applied original questionnaire should be reviewed at Appendix 3.

This questionnaire is applied, during the final exams of the 2004-2005 education year's first term. The sophomore, junior and senior students of Industrial and Management Engineering students are involved in the survey. 438 of the students gave proper answers to the questionnaire in order to be used during the data mining model studies. This data mining study will be based on these 438 different answers. Each answered questionnaire will form a separate record.

In the questionnaire there are four different categories of questions. These are, demographic category, educational category that searches the background and the success of the student before university, academic category that searches the adaptation and the success of the student during university education, and Personal category that asks for lifestyle of the student during university. Below, the questions, the answer choices and the distribution of the answers are explained in details.

Before staring the modeling studies, the data have to be evaluated. Data Evaluation is explained detaily in section 3.4. The objective of this chapter is to apply data evaluation of each variable that takes part in the questionnaire. In means of data evaluation the followings are performed; validity of chosen variables by visual inspection, handling of outliers values and variable selection and removal of redundant variables.

All the answers of the questionnaire are aggregated in an excel sheet. In excel the distinct values of each column is investigated. Hence, the validity of the variables is guaranteed by visual inspection.

That excel sheet is taken in to SAS Enterprise Guide tool. This tool is used to perform the table analysis, one way frequency analysis and bivariate analysis. The distributions of each variable are tabulated. If there are outlier values in the distribution the variable is discretized. Discretization is used to handle the outlier variables. Since outlier values are not proper to take part in the modeling unless discretization is applied, it is not useful to have these variables in the model. In this chapter, each variable are analyzed, and discretization is applied when it is necessary.

According to the distribution analysis, the redundant variables are selected. They are omitted from the modeling study. These variables are set as rejected in the data mining tool.

At the end of this chapter, potential variables would have been selected as input variables for the Data Mining modeling studies.

4.2 Demographic Questions

Demographic attributes are always very important attributes for data mining studies. Most of the companies and institutions use these kinds of attributes. It is tried to find out the success factor of the students. One of the most probable factors are the students' family and lifestyle at home. Besides it is believed that students' success mostly depends on family attributes. The demographic attributes that take place in the questionnaire are as follows.

Birth Date of the Students

One of the important demographic variables is the ages of the students. Their birth date is asked to calculate their ages. According to the answers:

Table 4.1 Age Distribution

Age	Frequency	Percent	Cumulative Frequency	Cumulative Percent
19	2	0.46	2	0.46
20	55	12.56	57	13.01

21	120	27.40	177	40.41
22	140	31.96	317	72.37
23	100	22.83	417	95.21
24	15	3.42	432	98.63
25	5	1.14	437	99.77
26	1	0.23	438	100.00

Table 4.2 Age Summary Statistics

Label	Maximum	Mean	Minimum	N
Age	26	21,8013	19	438

Gender of the Students

Since İTÜ is a technical university, and the major is engineering, so most of the students are male. The gender type will be examined if it has an effect on the success of the students or not.

Table 4.3 Gender Distribution

Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	147	33.56	147	33.56
Male	291	66.44	438	100.00

Marital Status

The marital statuses of the students are asked. The students were expected to be single. As it was expected, only two students are married, in the set of students that answered the questions.

Table 4.4 Marital Status Distribution

Marital Status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Married	2	0.46	2	0.46
Single	436	99.54	438	100.00

This variable will not be used during the data mining process. Because, this attribute is not giving any distinguishing information about the students. If this attribute is used as an active variable, then the model will be mistaken.

Child Ownership

The students were expected not to have a child. As it was expected, only one student has a child in the set of students that answered the questions.

Table 4.5 Child Ownership Distribution

Child_Ownership	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	437	99.77	437	99.77
Yes	1	0.23	438	100.00

Like Marital Status variable, this variable will not be used during the data mining process, either. Because, this attribute is not giving any distinguishing information about the students. If this attribute is used as an active variable, then the model will be mistaken.

Mother alive

In the questionnaire, it is asked if their mother is alive or not. Actually the answers are surprising. According to the answers all of the students' mothers are alive. By the way, Mother Alive attribute, does not give any valuable information about the students success so it will not be used in the data mining studies.

Father Alive

In the questionnaire, it is asked if their father is alive or not.

Table 4.6 Father Alive Distribution

Father_Alive	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	38	8.68	38	8.68
Yes	400	91.32	438	100.00

8,68 % of the students population's fathers are not alive. So this variable can affect the success of the students. This variable will be used during the data mining studies but the model will decide, if it is an effective variable or not.

Birth Province

Turkey has seven different geographical regions, and it is anticipated that the birth region should have an effect on the students' success. So, Birth Place Traffic code is asked in the questionnaire. Using the traffic codes, the birth regions are formed for all students.

Table 4.7 Birth Region Distribution

Birth_Region	Frequency	Percent	Cumulative Frequency	Cumulative Percent
AKDENIZ	46	10.50	46	10.50
DAB	24	5.48	70	15.98
EGE	65	14.84	135	30.82
GDA	12	2.74	147	33.56
KARADENIZ	50	11.42	197	44.98
MARMARA	148	33.79	345	78.77
OAB	66	15.07	411	93.84
ABROAD	27	6.16	438	100.00

% 33,79 of the students' birth place is in Marmara Region. Second is Middle Anatolian Region by % 15,07. Ege Region is following by % 14,84 of population.

Most Lived Province

It is examined that if the most lived region should have an effect on the students' success or segments results. So, most lived place traffic code is asked in the questionnaire. Using the traffic codes, the most lived regions are formed for all students.

Table 4.8 Most Lived Region Distribution

Live_Most_Region	Frequency	Percent	Cumulative Frequency	Cumulative Percent
AKDENIZ	53	12.10	53	12.10
DAB	20	4.57	73	16.67
EGE	70	15.98	143	32.65
GDA	7	1.60	150	34.25
KARADENIZ	42	9.59	192	43.84
MARMARA	187	42.69	379	86.53
OAB	52	11.87	431	98.40

ABROAD	7	1.60	438	100.00

% 33,79 of the students birth place is in Marmara Region, but % 42,69 of the students most lived in Marmara Region. The second most lived region is Ege by % 15,98. It is seen that, the most lived region ranks differs from birth region ranks.

High School Province

High School education is very important for university success. So, the region that the high school education takes place is very important. Unfortunately, the education conditions of the Turkey's seven geographical regions are not even. So, it will be seen during the data mining studies, if the high school region is an effective attribute on student success or not. So, high school place traffic code is asked in the questionnaire. Using the traffic codes, the high school regions are formed for all students.

Table 4.9 High School Region Distribution

High_School_Region	Frequency	Percent	Cumulative Frequency	Cumulative Percent
AKDENIZ	56	12.79	56	12.79
DAB	16	3.65	72	16.44
EGE	73	16.67	145	33.11
GDA	8	1.83	153	34.93
KARADENIZ	40	9.13	193	44.06
MARMARA	187	42.69	380	86.76
OAB	48	10.96	428	97.72
ABROAD	10	2.28	438	100.00

Mother Education and Father Education

There is a widespread belief that parent's education status has a great effect on the children growing up duration. Especially, in the traditional Turkish families the children spent most of their time with their mothers. So, mother education status has the probability of being more effective than father education status. Actually, after the data mining study is completed, all the questions will be answered.

According to the students' answers the education status of parents are as follows:

Table 4.10 Mother Education Distribution

Mother_Edu	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Bachelor	103	23.52	103	23.52
Doctorate	1	0.23	104	23.74
Lycee	126	28.77	230	52.51
Master	6	1.37	236	53.88
None	21	4.79	257	58.68
Primary	110	25.11	367	83.79
Secondary	36	8.22	403	92.01
Önlisans	35	7.99	438	100.00

Table 4.11 Father Education Distribution

Father_Edu	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Bachelor	191	43.61	191	43.61
Doctorate	7	1.60	198	45.21
Lycee	98	22.37	296	67.58
Master	17	3.88	313	71.46
None	3	0.68	316	72.15
Primary	65	14.84	381	86.99
Secondary	26	5.94	407	92.92
Önlisans	31	7.08	438	100.00

If None, Primary and Secondary education classes' participation of mothers and fathers are taken into account, it is clear to notify, only 94 fathers are in these classes. On the other, the number of mothers whom education status belongs to these classes is 167. The education status of fathers looks better than the education status of mothers. These variables will be used as active variables during the data mining studies.

Number of Older Brother and Number of Older Sister

Maybe families are the most important stuff in humans' lives. Families have so many effects in all parts of our lives. In this study the number of older siblings, male and female, are also asked to find out their importance in the success of university education.

According to the students' answers the numbers of older siblings are as follows:

Table 4.12 Number of Older Brother Distribution

Older_Brother	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	329	75.11	329	75.11
1	87	19.86	416	94.98
2	18	4.11	434	99.09
3	3	0.68	437	99.77
5	1	0.23	438	100.00

% 75,11 of the students do not have an older brother. So it will not be meaningful to use this variable in the data mining model with five different classes' outcomes. Also just one student has five brothers. This variable will be transformed in to a binary variable, which has two classes' outcomes. The resulting distribution is as follows:

Table 4.13 Transformed Number of Older Brother Distribution

Older_Brother	Frequency	Percent	Cumulative Frequency	Cumulative Percent
NO	329	75.11	329	75.11
YES	109	24.89	438	100.00

If a student has an older brother the class will be Yes, and if the student does not have an older brother then the class will be No. In means of avoiding correlations, both of these two variables should not be used as active variables in the data mining study. So only the transformed variable will be selected for the model.

Table 4.14 Number of Older Sister Distribution

Older_Sister	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	287	65.53	287	65.53
1	112	25.57	399	91.10
2	28	6.39	427	97.49
3	8	1.83	435	99.32
4	1	0.23	436	99.54
5	1	0.23	437	99.77
6	1	0.23	438	100.00

Like Number of Older Brother variable, Number of Older Sister variable distribution does not look proper for the data mining study. There are four outlier classes in the distribution. As a result this variable will be transformed by applying discretizing.

%6,39 of the students have two older sisters. But to transform this variable in to binary form is convenient for the study. Because, if the variable will have three classes outcomes as having no sisters, having one sister, and having more than one sister, the percentage of having more than one sister will be nearly 9. Hence, Number of Older Sister variable is transformed in to a binary variable, which has two classes' outcomes. The resulting distribution is as follows:

Table 4.15 Transformed Number of Older Sister Distribution

Older_Sister	Frequency	Percent	Cumulative Frequency	Cumulative Percent
NO	287	65.53	287	65.53
YES	151	34.47	438	100.00

If a student has an older sister the class will be Yes, and if the student does not have an older sister then the class will be No. Both of these two variables should not be used as active variables in the data mining study. So only the transformed variable will be selected for the model.

%34,47 of the students have older sister(s) and %24,89 of the students have older brother(s).

Number of Younger Brother and Number of Younger Sister

In the questionnaire the number of younger siblings, male and female, are also asked to find out their importance in the success of university education. Generally the communication between a person and her/his younger siblings and the communication between a person and his/her older siblings differs from each other. In this study, which one has an effect on the success of university students will be tried to find out.

Table 4.16 Number of Younger Brother Distribution

Younger_Brother	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	280	63.93	280	63.93
1	135	30.82	415	94.75
2	21	4.79	436	99.54
3	1	0.23	437	99.77
7	1	0.23	438	100.00

% 63,93 of the students do not have a younger brother. So it will not be meaningful to use this variable in the data mining model with five different classes' outcomes. Also just one student has three brothers and another one has seven brothers. This variable will be transformed in to a binary variable, which has two classes' outcomes. The resulting distribution is as follows:

Table 4.17 Transformed Number of Younger Brother Distribution

Younger_Brother	Frequency	Percent	Cumulative Frequency	Cumulative Percent
NO	280	63.93	280	63.93
YES	158	36.07	438	100.00

If a student has a younger brother the class will be Yes, and if the student does not have a younger brother then the class will be No. Both of these two variables should not be used as active variables in the data mining study. So only the transformed variable will be selected for the model.

Table 4.18 Number of Younger Sister Distribution

Younger_Sister	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	324	73.97	324	73.97
1	95	21.69	419	95.66
2	15	3.42	434	99.09
3	3	0.68	437	99.77
4	1	0.23	438	100.00

Like Number of Younger Brother variable, Number of Younger Sister variable distribution does not look proper for the data mining study. There are three outlier classes in the distribution. As a result this variable will be transformed in to binary variable by applying discretization.

Table 4.19 Transformed Number of Younger Sister Distribution

Younger_Sister	Frequency	Percent	Cumulative Frequency	Cumulative Percent
NO	324	73.97	324	73.97
YES	114	26.03	438	100.00

If a student has a younger sister the class will be Yes, and if the student does not have a younger sister then the class will be No. Both of these two variables should not be used as active variables in the data mining study. So only the transformed

variable will be selected for the model. %36,07 of the students have younger brother(s) and %26,03 of the students have younger sister(s).

Some other variables are generated using number of younger and older, brothers and sisters. The generated variables are analyzed.

Table 4.20 Total Number Brothers Distribution

Total_Brothers	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	191	43.61	191	43.61
1	190	43.38	381	86.99
2	45	10.27	426	97.26
3	8	1.83	434	99.09
4	2	0.46	436	99.54
5	1	0.23	437	99.77
7	1	0.23	438	100.00

Using Total Number of Brother variable, it can be examined if having a brother without considering his age, is effective on success or not. But the distribution of the variable is not proper for the study and should be applied discretization.

Table 4.21Transformed Total Number of Brothers Distribution

Total_Brothers	Frequency	Percent	Cumulative Frequency	Cumulative Percent
NO	191	43.61	191	43.61
YES	247	56.39	438	100.00

If a student has a brother the class will be Yes, and if the student does not have a brother then the class will be No. Both of these two variables should not be used as active variables in the data mining study. So only the transformed variable will be selected for the model.

Table 4.22 Total Number Sisters Distribution

Total_Sisteres	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	191	43.61	191	43.61
1	179	40.87	370	84.47
2	50	11.42	420	95.89
3	10	2.28	430	98.17
4	5	1.14	435	99.32
5	2	0.46	437	99.77

Using Total Number of Sister variable, it can be examined if having a sister without considering her age, is effective on success or not. But the distribution of the variable is not proper for the study and should be applied discretization.

Table 4.23 Transformed Total Number of Sisters Distribution

Total_Sisters	Frequency	Percent	Cumulative Frequency	Cumulative Percent
NO	191	43.61	191	43.61
YES	247	56.39	438	100.00

If a student has a sister the class will be Yes, and if the student does not have a sister then the class will be No. Both of these two variables should not be used as active variables in the data mining study. So only the transformed variable will be selected for the model.

With a great coincidence the percentage of having sister and the percentage of having brother are equal. %56,39 of all students have at least one brother and %56,39 of all students have at least one sister. Both of these generated variables will be used in the models. The equality of the variables percentages does not mean that, the marked students are the same. So, these two variables have to have different effect on the success.

Table 4.24 Total Number of Sibling Distribution

Total_Sibling	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	32	7.31	32	7.31
1	247	56.39	279	63.70
2	107	24.43	386	88.13
3	26	5.94	412	94.06
4	13	2.97	425	97.03
5	6	1.37	431	98.40
6	4	0.91	435	99.32
8	3	0.68	438	100.00

At last Total Number of Sibling variable is generated using the former four variables asked in the questionnaire. Like the other generated variables, the distribution of the variable will be improved by applying discretization. But this time the transformed variable should be discretized in to three subclasses.

Table 4.25 Transformed Total Number of Sibling Distribution

Total_Sibling	Frequency	Percent	Cumulative Frequency	Cumulative Percent
>ONE	159	36.30	159	36.30
ONE	247	56.39	406	92.69
ZERO	32	7.31	438	100.00

The data, the classes consists of, are very easy to guess by reading the class names. This variable will be used in the models and it will discover if having any siblings is effective on success or not.

Turkish Nationality

In the questionnaire the students are asked if their nationality is Turkish or not.

Table 4.26 Turkish Nationality Distribution

Turkish_Nationality	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	8	1.83	8	1.83
Yes	430	98.17	438	100.00

The variable will not be used in model because nearly all of the students that have been participated the questionnaire are Turkish. So, the variable does not have any effect on the model.

Mother Working Status and Father Working Status

Today, both of the parents work to earn money and to provide a better future for their families. But thirty years ago, the life styles were different. Men were not used to let their wives to work outside. Actually, a working mother is a powerful view of the family which is more democratic and more civilized. It is wanted to find out, if mother and father working status are effective on students' success. So, in the questionnaire, these are asked. The results are as follows:

Table 4.27 Mother Working Distribution

Mother_Working	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	225	51.37	225	51.37
Yes	213	48.63	438	100.00

According to the answers, nearly half of the İTÜ's students' mothers had worked before or are still working today. This attribute will be used as active variable in the model. On the other hand, father working status is as follows:

Table 4.28 Father Working Distribution

Father_Working	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	17	3.88	17	3.88
Yes	421	96.12	438	100.00

In the questionnaire, it was asked, if their parents had worked before or if they are still working. According to the answers, %3,88 of the students' fathers had not worked before. But, it is clear that asking father working status does not point to the target. Also, the perception level of the question is doubtful. On some of the questionnaire both of the choices were marked and then one if them was scribbled. As a result this variable will not be used in the data mining study.

Parents Status

Since the family life is very important for a person, the parents' status asked in the questionnaire. Parent's status naturally affects the life style of the students. According to the answers:

Table 4.29 Parents Status Distribution

Father_Working	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Divorced	9	2.24	9	2.24
Married	386	96.26	395	98.50
Separated	6	1.50	401	100.00

386 of the students' parents are married. This value is very high. Also there are 37 missing answers, which were led by the questionnaire. If their fathers have died before then these students do not answer this questions. The blank ones are the students whom fathers had died. But also, Father_Alive variable will be used in the model. Accordingly, Parent Status variable will not be an active variable in the predictive modeling.

4.3 Education Questions (Before University)

This category of the questionnaire searches the background and the success of the student before university.

High School Type

Table 4.30 High School Type Distribution

High_School_Type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Abroad	1	0.23	1	0.23
Anatolian	217	49.54	218	49.77
Private	35	7.99	253	57.76
Profession	2	0.46	255	58.22
Science	112	25.57	367	83.79
State	71	16.21	438	100.00

According to the Turkish education system, the curriculums of the different high school types differ from each other. Generally, Science and Anatolian type High Schools are believed to be more successful comparing university entrance percentages of the students than the other types of High Schools.

It is seen that, in İTU, nearly 50 percentage of the students are graduated from Anatolian high schools and nearly 25 percentage of the students are graduated from Science high schools.

There are some outliers in the variable. Just one student has graduated from an abroad high school and just two students have graduated from Profession high school. Since this variable's measurement is nominal and the variable's type is character, it will not be very helpful to apply discretization. So this attribute is used as active variable in the model without any adjustments.

High School Graduation GPA

In this study high school success of the students is expected to indicate university success. So, in the questionnaire high school success of the students are asked.

Table 4.31High School GPA Statistics

Label	Maximum	Mean	Minimum	N
High School GPA	5	4,803	3,45	437

The resulting attribute's type is interval. It was possible to apply discretization to this variable and transform it into a nominal variable. But the variable is given to the model as if it is. By the way, data mining algorithm should have the possibility of splitting the variable at the right point. Some decision tree algorithms do not handle continuous variables. But the SAS decision tree algorithms do handle interval variables. At the end of the model, the effect of this variable will be found out.

According to the distribution of the variable, only 38 students have less than 4,5 high school GPA which means that, nearly all the students have very high, high school GPAs. 125 students have 5,00 as high school GPA which means that they are the first runner ups of their high school.

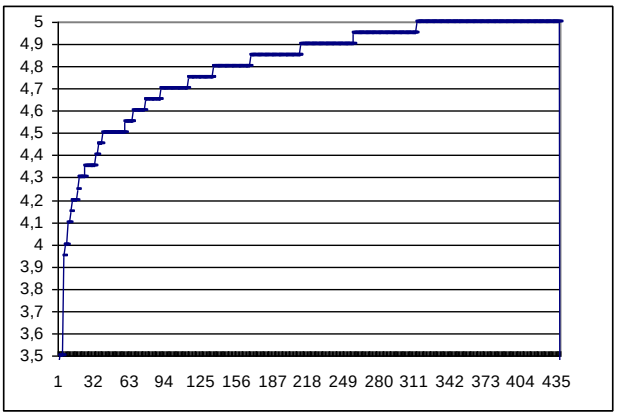


Figure 4.1 High School GPA Distribution

Entrance to University

In the questionnaire the students are asked if they get the right to enter İTÜ, at their first university entrance examination or not. The effect of this attribute will be investigated in the model. According to the answers:

Table 4.32 University Entrance First Distribution

Uni_Entrance_First	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	30	6.85	30	6.85
Yes	408	93.15	438	100.00

% 93,15 of the students get the right of entrance to İTÜ at their first university entrance examination trial. This variable will be used in the model to discover if it has an effect on success of the students or not.

Order of Department Selection

The order of students' department selection is asked in the questionnaire. The resulting distribution according to the answers is as follows:

Table 4.33 Order of Department Selection Distribution

Order_Of_Department_Selection	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	35	8.31	35	8.31
2	47	11.16	82	19.48
3	71	16.86	153	36.34
4	72	17.10	225	53.44
5	55	13.06	280	66.51
6	47	11.16	327	77.67
7	31	7.36	358	85.04
8	16	3.80	374	88.84
9	19	4.51	393	93.35
10	9	2.14	402	95.49
11	10	2.38	412	97.86
12	3	0.71	415	98.57
13	3	0.71	418	99.29
14	3	0.71	421	100.00

Over 421 students, only %8,31 of the students are studying their first choices. And, nearly %33,5 of the students are studying at the department which is not in their first five choices. There are 17 missing instances. Some of the students have been transferred from other universities or departments and some of the students are studying double departments. Since these students have not gotten the right to enter the department by the university entrance examination they left the answer of this question blank. Order of Department Selection attribute is a nominal attribute. There are some outliers in the distribution. It will not be proper to use the variable as it is, in the model. This variable will be discretized and the blank answers will be marked as transfer students. The resulting variable is as follows:

Table 4.34 Transformed Order of Department Selection Distribution

Order_Of_Department_Selection	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1-2	82	18.72	82	18.72
3-4	143	32.65	225	51.37
5-6	102	23.29	327	74.66
>=7	94	21.46	421	96.12
Other	17	3.88	438	100.00

If the students are studying their first or second choices they are marked as ‘1-2’ if the students are studying their third or fourth choices they are marked as ‘3-4’, if the students are studying their fifth or sixth choices they are marked as ‘5-6’, if the students did not attend the department by university entrance examination they are marked as ‘Other’, and the rest is marked as ‘>=7’. The transformed variable will be used in the data mining study as an active variable.

4.4 Academic Questions (During University)

There are different questions that ask for students’ adaptation to university education. These variables should have an important effect on students’ success.

Library Usage

During university education, since most of the projects and home works yield students to make research about specific subjects related their major, they generally utilize libraries or internet. But also libraries provide access to online catalogs and journals. In the questionnaire it is asked if the students utilize university library.

Table 4.35 Library Usage Distribution

Library_Usage	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	35	7.99	35	7.99
Yes	403	92.01	438	100.00

This variable will be used as active variable in the study.

Nonacademic Reading Interest

In the questionnaire the students are asked how often they read nonacademic materials such as books, newspapers and magazines. University students are expected to be interested in contemporary events, and news. Successful students are anticipated to have strong links with social life.

Table 4.36 Nonacademic Interest Distribution

Nonacademic_Interest	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Never	3	0.68	3	0.68
Rarely	41	9.36	44	10.05
Sometimes	185	42.24	229	52.28

Usually	209	47.72	438	100.00

This attribute will be used as an active variable in the model.

Academic Reading Interest

In the questionnaire the students are asked if they read academic materials or not.

Table 4.37 Academic Interest Distribution

Academic_Interest	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	230	52.51	230	52.51
Yes	208	47.49	438	100.00

Nearly half of the population reads academic materials and the others do not read. This variable will be used in the model as active variable. According to the data mining results it will be discovered if this variable is effective on students success or not.

Conference Involvement

The students are asked if they joined any conference or symposium about their major during their university duration.

Table 4.38 Conference Involvement Distribution

Conference_Involvement	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	92	21.00	92	21.00
Yes	346	79.00	438	100.00

% 79 of all the students have participated to a conference or a symposium. The effect of this attribute on student success will be examined during the data mining study.

Master of Science Degree

After four years of undergraduate education, university graduates sometimes prefer to get a Master of Science degree. In order to be accepted to a master degree, the students should have outstanding records. In the questionnaire the students are asked if they plan to get a master degree after graduation.

Table 4.39 Master Degree Distribution

Master_Degree	Frequency	Percent	Cumulative	Cumulative

			Frequency	Percent
No	74	16.89	74	16.89
Not Decided Yet	164	37.44	238	54.34
Yes	200	45.66	438	100.00

% 45 percent of the students have decided to get a master degree. This attribute will be used as an active variable in the model.

Recent Year's Cumulative Grade Point Average

The aim of this study is to find out the success drivers of the Management Faculty students. But the students that are successful and unsuccessful have to be determined to continue and make a progress. The rule of being a successful student has to be defined, and the study should be continued using this borderline.

Accordingly, it is decided to call a student successful if his recent cumulative GPA is equal or over 3,00 and otherwise unsuccessful. So, in the questionnaire the students are asked for their recent year's cumulative GPAs.

Two classes of students are formed according to the students recent cumulative GPAs. This variable is very important for the data mining study. This variable is the target variable of the decision tree model. The model will try to find out the best paths that will result in a successful or unsuccessful student sets.

Table 4.40 Recent_GPA_Success Distribution

Recent_GPA_Success	Frequency	Percent	Cumulative Frequency	Cumulative Percent
US	224	52.83	224	52.83
S	200	46.93	424	100.00

According to the distribution it is obvious that only 424 students have give answer to that question. After applying the questionnaire, it is understood that, some of the students did not know the abbreviation of GPA which was used in the question. So, some of the students did not answer that question. Since, this attribute is the target value, 14 students that did not give answer to that question, are disregarded from the study.

Expected Graduation GPA

In the questionnaire the students are asked for their expected graduation GPA. But generally the recent years cumulative GPA and the expected graduation GPA

answers were the same. Hence, it is understood that this attribute will have contribution to the model. As a result, this variable is not examined and omitted from the study.

School Projects

The students are asked if their school projects or home works make contribution for their success in their academic or professional lives.

Table 4.41 Projects Contribution Distribution

Project Contribution	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	59	13.47	59	13.47
No Idea	38	8.68	97	22.15
Yes	341	77.85	438	100.00

% 77,85 of the students believe that their projects or home works have an efficient effect in their academic or professional lives. This variable will be used in the model.

Attendance Percentage

In İTÜ at least %70 of attendance is mandatory in order to pass a course. The attendance percentages of the students are asked. It is anticipated that if the attendance percentage of the students have a positive effect on their success or not. The distribution of the attendance according to answers is as follows:

Table 4.42 Attendance Percentage Distribution

Attendance Percentage	Frequency	Percent	Cumulative Frequency	Cumulative Percent
70	165	37.67	165	37.67
80	94	21.46	259	59.13
90	138	31.51	397	90.64
100	41	9.36	438	100.00

%37,67 of the students attend classes just as necessary to pass a course. Besides, %9,36 percent of the students attend all the courses. This attribute will be used as an active variable in the model.

4.5 Personal Questions

In the questionnaire some of the questions are about life styles, hobbies, monetary status, and how they pass a day. These answers will provide important information for the data mining study.

Part Time Working

Part time working status could have an effect on students' success. So in the questionnaire it is asked if the students work part time or do not work.

Table 4.43 Part Time Working Distribution

Part Time Working	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	337	76.94	337	76.94
Yes	101	23.06	438	100.00

According to the answers, only % 23 of the students work part time. This attribute will be used as an active variable in the study.

Cinema Frequency

One of the popular hobbies of the university students are going to the cinema. Students different hobbies are tried to be uncovered if they have an effect on success or not. In the questionnaire the students are asked how often they go to cinema.

Table 4.44 Cinema Frequency Distribution

Cinema Frequency	Frequency	Percent	Cumulative Frequency	Cumulative Percent
>Once/Week	9	2.05	9	2.05
Never	49	11.19	58	13.24
Once/Month	312	71.23	370	84.47
Once/Week	68	15.53	438	100.00

Most of the students go to the cinema once a month. This variable will be used in the data mining study.

Doing Sports

In the questionnaire the students are asked if they do sports or not. This is an interesting variable, and it is anticipated if there is a relation between doing sports and being successful at school.

Table 4.45 Sports Distribution

Sports	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	191	43.61	191	43.61
Yes	247	56.39	438	100.00

% 56,39 of the students do sports. This variable will be used in the data mining study.

Hobby Course

In the questionnaire the students are asked if they follow a hobby course or not. It is tried to find out, if they spent time for their hobbies or personal improvement.

Table 4.46 Hobby Course Distribution

Hobby Course	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	293	66.89	293	66.89
Yes	145	33.11	438	100.00

Only % 33,11 of the students involve a hobby or personal improvement course. This yields the result that either these students do not have enough time or money or willingness. This variable will be used in the study.

Average Internet Usage

Internet is very widely used nowadays. Especially for students, internet is an inevitable source of knowledge. Nearly all the universities have computer laboratories that have internet connections. Students do not always use internet for their success on courses. Sometimes they use internet for chatting, playing games or just surfing or shopping.

In the questionnaire on average how many hours per day students spend their time using internet is asked for. Actually, students' reasons for spending time on internet should be asked for, but in the means of privacy, it is thought that this question will not be proper.

Table 4.47 Internet Hours Distribution

Internert Hours	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2-4 Hours	139	31.74	139	31.74

<2 Hours	216	49.32	355	81.05
>4 Hours	83	18.95	438	100.00

According to the answers, nearly %50 of the students spend only less than two hours per day using internet. Only % 18,95 of the students pass more than four hours per day using internet. This variable will be used as an active variable in the study.

Average Television Watching Duration

In Turkey, watching television is one of the habits of the families. Turkish people like watching television. But for university students watching television can be a very time consuming habit. In the questionnaire the students are asked for on average how many hours per day they spend time watching television.

Table 4.48 Television Hours Distribution

Television Hours	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2-4 Hours	64	14.61	64	14.61
<2 Hours	368	84.02	432	98.63
>4 Hours	6	1.37	438	100.00

It is pleasant to see that, % 84 of the students spend only less than two hours per day watching television. This variable will be used as an active variable in the study.

Average Studying Hours

University students should spend most of their time studying. In the questionnaire the students are asked for on average how many hours per day they spend time studying.

Table 4.49 Studying Hours Distribution

Studying Hours	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2-4 Hours	113	25.80	113	25.80
<2 Hours	314	71.69	427	97.49
>4 Hours	11	2.51	438	100.00

Most of the students study less than two hours per day. Just %2,5 of the students study more than four hours a day. This variable will be used as an active variable in the study.

Meeting University Friends

This questions aim at evaluating the socialization level of the students, and its effect on success.

Table 4.50 Meeting University Friends Distribution

Meeting Uni. Friends	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	23	5.25	23	5.25
Yes	415	94.75	438	100.00

%5,25 of the students do not meet friends from university. This variable will be used as an active variable in the study. Most of the students see friends from university. Accordingly the socialization level of the campus can be thought as high. This variable will be used as an active variable in the study.

School Club Participation

In the questionnaire the students are asked if they take place in the activities of the school clubs.

Table 4.51 School Club Participation Distribution

School Club Participation	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	299	68.26	299	68.26
Yes	139	31.74	438	100.00

%31,74 of the students take place in the activities of school clubs. This variable will be used as an active variable in the study.

Residence Type in İstanbul

İTÜ is placed in İstanbul. Only %42,69 of the students most lived region is Marmara. 116 students' most lived city is Istanbul. If it is assumed that, these 116 students were still living in İstanbul when they attend to university; at least 332 students had to provide a residence for themselves in İstanbul.

The type of residence should have an important effect on the success of the students. So this question is asked in the questionnaire.

Table 4.52 Residence Type Distribution

Residence Type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Home Alone	12	2.74	12	2.74
Hostel	121	27.63	133	30.37
Own Family	92	21.00	225	51.37
Roommate	183	41.78	408	93.15
With a family	30	6.85	438	100.00

%41,78 of the students stay with a roommate. The following is %27,63 of the students stay in hostel. This variable will be used as an active variable in the study.

Arrival to School

In the questionnaire the students are asked how they arrive to school. This attribute could have an effect on success of the students.

Table 4.53 School Arrival Distribution

School Arrival	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Public Transport	346	79.00	346	79.00
Private Vehicle	24	5.48	370	84.47
Walking	68	15.53	438	100.00

%79 of the students arrive school by using public transport. Just %5,48 of the students use their private vehicles to arrive school. This variable will be used as an active variable in the study.

Average Transportation Duration

Generally students spend most of their time on way going to school. The average duration of the time that was spent on way to school is asked in the questionnaire.

Table 4.54 Arrival Duration Distribution

School Arrival	Frequency	Percent	Cumulative Frequency	Cumulative Percent
15-30 min.	79	18.04	79	18.04
30-60 min.	203	46.35	282	64.38
<15 min.	48	10.96	330	75.34
>60 min.	108	24.66	438	100.00

The class values of the attribute are very easy to understand when reading the class names. “15-30 min” means that the student passes more than 15 and less than 30

minutes on way to school. Only %10,96 of the students spend less than 15 minutes on way. %24,66 of the students spent more than an hour on way to school. This variable will be used as an active variable in the study.

Studying Environment

For the success of the students the studying conditions are very important. So, in the questionnaire the students' level of pleasure about their studying environment is asked.

Table 4.55 Studying Environment Distribution

School Arrival	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Not Pleased	61	13.93	61	13.93
Pleased	277	63.24	338	77.17
Very Pleased	100	22.83	438	100.00

This variable will be used as an active variable in the study.

Private Computer Usage

During university education home works and projects are generally done using computers. Also, a computer is needed to access to internet. In school it is possible to use common computers in a laboratory. But it is more functional to have a computer access at the residence place.

Table 4.56 Private Computer Usage Distribution

Private Computer Usage	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	57	13.01	57	13.01
Yes	381	86.99	438	100.00

% 86,99 of the students have private computer usage. This variable will be used as an active variable in the study.

Average Monthly Expenditure

The income level of the students can affect their success at school. In the questionnaire the average monthly expenditure of the students are asked.

Table 4.57 Average Monthly Expenditure Distribution

Average Monthly Expen.	Frequency	Percent	Cumulative	Cumulative

			Frequency	Percent
200-400 mio	194	44.29	194	44.29
400-600 mio	140	31.96	334	76.26
<200 mio	46	10.50	380	86.76
>600 mio	58	13.24	438	100.00

Most of the students spend between 200 and 400 YTL per month. This variable will be used as an active variable in the study.

Scholarship

There are different scholarships opportunities are available to support students in financial means.

Table 4.58 Having Scholarship Distribution

Having Scholarship	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	144	32.88	144	32.88
Yes	294	67.12	438	100.00

%67,12 of the students get scholarship. This variable will be used as an active variable in the study.

Tuition Credit

In Turkey, the students have to pay tuition even for public universities. Sometimes students prefer to use tuition credit, and begin to pay back after their graduation.

Table 4.59 Tuition Credit Distribution

Tuition Credit	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	141	32.19	141	32.19
Yes	297	67.81	438	100.00

According to the answers, %67,81 of the students get tuition credit. This variable will be used as an active variable in the study.

Finding Job

Nowadays to find a job is a very though process. The rate of unemployment is nearly %10. In the questionnaire the students are asked, if they could find a job whenever they want after graduation.

Table 4.60 Finding Job Distribution

Finding Job	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	112	25.57	112	25.57
Yes	326	74.43	438	100.00

According to the answers %74,43 of the students believe that they will fing a job whenever they want. This variable will be used as an active variable in the study.

Finding Job Duration

The students are asked for their opinion about finding job duration.

Table 4.61Finding Job Duration Distribution

Finding Job Duration	Frequency	Percent	Cumulative Frequency	Cumulative Percent
12 months	70	15.98	70	15.98
6 months	216	49.32	286	65.30
Immediately	77	17.58	363	82.88
No Idea	75	17.12	438	100.00

%67 of the students believe that they would find a job within six months. This variable will be used as an active variable in the study.

Social Environment of İTÜ

In the questionnaire the students are asked their opinion about İTÜ's social environment.

Table 4.62 Social Environment Distribution

Social Environment	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Bad	253	57.76	253	57.76
Good	15	3.42	268	61.19
So-So	170	38.81	438	100.00

%57,76 of the students think that İTÜ's social environment is Bad. Generally the students are not pleased about İTÜ's social environment. This variable will be used as an active variable in the study.

Most Liked Course Group

There are different course groups in İTÜ. In the questionnaire the students are asked for their favorite course group.

Table 4.63 Most Liked Course Group Distribution

Finding Job Duration	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Design	183	42.66	183	42.66
Fundamental Engineering	98	22.84	281	65.50
Fundamental Sciences	29	6.76	310	72.26
Human and Social Sciences	119	27.74	429	100.00

Most liked course group of students is Design by %42,66 of the students. This variable will be used as an active variable in the study.

Pleasure Level of Being in İTÜ

The general pleasure level of being in İTÜ is asked to the students in the questionnaire.

Table 4.64 General Level of Pleasure Distribution

General Level of Pleasure	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Not Pleased	58	13.24	58	13.24
Pleased	309	70.55	367	83.79
Very Pleased	71	16.21	438	100.00

%70,55 of the students are pleased to be in İTÜ. This variable will be used as an active variable in the study.

4.6 Formed Variables

In data mining studies, all of the input attributes are investigated as done in this chapter. Accordingly, if the distribution of one of the variables is seemed to mistake the result of the study or is seemed not to have any effect on the result, then some adjustments are performed. Under this concept, some attributes are omitted from the study. On the other hand, some attributes are discretized to reuse in the study. Discretization is widely explained in section 5.4.

In this study the variables that are applied discretization are as follows:

- Number of Older Brother Distribution

- Number of Older Sister Distribution
- Number of Younger Brother Distribution
- Number of Younger Sister Distribution
- Total Number of Brothers Distribution
- Total Number of Sisters Distribution
- Total Number of Sibling Distribution
- Order of Department Selection
- Recent_GPA_Success (target variable)

5. DECISION TREES FOR PREDICTIVE MODELING

In this study, the critical success factors of students are aimed to be found out. The mapping of the inputs to the target is a predictive modeling. This objective will be fulfilled by applying a predictive data mining method. The data used to estimate a predictive model is a set of cases (observations, examples) consisting of values of the inputs and target. In the model, the target variable has two classes: one class is successful students and the other is unsuccessful students. The success variable is the target variable of the model and using the independent variables, which independent variables are important in the success model of students will be found out.

Decision trees are a way of representing a series of rules that lead to a class or value. Student classification analysis aims to identify the background characteristics that indicate the predefined group (successful or unsuccessful) to which each student belongs. It links the value of success in Istanbul Technical University to the background information. Decision tree classification of data mining creates classification models by examining already classified data from a historical data source and inductively finding a predictive pattern. This pattern can be used to understand the existing data.

5.1 What a Decision Tree Is

It is a predictive model viewed as a tree consisting of the decision nodes, branches and leaves. A decision node specifies a test to be carried out. The results of this test cause the tree to split into branches without loosing any of the data. The split decision is made at the node and it is never revisited. All splits are made sequentially, so each split is dependent on its predecessor. Thus all future splits are dependent on the first split, which means the final solution could be very different if a different first split is made. Each branch of the tree is a possible answer to the classification question and will lead either to another decision node or to the bottom of the tree, called a leaf node as seen in Figure 5.1 [20].

A decision tree is read from top-down starting at the root node. Each internal node represents a split based on the values of one of the inputs. The inputs can appear in any number of splits throughout the tree. Cases move down the branch that contains its input value. In a binary tree with interval inputs, each interval internal node is a simple inequality. A case moves left if the inequality is true and right otherwise. The terminal nodes of the tree are called leaves. The leaves are the partitions of the data set with their classification and represent the predicted target. Decision tree process starts at the root node and moves to each subsequent node until a leaf node is reached. All cases reaching a particular leaf are given the same predicted value. The leaves give the predicted class as well as the probability of class membership.

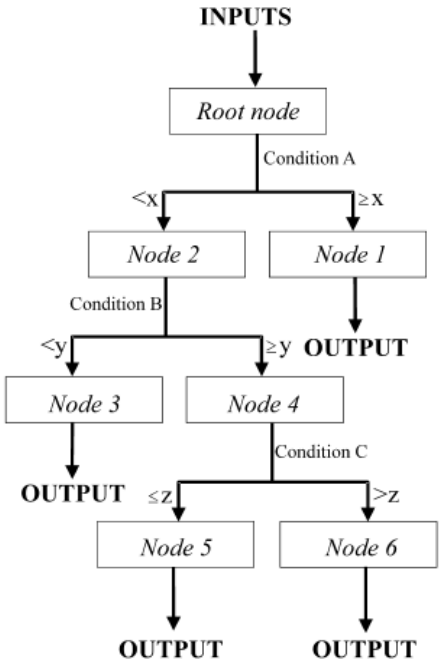


Figure 5.1 General Structure of a Data Mining Decision Tree

When the target is categorical, the model is called a classification tree. When the target is continuous, the model is called a regression tree.

Decision trees handle non-numeric data very well. This ability to accept categorical data minimizes the amount of data transformations. Continuous predictors can frequently be used even in these cases by converting the continuous variable to a set of ranges (binning). Some decision trees do not support continuous variables, in which case the response variables in the data set must also be binned to output classes (discretizing).

The advantage of decision tree is that they can be read like English sentences. However, with a large number of data factors it might be very difficult to actually understand what is being said. The disadvantage of these representations is that they are not suited for numbers covering large intervals. This is because each rule or point in the decision tree represents one relationship. To represent the relationships over a large interval many rules or decision tree points are required.

Trees are also appealing because they accept several types of variables: nominal, binary, ordinal, and interval. In the SAS dataset, a variable is listed as ordinal when it is a numeric variable with more than two but no more than ten distinct non missing levels in the data sample. These are categories that can be rationally listed in some order. The variables have the measurement level interval because they are numeric variables in the SAS data set and have more then ten distinct levels in the data sample. The miner identifies a variable as binary when it has only two distinct non missing levels in the data sample. If an attribute is a character variable with more then two levels then the model role will be nominal [17, 21, 22].

5.2 How Decision Tree Are Built

Decision trees are built through a process known as recursive partitioning. Recursive partitioning is an iterative process of splitting the data up into partitions. It is top-down greedy algorithm. Initially all of the records are together in one big box. Starting at the root node, a number of splits that involve a single input are examined. For interval inputs, the splits are disjoint ranges of the input values. For nominal inputs, the splits are disjoint subsets of the input categories. Various split-search strategies can be used to determine the set of candidate splits. A splitting criterion is used to choose the split. The splitting criterion measures the reduction in variability of the target distribution in the child nodes. The cases in the root node are then partitioned according to the selected split [4,23].

5.2.1 Splitting Criteria

The process starts with a data set consisting of predefined records. Preclassified means that the target field, or dependent variable, has a known class. The goal is to build a tree that distinguishes among the classes.

The first task is to decide which of the independent fields makes the best splitter. The best split is defined as one that does the best job of separating the records into groups where a single class predominates. The measure used to evaluate a potential splitter is the reduction in diversity, which is just another way of saying "the increase in purity".

To choose the best splitter at a node, the decision tree algorithm considers each input field in turn. In essence, each field is sorted. Then, every possible split is tried. The diversity measure is calculated and the best split is the one with the largest decrease in diversity. This is repeated for all the fields. The winner is chosen as splitter for that node.

In contrast, if a split results in pure child nodes, the split is undisputedly best. For classification trees, the three most widely used splitting criteria are based on the Pearson chi-squared test, the Gini index, and entropy. All three measure the difference in class distributions across the child nodes. The three methods usually give similar results [4, 23].

Gini Impurity

The Gini Index is a measure of variability for categorical data (developed by the eminent Italian statistician Corrado Gini in 1912). The Gini index can be used as a measure of node impurity where p1, p2,……., pr are the relative frequencies of each target class in a node [23].

CART (Classification and Regression Trees) are developed in 1984 [24]. In building a CART tree, best predictor is picked according to how well it splits apart the records with different predictions. The measure is the reduction in diversity for choosing the best split. Gini index is a diversity metric used by CART algorithm.

The Gini value for a given segment is calculated to be one minus the sum of squared probabilities for each prediction. In other words, Gini index is the probability that the second event chosen belongs to a different class than the first. Therefore, Gini index will be maximum when the proportions of each prediction value are equivalent, and it will be the minimum when the segment is homogenous. The limiting value is 0.5 for binary events or 1/n when there is n number of categories. This value is reached when each class has exactly the same number of members. The total reduction in

diversity is the diversity at root minus the weighted average of the diversity of the children [4].

Probability of the class (i) being chosen twice is P_i^2. The diversity index is simply one minus the sum of all P_i^2.

$$\text{Gini (T)} = 1 - \sum P_i^2 \quad (5.1)$$

When there are only two classes the formula is found as:

$$\text{Gini (T)} = 2 P_1(1-P_1) \quad (5.2)$$

Entropy

Entropy is a measure of variability for categorical data. Consider r mutually exclusive events with probabilities p_1, p_2, ……., p_r . The rarity of a particular outcome can be measured as:

$-\log_2 (p_i)$.

Entropy is the average rarity and thus measures the uncertainty of outcome. Entropy is a measure commonly used in information theory. The higher the entropy of an attribute, the more uncertainty there is with respect to its outcomes. Thus, attributes are selected in order of increasing entropy, where the root node of tree would correspond to the attribute with the lowest entropy value. The formula for the entropy of any given attribute, A_k, is given as:

$$H(C / A_k) = \sum_{j=1}^{M_k} p(a_{k,j}) \times \left[- \sum_{i=1}^{N} p(c_i / a_{k,j}) \log p(c_i / a_{k,j}) \right] \quad (5.3)$$

where

$H (C / A_k)$ = entropy of the classification property of Attribute A_k.

$P (a_{k,j})$ = probability of attribute k at value j.

$P (c_i / a_{k,j})$ = probability that the class value is c_i when attribute k is at its j^{th} value.

M_k = total number of values for attribute A_k; j = 1,2,…. M_k

N = total numer of different classes; i= 1,2,…N

K = total number of attributes; k= 1,2,..K

The terms in the brackets called the information. Thus, entropy is the expected information that is sum of the information in the several possible outcomes multiplied by their probability [17].

Chi –Square Test

The Pearson chi – squared test can be used to judge the worth of the split. It tests whether the class proportions are the same in each child node. The test statistics measures the difference between the observed cell counts and what would be expected if the branches and target classes were independent.

The statistical significance of the test is not monotonically related to the size of the chi-squared test statistics. The degrees of freedom of the test is (r-1)(B-1) where r is number of target levels and B is number of branches. The expected value of a chi-square test statistics with v degrees of freedom equals v. Consequently, tree with more branches, will naturally have larger chi-squared statistics [23].

5.2.2 Stopping Rules

The size of a tree may be the most important single determinant of quality, more important, perhaps, than creating good individual splits. Trees that are too small do not describe the data well. Trees that are too large have leaves with too little data to make any reliable predictions about the contents of the leaf when the tree is applied to a new sample. Splits deep in a large tree may be based on too little data to be reliable.

The easy algorithm tries to avoid making trees too small or too large by splitting any node that contains at least twenty percent of the original data. This strategy is easily foiled. Twenty percent of the data will be too few or too many when the original data contain few or many cases. With some data sets, the algorithm will split a node into branches in which one branch has only a couple of cases. Many splits are required to whittle the thick branches down to the limit of twenty percent, so most of the data will end up in tiny leaves.

Certain limitations on partitioning seem necessary:

- the minimum number of cases in a leaf
- the minimum number of cases required in a node being split

- the depth of any leaf from the root node

Appropriate values for these limits depend on the data. For example, a search for a reliable split needs less data the more predictable the data are. A stopping rule that accounts for the predictive reliability of the data seems like a good idea. For this purpose, the CHAID algorithm applies a test of statistical significance to candidate splits. Partitioning stops when no split meets a threshold level of significance. Unfortunately, this stopping rule turns out to be problematic in three related aspects: choice of statistical test, accommodations for multiple comparisons, and the choice of a threshold.

5.3 Decision Tree Algorithms

Various decision tree algorithms exist like ID3/C4.5, CART and many others. Each produces trees that differ from one another in the number of splits allowed at each level of tree, how those splits are chosen when the tree is built, and how the tree growth is limited to prevent over fitting.

Decision tree algorithms are a form of supervised learning. The first step in the algorithm is the process of making the tree grows. Algorithm seeks to create a tree that works as perfectly as possible on all the available data. The goal of decision trees is to have homogenous leaves with respect to the prediction value. Decision trees are built through a process known as recursive-partitioning. It is an iterative process of splitting the data into partitions until some stopping points [24].

Step 1: Independent and dependent variables are chosen from a data source. According to the goal of the data mining, the user chooses a dependent, in other words, a target variable.

Step 2: A variable among independent variables is deemed to be the most predictive for the dependent variable and is used to split the data. An obvious question at this point would be how a decision tree will pick one predictor among others and make the split. The algorithm chooses the split that partitions the data into parts that are purer than the original. In other words, decision tree algorithms choose the best split which decreases the disorder of the data set more than the other splits. The best split reduces the disorder of the whole data by creating more ordered smaller partitions. Splitting criteria are explained in section 5.2.1.

Step 3: Then this splitting is applied to each new partition. Each new partition will now be the input of the algorithm recursively until the stopping criterion is achieved.

Step 4: Most of the decision tree algorithms stop growing the tree when each new partition is completely organized into just one value for the target variable and pure or when the partition contains algorithmically defined minimum number of records. This is because of the statistical reasons. Few records are not enough to make predictions based on historical data. These rules are determined at the initial step of the parameter setting. All the stopping rules that the tool allow the miner is set, before the model is run.

5.3.1 Well Known Algorithms

The following list contains the best known algorithms and describes how they address the main issues. Each algorithm was invented by a person or group inspired to create something better than what currently existed. Some of their opinions on strategies for creating a good tree are included.

AID, SEARCH

James Morgan and John Sonquist proposed decision tree methodology in an article in the Journal of the American Statistical Association in 1963. Their original program was called AID, which is an acronym for "automatic interaction detection." AID represents the original tree program in the statistical community. The third version of the program was named SEARCH and was in service from about 1971 to 1996.

James Morgan opposes splitting nodes into more than two branches because the data thins out too quickly. He also opposes subjecting splits to tests of statistical significance because the large number of required tests renders them useless. Morgan and Sonquist advocate using test data to validate a tree.

CHAID

AID achieved some popularity in the early 1970s. Some nonstatisticians failed to understand over-fitting and got into trouble. The reputation of AID soured. Gordan Kass (1980), a student of Doug Hawkins in South Africa, wanted to render AID safe and so invented the CHAID algorithm. CHAID is an acronym for "Chi-Squared Automatic Interaction Detection.”

CHAID recursively partitions the data with a nominal target using multiway splits on nominal and ordinal inputs. A split must achieve a threshold level of significance in a chi-square test of independence between the nominal target values and the branches, or else the node is not split.

CART

Leo Breiman and Jerome Friedman, later joined by Richard Olshen and Charles Stone, began work with decision trees when consulting in southern California in the early 1970s. They published a book and commercial software in 1984.

Their program creates binary splits on nominal or interval inputs for a nominal, ordinal, or interval target. An exhaustive search is made for the split that maximizes the splitting measure. The available measures for an interval target are reduction in square error or mean absolute deviation from the median. The measures for a nominal target are reduction in the Gini index.

ID3, C4.5, C5

While all decision-tree algorithms have the similar type of process, they employ different mathematical algorithms to determine the splitting criterion. They use different methods to find the best split that decrease the disorder of the data set.

ID3, C4.5, C5.0 are algorithms introduced by J. Ross Quinlan for inducing decision trees. The basic ideas behind ID3 are that:

- In the decision tree, each node corresponds to a non-categorical attribute, in other words a predictor, and each arc to a possible value of that attribute. A leaf of the tree specifies the expected value of the categorical attribute, independent variable, for the records described by the path from the root to that leaf.

- In the decision tree at each node, a non-categorical attribute which is the most informative is chosen among the attributes not yet considered in the path from the root.

- Entropy is used to measure how informative is a node. ID3 chooses a non-categorical variable on the basis of the information gain that this split provides. Gain represents the difference between the information needed to

make a prediction correctly before and after the split. In other words, if the information required is much lower after the split is made, then it means that the split has decreased the disorder of the initial single data segment.

Information gain is the difference between the entropy of the original segment and the accumulated entropy of the resulting split segments. Entropy is a well-defined measure of disorder or information found in the data.

In general, if the following probability distribution is given, P = (p1, p2... pn), then the information found in this distribution, also called entropy of P is:

$$Info(P) = -(p_1 \log(p_1) + p_2 \log(p_2) + \ldots + p_n \log p_n) \quad (5.4)$$

The ID3 algorithm is used to build a decision tree, given a set of R which is composed of non-categorical attributes C1, C2...Ck, the categorical attribute C, and a training set T of records. At the beginning, only the root is present. At each node the following divide and conquer algorithm is executed, trying to choose the best split, with no backtracking allowed as seen in Figure 5.2.

```
Function ID3  (R: a set of non-categorical attributes,
                    C: the categorical attribute,
                    T: a training set) returns a decision tree;
    begin
    If T is empty, return a single node with value Failure;
    If T consists of records all with the same value for
    the categorical attribute, return a single node with that value;
        If R is empty, return a single node with as value the most frequent of
        the values of the categorical attribute that are found in
        records of T;
            Let D be the attribute with largest Gain(D,T) among attributes in R;
            Let {dj| j=1,2, .., m} be the values of attribute D;
            Let {Tj| j=1,2, .., m} be the subsets of T consisting  respectively of
                    records with value dj for attribute D;
            Return a tree with root labeled D and arcs labeled  d1, d2, .., dm going
                    respectively to the trees
            ID3(R-{D}, C, T1),
            ID3(R-{D}, C, T2),
             ...,
            ID3(R-{D}, C, Tm);
    end ID3;
```

Figure 5.2 ID3 algorithm by Quinlan

High cardinality predictors in ID3 affect the accuracy of the resulting model. This is because of the many small segments that will be formed with little data in them. A high cardinality predictor is the one that has many possible values to perform splitting. A field with a customer name or a zip code can be taken as examples. An experienced data mining specialist will discard those variables. For instance, the number of records in data set is 20 and there are 20 different customer names. If the splitting criterion is customer name for churn prediction, then there will be 20 different small segments with only one record in each of them. The segments are fully homogenous. Entropy of those resulting 20 segments is zero meaning no disorder anymore, and there will be no other splits that will be better than this. So the model will choose customer name as the best split, but this model will never work well for the data other than this historical data.

To deal correctly with those high cardinality predictors, ID3 improved the gain theory and introduced gain ratio. J.R. Quinlan suggests using the following ratio instead of using only gain:

$$GainRatio(D,T) = \frac{Gain(D,T)}{SplitInfo(D,T)} \qquad (5.5)$$

C4.5 is an enhancement of ID3 algorithm in several subjects. The algorithm can deal with training sets that have records with unknown attribute values by evaluating the gain, or the gain ratio, for an attribute by considering only the records where that attribute is defined. Using produced decision tree, records that have unknown attribute values can be classified by estimating the probability of the various possible results [11, 17, 23-24].

5.4 Discretizing Numeric Attributes

Some classification and clustering algorithms deal with nominal attributes only and can not handle ones measured on a numeric scale. To use them on general datasets, numeric attributes must first be discretized into a small number of distinct ranges. Even learning algorithms that do handle numeric attributes sometimes process them in ways that are not altogether satisfactory. Statistical clustering methods often assume that numeric attributes have a normal distribution, often not a very plausible assumption in practice.

Although most decision tree and decision rule learners can handle numeric attributes, some implementations works much more slowly when numeric attributes are present because they repeatedly sort the attribute values. There is a global method of discretization that is applied to all continuous attributes before learning starts. It is a simple but an effective technique: sort the instances by attributes values and assign the value into ranges at the point that the class value changes.

When using global discretization, there are two possible ways of representing the discretized data to the learner. The most obvious way is to treat discretized attributes like nominal ones: each discretization interval is represented by one value in the nominal attribute. However since a discretized attribute is derived from a numeric one, its values are ordered, and treating it as nominal discards this potentially valuable ordering information. Of course, if a learning scheme can handle ordered attributes directly, the solution is obvious: each discretized attribute is declared to be of type ordered. [25]

5.5 Decision Tree Pruning

The algorithm should be checked if the model overfits the data. To overcome this problem, decision tree algorithms are using validation approach and trying many different simpler versions of the tree on a validation set. Pruning is the process of improving the performance of decision tree by removing leaves and branches. A pruned tree is in fact a subset of the full decision tree. Tree building algorithms make their best split at the beginning. Each partition is smaller then the whole population and undergoes to the same splitting criteria. This process goes on until the stopping criteria. As the partitions get smaller, they are being less representative of the population and overfit the data. There are some pruning techniques referred as bonsai techniques for avoiding overfitting. Bonsai techniques [4] try to stunt the growth of the tree before it gets too deep. This is also called top down pruning. For instance, setting a limit for the minimum number of records that must be in a node is a type of bonsai techniques.

Pruning methods let the decision tree to grow quite deep and then to prune off the branches that fail to generalize the data. This is called bottom up pruning. One common approach is to find the error rate associated with the subtrees of the initial tree. But these error rates should not be calculated with the same data set. Error rate

decreases as the tree gets more complex. So a complexity term is added to the error rate to discourage greater complexity. An addition of a branch is allowed only when its improvement in tree performance is large enough to overcome the extra complexity. When the data is large enough to build some test data sets, the performance of initial tree and subtrees are measured on separate test data sets. Figure 5.2 shows how pruning works with training and test data sets [4].

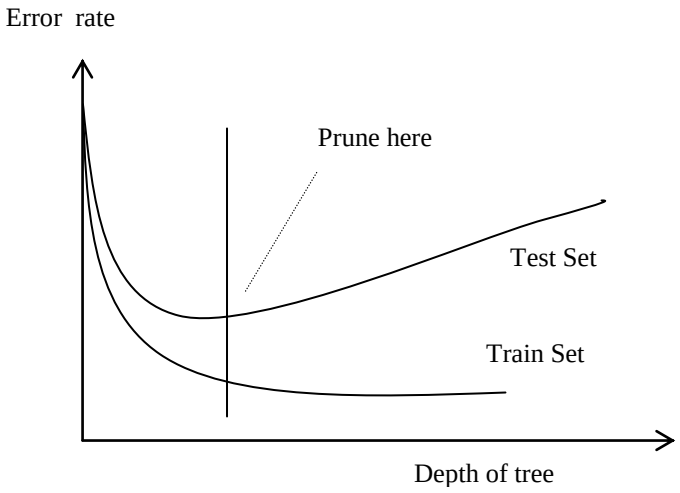


Figure 5.3 Pruning of a tree

5.6 Measure of Performance

At the end of model building phase, there are several models produced by different algorithms and ready for the assessment.

5.6.1 Response Curves

The performance of each algorithm is measured by a response curve, also called gains chart. To make a response curve, events are ranked by the predicted probability of the response in descending order and divided into deciles. Within each decile, actual percentage of the responders is calculated using the validation data set. Afterwards, the deciles are plotted on the horizontal axis and the actual percentage of responders in each decile is put on vertical axis. Cumulative response curve has the cumulative percentage of the respondents on the vertical axis.

If the performance of the model is good, the proportion of actual responders will be relatively high in the first decile. Figure 5.4 illustrates both the noncumulative and cumulative response curves. The line passing through both curves is the baseline that reveals the percentage of responders if a random sample of the customers is taken in each decile. In Figure 5.3, the ratio of actual responders over the population in the first decile is 70 per cent. The ratio of actual responders is 20 per cent if a random sample of the same size as in first decile is taken.[11]

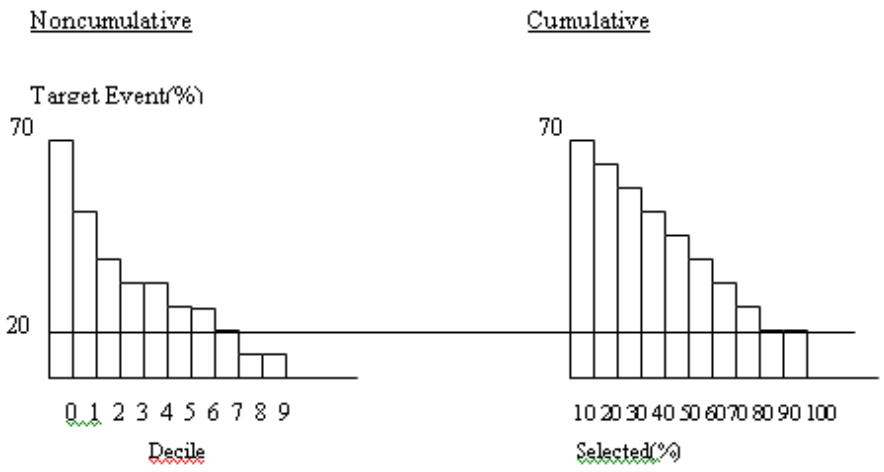


Figure 5.4 Non cumulative and cumulative response curves

5.6.2 Lift Curves

Lift curves represent similar information as response curves but on a different scale. This curve is obtained by dividing the response rate in each decile by the population response rate. The lift chart plots relative improvement over baseline.

$$LiftValue = \frac{Response\ rate\ in\ each\ decile}{Response\ rate\ in\ population} \qquad (5.6)$$

Lift measures the degree to which the prediction model increased the density of responses for a given subset of the database over what would be achieved by no model (random selection). The performance improvement is usually measured for lift by stating the percentage of the population for which the prediction will be used and the lift for that subset. For instance if the normal density of response in the population to a targeted mailing were increased to 30 percent but by focusing in on just the top quarter of the population predicted to respond by the predictive model, the response were increased to 30 percent, then the lift would be 3 for the first

quartile (quarter of the database, lift = 30%\10% = 3). Since this method is for measuring lift, it gives only a limited view of the improvement for one particular subset of the entire population [24].

5.6.3 Captured Response Curves

Response curve provides information about the percentage of responders in each decile. In captured response curves, the percentage of the responders in a decile over the total number of responders is investigated. In Figure 5.5 the curve stands for the captured response curve. The captured response for first decile is 35 per cent which means if 10 per cent of the customers are reached, 35 per cent of responders are obtained.

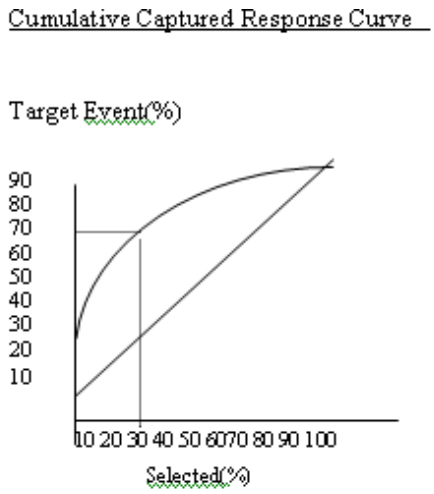


Figure 5.5 Cumulative captured response curve

When the lift curve is close to the baseline, it means that model is not good. The more is the lift value the better is the model. The model two represented in Figure 5.6 is better than the model one. If 30 per cent of the customers are contacted, 85 per cent of the actual responders are obtained in model 2, and 55 per cent of the responders in model one.

To select the best model, different criteria are considered also other than the curves described. A model can perform better in all response curves, but its complexity may overcome this advantage. The more complex a model is, the more difficult it is to understand it. The model must be stable and behave similar on the different data sets. For instance, it is possible to get an unstable model on training and test sets because of over fitting. Moreover, performance of the selected model must be observed over

time and if there is a decrease in the accuracy of the model, it is necessary to develop a new one [11].

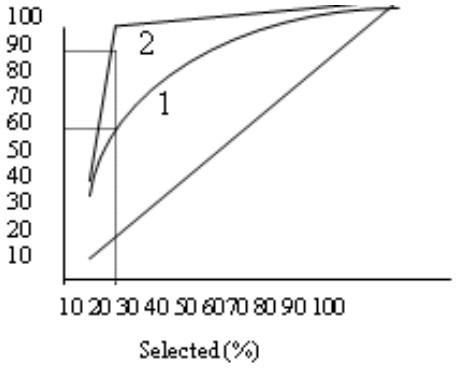


Figure 5.6 Captured response curves for two different models

5.6.4 Confusion Matrix

A two class classification model can produce two classifications leading to four results. Taking "1" to indicate class membership and "0" to indicate class nonmembership, Table 5.1 shows the possibilities. The model can classify as 1 when the actual result is 1, or it can classify as 1 when the actual result is 0. Similarly it can indicate 0 when the actual result is 0, or indicate 0 when the actual result is 1. This table forms the basis of what is called a confusion matrix because it allows an easy indication of where the model is confused (classifies 1 for 0 and 0 for 1) and where it isn't confused.

Table 5.1 Confusion Matrix

Model	**Class1**	**Class0**	**Total**
Is 1	189	35	224
Is 0	27	173	200
Total	216	208	424

Table 5.1 summarizes the performance of the model. The column labeled "Class 1" contains a count of the number of times the model predicted 1 as the class. Similarly column "Class 0" contains count of the number of times the model predicted 0. The row headed "Is 1" contains counts for the number of instances that actually are 1, and similarly row "Is 0" contains count for the number of instances that actually are 0. At the bottom in row "Total" is the column total, and to the right is the row total.

The interpretation is very straightforward. The 189 at the intersection of “Class 1” and “Is 1” indicates that 189 of the instances that the model predicted to be in Class 1 actually were in that class. The cell at the intersection of “Class 0” and “Is 1” shows that 35 instances were predicted to be 0, but actually were 1. Similarly, the “Is 0” row shows the appropriate counts for instances that actually are 0. Column totals add up to the predicted counts of 0’s and 1’s, row totals sum the totals for the number of actual instances in each class [26].

According to the above confusion matrix, it is calculated that, 35 plus 27 instances are misclassified. The total number of instances that create error is 62. The error rate of the model is easily calculated as 62 over 424 which is the total number of instances that are put into the model. As a result the error rate of the model equals to 0,15.

6. DEFINING CRITICAL SUCCESS FACTORS

After explaining the data mining methodology and related algorithms in the previous sections, this chapter presents the modeling and analysis for defining critical success factors. The methodology will be applied to a success prediction study carried out for the İTU's Industrial Engineering and Management Engineering students.

Data mining projects start with choosing a data mining software tool according to the business needs. SAS tools are used as the data mining software in this study. This tool processes the given data by applying transformations, handling missing values, selecting and grouping the variables. SAS Enterprise Miner (SEM) software is an integrated product that provides business solutions for data mining. Enterprise Miner performs functions such as clustering, associations, generating decision trees. It supports the total data mining process for creating and analyzing a data source.

The steps of the study will be explained by following Generic Data Mining Method explained in Chapter 3.

6.1 Problem Definition

Success analysis is based on a predictive modeling approach aimed to discover the likelihood of the students who are successful. This success model will predict the success propensity of the students who have been at İTU more then one year and have a recent Cumulative Grade Point Average.

There is a target variable for success prediction which indicates if the students are successful or not in the recent academic year. In this section, success prediction model is developed using data gathered by a questionnaire. The information about a student has been successful or not is formed, by using the recent cumulative grade point average of each student, asked during the questionnaire that is applied at the end of fall semester of 2004-2005 academic years. If the students GPA is equal or

over 3 then the student is marked as successful, if GPA is less than 3 then the student is marked as unsuccessful. This modeling process is handled in March 2005.

Success target is a binary variable in the data set indicating either a student has been successful or not at the end of recent academic year. If the student has been successful the value of target variable is "S" for that student and "US" otherwise. The aim of the model is to predict the successful students, because of that, value of "S" for success target is called target event.

In this study, students that are successful and unsuccessful will be mined according to their answers to the questionnaire. Using this information, critical success driving factors will be found out. According to the results obtained at the end of success modeling, necessary inputs for guidance systems of students will be developed. The target object is the student. Furthermore, data is associated on the student base.

6.2 Data Model Used

All the data used in this study is sourced from the answers of the students that have been participated the questionnaire of this thesis. There are 52 questions in the questionnaire and nearly 15 more attributes are generated.

This questionnaire is applied, during the final exams of the 2004-2005 education year's first term. The sophomore, junior and senior students of Industrial and Management Engineering are involved in the survey. 438 of the students gave proper answers to the questionnaire in order to be used during the data mining model study. This data mining study will be based on these 438 different answers. Each answer will form a separate record.

6.3 Sourcing and Preprocessing the Data

Data sourcing and preprocessing comprises the stages of identifying, collecting, filtering and aggregating (raw) data into a format required by the data models and the selected mining functions.

All the paper based questionnaires answers are entered and aggregated in the Microsoft Excel format. The data mining tool that is used, SEM can use Excel based files as input.

In this study, the mere data source is the Excel based file. So, during the study, there are no identifying, collecting and filtering data processes are taken place. But generally, these steps are very hard and time consuming phases.

In this step, the source data is examined and necessary transformations are done. For example, the target variable of Recent_GPA_Success is formed.

6.4 Evaluating the data model

The questions of the questionnaire form the attributes of the data mining model. All of the attributes are examined statistically. Each questions' answers distribution is checked. According to the distributions:

- Some of the attributes are labeled as redundant and omitted from the model.
- Some of the attributes are transformed in to new attributes
- Some of the attributes are discretized.

These steps are explained at the Student Questionnaire, Chapter 4, in detail. Also the correlations of the attributes are searched. And the ones that are highly correlated are not used in the model.

Rejected Variables

In the Table 6.1, all the rejected attributes are shown.

Table 6.1 Rejected Variables

Name	Model Role	Measurement Level	Type
Turkish_Nationality	rejected	binary	Char
Total_Number_Of_Sisters	rejected	ordinal	Num
Total_Number_Of_Sibling	rejected	ordinal	Num
Total_Number_Of_Brothers	rejected	ordinal	Num
Smaller_Sister	rejected	ordinal	Num
Smaller_Brother	rejected	ordinal	Num
Recent_Gpa	rejected	interval	Num
Parents_Status	rejected	nominal	Char
Order_of_Department_Selection	rejected	interval	Num
Mother_Alive	rejected	unary	Char
Marital_Status	rejected	binary	Char
Live_Most_Traffic_Code	rejected	interval	Num
Live_Most_Time	rejected	nominal	Char
High_School_Province_Traffic_Cod	rejected	interval	Num

High_School_Province	rejected	nominal	Char
Graduate_Gpa	rejected	interval	Num
Father_Working	rejected	binary	Char
Child_Ownership	rejected	binary	Char
Birth_Traffic_Code	rejected	interval	Num
Birth_Place	rejected	nominal	Char
Birth_Month	rejected	interval	Num
Bigger_Sister	rejected	ordinal	Num
Bigger_Brother	rejected	ordinal	Num
Age	rejected	ordinal	Num

The reject reasons of the variables one by one are as follows:

Turkish_ Nationality: It does not have a proper distribution and it was not distinguishing.

Total_Number_Of_Sisters: This is a generated variable. There were many classes in the attribute and also there were outlier classes. So, a new variable is created by discretizing.

Total_Number_Of_Sibling: This is a generated variable. There were many classes in the attribute and also there were outlier classes. So, a new variable is created by discretizing.

Total_Number_Of_Brothers: This is a generated variable. There were many classes in the attribute and also there were outlier classes. So, a new variable is created by discretizing.

Smaller_Sister: It does not have a proper distribution, and to handle the outliers it is discretized. A new variable is created. Only the discretized one is used in the model in order not to increase the effect of this variable.

Smaller_Brother: It does not have a proper distribution, and to handle the outliers it is discretized. A new variable is created. Only the discretized one is used in the model in order not to increase the effect of this variable.

Recent_Gpa: The target variable, Recent GPA Success is generated using, Recent Gpa , and these two variables are highly correlated. If Recent_Gpa variable is used in the model, then the model will have just one node.

Parents_Status: It does not have a proper distribution and it was not distinguishing.

Order_of_Department_Selection: It does not have a proper distribution, and to handle the outliers it is discretized. A new variable is created. Only the discretized one is used in the model in order not to increase the effect of this variable.

Mother_Alive: It is a unary variable, so this attribute is automatically eliminated.

Marital_Status: It does not have a proper distribution and it was not distinguishing.

Live_Most_Traffic_Code: It is high cardinality predictor and has many possible values to perform splitting. An experienced data mining specialist will discard those variables.

Live_Most_Time: Using Live_Most_Traffic_Code attribute, the name of the provinces are found out. Namely, Live_Most_Traffic_Code and Live_Most_Time consider same information in different measurement levels.

High_School_Province_Traffic_Cod: It is high cardinality predictor and has many possible values to perform splitting. An experienced data mining specialist will discard those variables.

High_School_Province: Using High_School_Province_Traffic_Cod attribute, the name of the provinces are found out. Namely, High_School_Province_Traffic_Cod and High_School_Province consider same information in different measurement levels.

Graduate_Gpa: This question was asked to the students. But, also, the students Graduate_Gpa expectations are correlated with their recent cumulative GPA's. And also, it is not proper to use this kind of information in a predictive modeling. So, this attribute is omitted from the model.

Father_Working : It does not have a proper distribution and it was not distinguishing.

Child_Ownership: It does not have a proper distribution and it was not distinguishing.

Birth_Traffic_Code : It is high cardinality predictor and has many possible values to perform splitting. An experienced data mining specialist will discard those variables.

Birth_Place: Using Birth_Traffic_Code attribute, the name of the provinces are found out. Namely, Birth_Traffic_Code and Birth_Place consider same information in different measurement levels.

Birth_Month: It is high cardinality predictor and has many possible values to perform splitting. An experienced data mining specialist will discard those variables.

Bigger_Sister: It does not have a proper distribution, and to handle the outliers it is discretized. A new variable is created. Only the discretized one is used in the model in order not to increase the effect of this variable.

Bigger_Brother: It does not have a proper distribution, and to handle the outliers it is discretized. A new variable is created. Only the discretized one is used in the model in order not to increase the effect of this variable.

Age: This attribute has some outliers. Actually, during the iterations, it is seen that, the model is more successful when omitting this attribute.

The input variables of the model are tabulated in Table 6.2

Table 6.2 Input Variables

Name	**Model role**	**Measurement**	**Type**
Student_Id	Id	Interval	Num
Acedemic_Interest	Input	Binary	Char
Arrival Duration	Input	Nominal	Char
Attendance_Perc	Input	Ordinal	Num
Average_Monthly_Expenditure	Input	Nominal	Char
Birth_Region	Input	Nominal	Char
Cinema_Frequency	Input	Nominal	Char
Class	Input	Nominal	Char
Conference_Involvement	Input	Binary	Char
Department	Input	Binary	Char
Father_Alive	Input	Binary	Char
Father_Edu	Input	Nominal	Char
Fee_Credit	Input	Binary	Char
Finding_Job	Input	Binary	Char
Gender	Input	Binary	Char
General_Pleasure_of_ITU	Input	Nominal	Char
High_School_GPA2	Input	Interval	Num
High_School_Region	Input	Nominal	Char
High_School_Type	Input	Nominal	Char
Hobby_Course	Input	Binary	Char
Internet_Hours	Input	Nominal	Char
ITU_Social_Environment	Input	Nominal	Char
Library_Usage	Input	Binary	Char
Live_Most_Region	Input	Nominal	Char
Master_Degree	Input	Nominal	Char
Meetıng_Uni_Friends	Input	Binary	Char

Most_Liked_Course_Group	Input	Nominal	Char
Mother_Edu	Input	Nominal	Char
Mother_Working	Input	Binary	Char
Nonacedemic_Interest	Input	Nominal	Char
Parttime_Working	Input	Binary	Char
Project_Aid	Input	Nominal	Char
Residence	Input	Nominal	Char
Scholarship	Input	Binary	Char
School_Arrival	Input	Nominal	Char
School_Clubs	Input	Binary	Char
Special_Computer_Usage	Input	Binary	Char
Sports	Input	Binary	Char
Studying_Environment	Input	Nominal	Char
Studying_Hours	Input	Nominal	Char
TV_Hours	Input	Nominal	Char
Uni_Entrance_First	Input	Binary	Char
X_Bigger_Brother	Input	Binary	Char
X_Bigger_Sister	Input	Binary	Char
X_Order_Of_Department_Selection	Input	Nominal	Char
X_Smaller_Brother	Input	Binary	Char
X_Smaller_Sister	Input	Binary	Char
X_Total_Number_Of_Brothers	Input	Binary	Char
X_Total_Number_Of_Sibling	Input	Nominal	Char
X_Total_Number_Of_Sisters	Input	Nominal	Char

48 variables are used as input variables in the model. Actually, while the training data set consists of 438 instances, 48 numbers of attributes seems so much for the prediction study of data mining. But when considering the object of this study which is to find out the most effective factors that drive the classification model, all of the variables should be used in the model as much as possible.

6.5 Applying the data mining technique

In this study, the critical success factors of students are aimed to be found out. The mapping of the inputs to the target is a predictive modeling. This objective will be fulfilled by applying a predictive data mining method.

Decision trees are a way of representing a series of rules that lead to a class or value. Student classification analysis aims to identify the background characteristics that indicate the predefined group (successful or unsuccessful) to which each student belongs. It links the value of success in Istanbul Technical University to the background information. Decision tree classification of data mining creates classification models by examining already classified data from a historical data

source and inductively finding a predictive pattern. This pattern can be used to understand the existing data.

Input Data Source

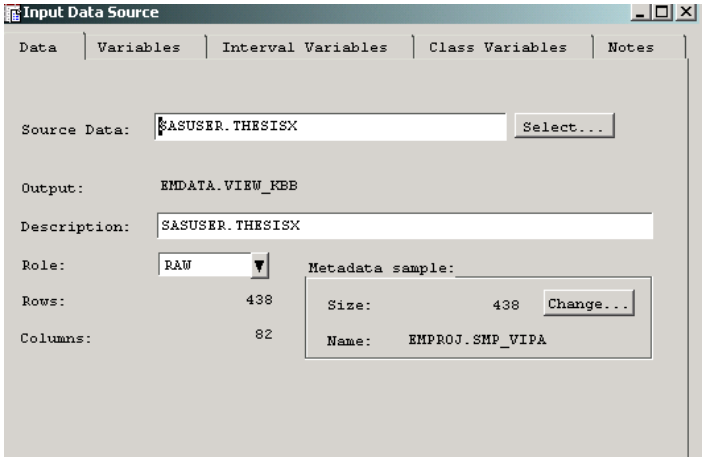


Figure 6.1 Input Data Source

The initial step of a Decision Tree model is to select the data source. In Figure 6.1 the selected data source of the model is seen. As seen above, the size of the file is 438 records. The name of the data source is Sasuser.Thesisx.

Selecting the Splitting Criteria

In the SEM tool, Chi Square test, Gini Reduction and Entropy Reduction can be selected as splitting criteria. These splitting methods are explained in section 5.2.1. Using all three different splitting criteria, with the same input data model and the same stopping rules, three models are generated.

In Figure 6.2 Chi-square test criteria is seen.

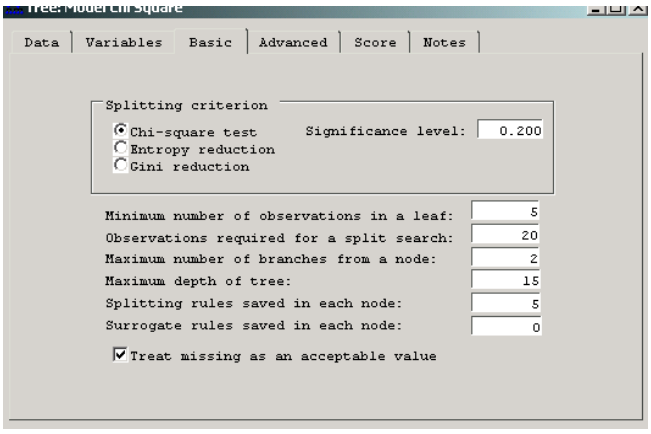


Figure 6.2 Chi Square Test

In Figure 6.3 Gini Reduction is seen.

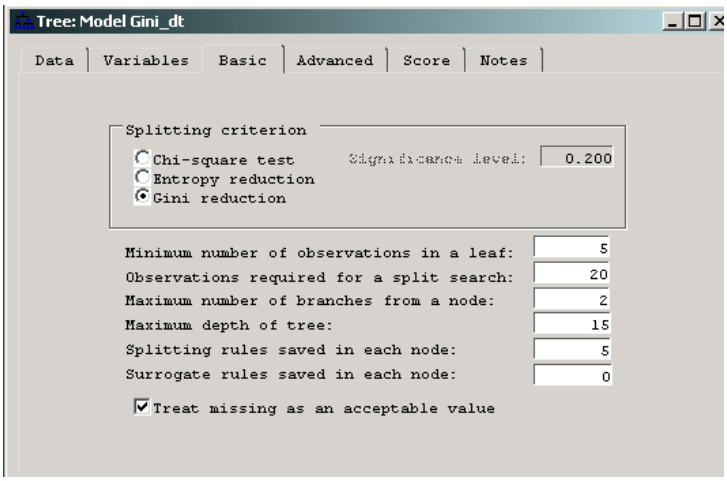


Figure 6.3 Selecting Gini Reduction

In Figure 6.4 Entropy Reduction is seen.

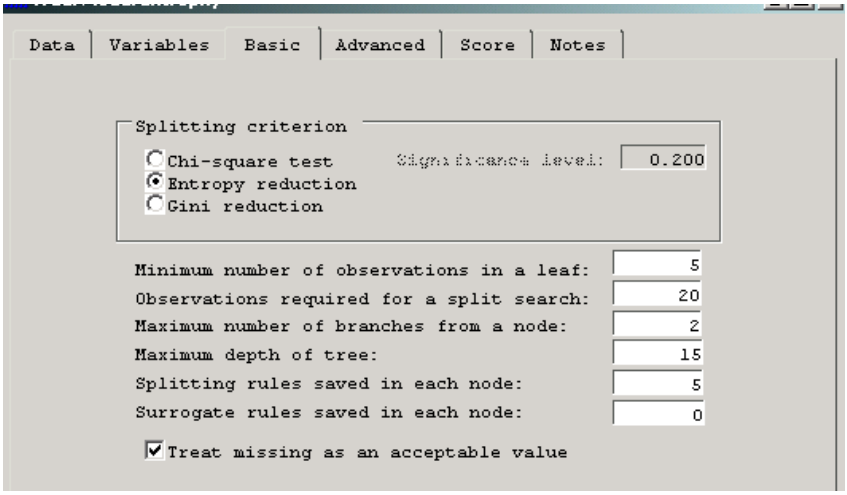


Figure 6.4 Selecting Entropy Reduction

The results of these three models are compared in the Assessment mode. The views of the nodes are seen in Figure 6.5.

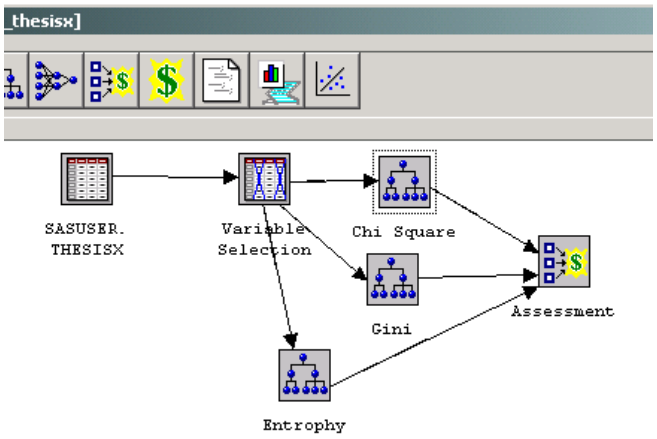


Figure 6.5 View of the Selecting Splitting Criteria Model

In the Assessment node, the three models are selected and the lift charts for the three models are drawn. In Figure 6.6 the comparison lift chart of the models is seen.

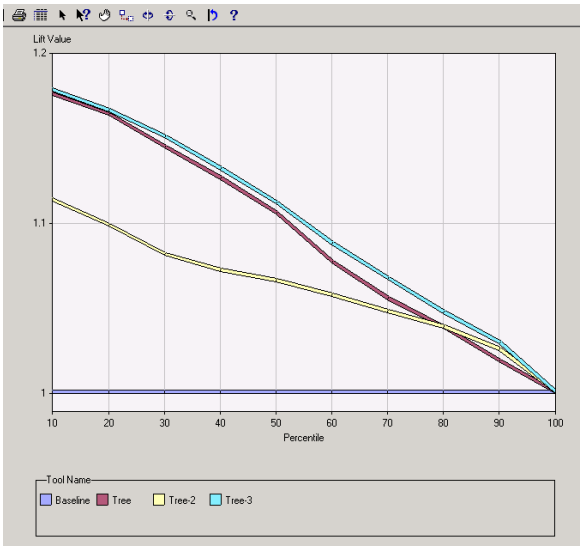


Figure 6.6 The Comparison Lift Chart

According to the chart, Chi Square model lies in the middle. Gini model lift chart is the closest one to the base line. Entropy reduction model lift chart is at the top of the models.

As seen from the chart, the third model, which is Entropy Reduction splitting based model, is chosen to be the best model. So, the modeling studies will be continued using Entropy Reduction splitting criteria.

Stopping Rules

In the basic tab of Tree Node, the stopping rules are determined. While comparing the splitting criteria, the stopping rules are set as:

- Minimum number of observations in a leaf is 5.
- Observations required for a split search is 20.
- Maximum Depth of Tree is 15.

Also there is one more parameter, but it is not a stopping rule.

- Maximum Number of Branch from a Node is 2.

Number of branch defines, how many child nodes, a predecessor node can be split into. So the models that are developed, during deciding the best splitting criteria that suits the data, are binary trees.

Model Development

Entropy splitting is chosen as the best split that suits the data. But while examining this, only binary trees were used. After deciding the best splitting criteria, the following one is to decide, the number of branches of the tree. So setting maximum number of branch 2, 3, and 4, different tree models are developed. The views of the nodes are seen in Figure 6.7.

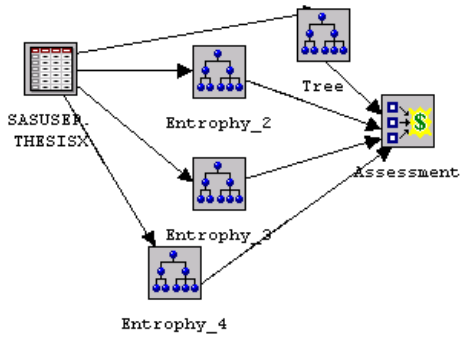


Figure 6.7 View of the Selecting Number of Branches of the Tree

The model named “Entrohy_2” means, Entropy Reduction is chosen as splitting criteria and the number of branches is 2. The model named “Entrohy_3” means, Entropy Reduction is chosen as splitting criteria and the number of branches is 3. The model named “Entrohy_4” means, Entropy Reduction is chosen as splitting criteria and the number of branches is 4. The model named “Tree” means, Chi-square Test is chosen as splitting criteria and the number of branches is 2.

The Chi-Square Test splitting criteria is added to models that are to be measured. Because, during some iterations, the input variables are changed. The resulting input variables are listed in Table 6.2. In Figure 6.8 the comparison lift chart of the four models are seen.

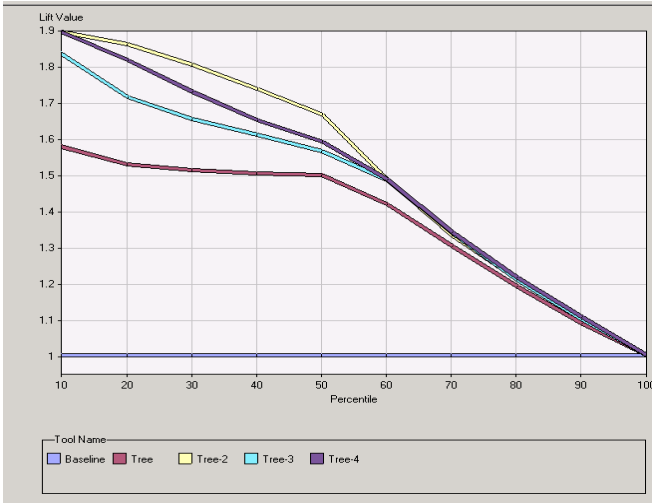


Figure 6.8 The Comparison Lift Chart for Selecting Number of Branches of Tree

The lift chart of the model named Tree is the closest line to the baseline. The lift charts of the models named “Entropy_3” and “Entropy_4” are appeared in the middle part of the Figure 6.8. The lift chart for model “Entropy_2” is the upper one.

It is obvious that the second model which is still “Entropy_2” is the best model. In section 5.3 measures of performance are explained. In this section it is seen that not only the Lift Chart but also, Captured Response and Response charts are used during model assessment. In Figure 6.9 the comparison Response Chart and in Figure 6.10 the comparison Captured Response Chart for the four models are seen.

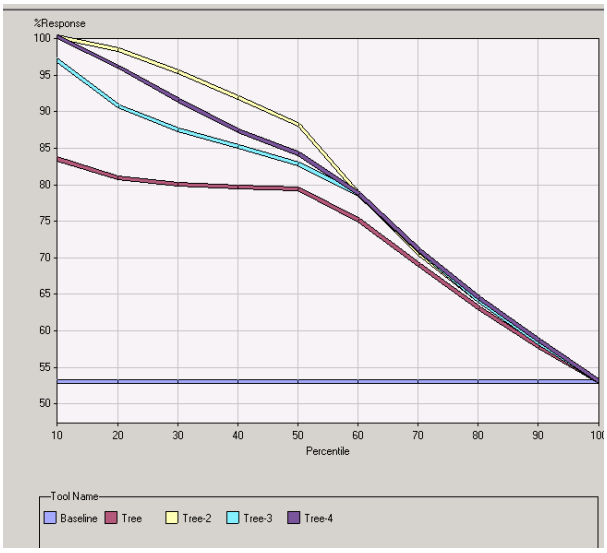


Figure 6.9 The Comparison Response Chart for Selecting Number of Branches of Tree

Lift Chart and the Response Chart look similar to each other.

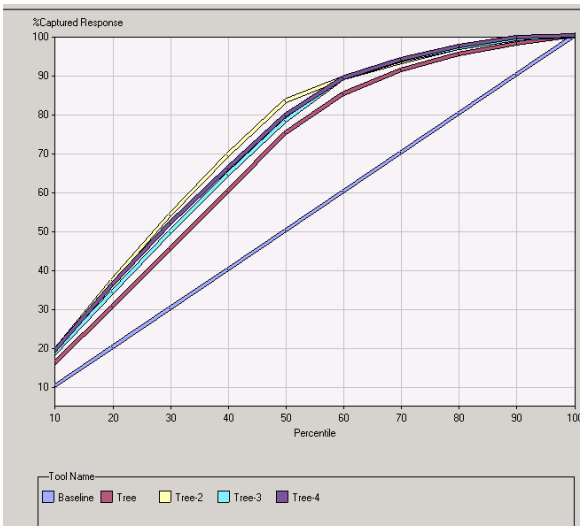


Figure 6.10 The Captured Response Chart for Selecting Number of Branches of Tree

In Figure 6.9 and 6.10 The lift chart for model "Entropy_2" is the upper one. In Figure 6.10, it is again clear to say that the second model is best one, out of the four models. Actually the models look very similar. There are slight differences between the charts. But as a result, the tree of the model named Entropy_2 will be explained briefly.

Result of Decision Tree

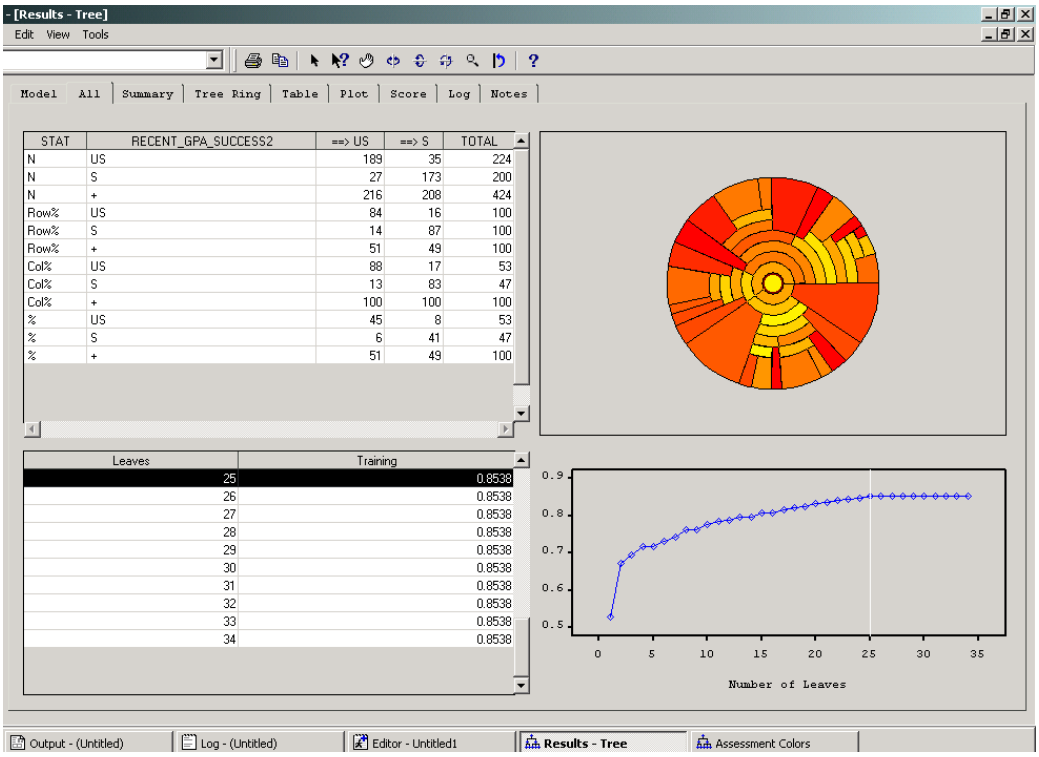


Figure 6.11 The Results Of Selected Model

In Figure 6.11, all the results of the tree are seen combined together in one tab. The selected three has 25 leaves. In Table 6.3 the confusion matrix of the model is seen.

Table 6.3 Confusion Matrix of the Model

STAT	**RECENT_GPA_SUCCESS**	**=== >US**	**=== >S**	**TOTAL**
N	US	189	35	224
N	S	27	173	200
N	+	216	208	424
Row%	US	84	16	100
Row%	S	14	87	100
Row%	+	51	49	100
Col%	US	88	17	53
Col%	S	13	83	47
Col%	+	100	100	100
%	US	45	8	53
%	S	6	41	47
%	+	51	49	100

In the model 424 records are used as input. There are 438 respondents to the questionnaire but only 424 of them have answered the recent GPA question. The target variable of the tree, is based on recent GPA. If a record does not have the recent GPA attribute, then it is omitted from the model. So, 14 missing variables of recent GPA are omitted from the model.

In the data set, 224 of all the students are unsuccessful and 200 of the students are successful. Initially the ratio of the target variable is 200/424 = 0,47 .

The model predicts 216 of the students as unsuccessful. 189 students out of 216 students are actually unsuccessful and predicted as unsuccessful. But that are 27 wrongly predicted students form the unsuccessful class error ratio of the model.

The model predicts 208 of the students as successful. 173 students out of 208 students are actually successful and predicted as successful. But that are 35 wrongly predicted students form the successful class error ratio of the model.

The total error ratio of the model is : (27 +35) / 424 = 0,14. The error of the model is %14. That means the results of the model are right by %86.

In Figure 6.12, the counts of actual and predicted targets can be seen. According to the graphic, visually it can be concluded that the error ratio of the model is low and the model reliability is high.

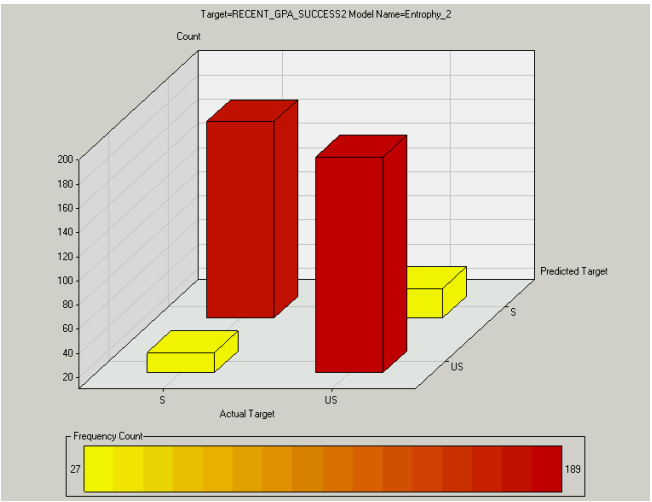


Figure 6.12 The Graph of Confusion Matrix

The model goes in deeper by splitting the nodes, until one of the stopping rules is reached or until the tree can not improve itself. At the end of the model 25 different leaves are attained. During the development phase of the tree, after formation of each leaves, an improvement on the trees' overall purity level is calculated. This is shown in Table 6.5.

Table 6.4 The development of the leaves

Leaves	Overall Correctness Percentage
1	0.5283
2	0.6722
3	0.6958
4	0.7170
5	0.7170
6	0.7311
7	0.7429
8	0.7618
9	0.7618
10	0.7759
11	0.7854
12	0.7877
13	0.7972
14	0.7972
15	0.8066
16	0.8066
17	0.8160
18	**0.8208**
19	0.8255
20	0.8325
21	0.8349
22	0.8420
23	0.8443
24	0.8467
25	**0.8538**
26	0.8538
27	0.8538

The graphical illusion of the Table 6.5 can be seen at Figure 6.13

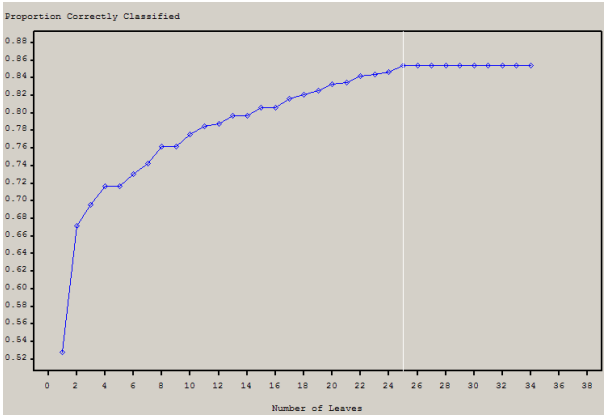


Figure 6.13 The Graph of Proportion Correctly Classified

The decision tree model can be seen in Appendix 1 and all the rules of each leaves can be seen in Appendix 2.

The model's correctness percentage is nearly 86%. But the complexity of the tree gets higher as the number of leaves increases. The tree should be pruned to decrease the complexity and increase the understanding. Pruning is the process of improving the performance of decision tree by removing leaves and branches. A pruned tree is in fact a subset of the full decision tree. To balance the depth of the tree and the error ratio, the best point for pruning is selected. According to Table 6.5, after the 18th leaves, the overall correctness ratio of the tree is 0.8208. That is an enough ratio for the model so the tree can be pruned right after 18th leaves.

Critical Factors Affecting Student Success

According to the resulting tree, the attributes that forms the nodes are the critical factors that affects the students' success. The result of the pruned Decision Tree modeling has 18 different leaves. Each leaf has its own rules. In Table 6.4 the critical factors affecting student success can be found.

In Table 6.4 the column named "Importance" discloses the level of importance of the variable for the model. The Importance value is calculated by the model. The column named "Rules" shows how many times the attribute appear in the tree. If the value is greater than 1 then this means that variable is used at different levels of the tree.

Table 6.5 Critical Factors Affecting Student Success

Name	**Importance**	**Rules**
Attendance_Perc	1.0000	2
Gender	0.6192	1
Residence	0.5022	1
Most_Liked_Course_Group	0.4518	1
Find_Job_Duration	0.3755	2
Birth_Region	0.3745	1
Class	0.3264	1
Scholarship	0.3056	1
Father_Edu	0.2792	1
Studying_Hours	0.2687	1
General_Pleasure_Of_ITU	0.2546	1

Master_Degree	0.2403	1
Father_Alive	0.2362	1
Cinema_Frequency	0.2204	1
Average_Monthly_Expenditure	0.1976	1

424 students answers are involved in this data mining study. There are some examples that similar numbers of record are used to make a data mining study. In Boğaziçi University a resembling Classification analysis is carried on with 500 records of train data set. [17]

7. CONCLUSION

In this study a different approach that employs a distinctive objective is examined. A data mining application took place to find out the critical success factors of the students. As the initial step to figure out the student characteristics, demographic situation, academic background and students personal lives and interests are asked by a student questionnaire. Second, the analysis of the answers are performed. Accordingly the data relevant for data mining is selected. Under the recent cumulative GPA answer of the questionnaire, the successful and unsuccessful students' classes are formed. 438 students answered the questionnaire, since 424 students answered the recent GPA question; only 424 students are involved in the analysis.

The aim of the study is to find out the attributes that yields successful and unsuccessful student classes. Decision Tree technique is the most suitable predictive data mining technique for this aim.

Results of Decision Tree

Considering the results of decision tree the most effective attribute that effects students' success is Attendance Percentage. The root node of the tree is the Attendance Percentage. After applying this attribute, two new child nodes appear. One of the nodes consists of 70 or 80 percentage of attendance. This node contains 253 students. 169 students out of 253 are unsuccessful (66 %). On the other hand, second node consists of 90 or 100 percentage of attendance values. But the succeeding attribute of this branch is Attendance Percentage, again. As a result, a new leaf node occurs. This node consists of 100 percentage of attendance. This node contains 40 students. 36 students out of 40 students are successful (90 %). The other node consists of 90 percentage of attendance. This node contains 131 students. 80 students out of 131 students are successful (61 %). As a result when the Attendance Percentage is one of 70 or 80, the dominant class of the node becomes

“Unsuccessful”. When the Attendance Percentage is one of 90 or 100 the dominant class of the node becomes “Successful”.

The second critical success factor is “Gender”. Gender is the succeeding attribute of the node which consists of 70 or 80 Attendance Percentage. Gender is effective only on the left side of the tree. It is not an effective variable for the right side. But it is applied to 253 students, which is nearly 60% of all the population. As a result, two new child nodes appear. One is formed by male students, 140 students out of 185 students are unsuccessful. The other is formed by female students, 39 out of 68 students are successful. The dominant class of Female node is “Successful”. The dominant class of Male node is “Unsuccessful”.

The third critical success factor is “Residence”. The dominant class of child nodes of this attribute is both “Unsuccessful”. The child nodes do not have to have different dominant classes. But, applying this attribute will result in having more pure nodes in the lower level of the tree for both target classes. This is a very important variable because it has an important effect on the flow of the decision tree. And one of the child nodes is a leaf node with 82.6% purity of unsuccessful class which is formed by 76 out of 92 students.

The fourth critical success factor is “Most Liked Course Group”. If the Attendance Percentage equals 90, the succeeding node’s attribute is “Most Liked Course Group”, applied on 131 students. One of the resulting two nodes is a leaf node. Its’ dominant class is successful by 81.4% purity. The rule of this node is as follows:

IF Most_Liked_Course_Group IS ONE OF: FUNDAMENTAL ENGI
FUNDAMENTAL SCIE
AND Attendance_Perc EQUALS 90

Hence the other node’s selections as Most Liked Course Group are Design or Human and Social Sciences. It is concluded that, if the students’ attendance percentage equals to 90 percentage then their success situation depends on their Most Liked Course Group selection. If Fundamental Engineering or Fundamental Sciences are selected then the students are successful by 81.4 % probability.

The fifth most important factor for the decision tree is “Find Job Duration”. It is seen on both left and right branches of the tree. On the left side of the tree, this attribute is seen after Father Education attribute. It is applied to 20 students. It produces two leaf nodes. One of the leaf nodes’ dominant class is successful by 100 % purity. The

other's dominant class is unsuccessful by 61.5% purity. When the Find Job Duration is one of 12 Months or No Idea then the node's class is successful. When the Find Job Duration is one of 6 months or Immediately then the node's class is unsuccessful. On the right side of the tree, this node is the succeeding node of Most Liked Course Group. If the value of Most Liked Course Group is one of Design or Human and Social Sciences then Find Job Duration node is following. It is applied to 88 students. One of the two resulting nodes is a leaf node. The leaf node's value is No Idea and dominant class is unsuccessful by 86.7% purity. The other resulting node's purity is 58.9% and class is successful.

The sixth most important factor for the decision tree is "Birth Region". Birth Region attribute is applied on 73 students. The resulting two nodes of the attribute are leaf nodes. One of the leaf nodes' dominant class is successful by 73.3 % purity. The other's dominant class is unsuccessful by 64.3% purity. When the Birth Region is one of Akdeniz, Ege or Marmara then the node's class is successful. When the Birth Region is one of DAB, Karadeniz, OAB or ABROAD then the node's class is unsuccessful.

The seventh most important factor for the decision tree is "Class". Class attribute is applied on 65 students. One of the resulting two nodes is a leaf node. The leaf nodes' dominant class is successful by 72.2 % purity. The other's dominant class is unsuccessful by 59% purity. When the Class is 2 then the node's class is successful. When the Class is one of 3 or 4 then the node's class is unsuccessful.

The eighth critical success factor is "Scholarship". This attribute is succeeded after Gender attribute. If the Gender is Female then the succeeding attribute is Scholarship. It is applied to 68 students. One of the resulting two nodes is a leaf node. The leaf nodes' dominant class is unsuccessful by 84.6 % purity. The other's dominant class is successful by 67.3% purity. When the Scholarship is Yes then the node's class is successful. When the Scholarship is No then the node's class is unsuccessful.

The ninth critical success factor is "Father Education". This attribute is succeeded after Average Monthly Expenditure attribute. If the Average Monthly Expenditure is 200-400 mio or 400-600 mio then the succeeding attribute is Father Education. It is applied to 39 students. One of the resulting two nodes is a leaf node. The leaf nodes'

dominant class is unsuccessful by 78.9 % purity. The other's dominant class is successful by 60% purity. When the Father Education is Bachelor then the node's class is successful. When the Father Education is different from Bachelor then the node's class is unsuccessful.

The tenth most important factor for the decision tree is "Studying Hours". Studying Hours attribute is applied on 55 students. One of the resulting two nodes is a leaf node. The leaf nodes' dominant class is successful by 93.8 % purity. The other's dominant class is successful by 56.4% purity. When the Studying Hours is <2 then the node's class is successful by a lower purity. When the Studying Hours is one of 2-4 or >4 then the node's class is successful by a higher purity.

7.1 What has been done similar to this study?

According to the literature, students' success factors were investigated for different undergraduate educations, for primary or high school students.

For example, in a study self-directed or other-directed career choices are examined if they affect academic success. The construct of "other-directed versus self-directed career choice" has existed for quite some time. The current focus of vocational psychologists and counselors has made them question the relevance of this construct for contemporary American society. Many counselors today challenge the assumption that a career choice based on others' expectations is problematic. The findings showed that self-directed or other-directed career choice did not predict academic success. [27]

The relations between achievement motives, achievement goals, and motivational outcomes on a math task were explored in a relational study of Asian American (n=105) and Anglo American (n=98) college students. Students completed pretest questionnaires about their two motives (motive to approach success and fear of failure) and three achievement goals (mastery, performance-approach, performance-avoidance) prior to working on a mathematics task, which was then followed by a post-test questionnaire that assessed students' competence perceptions, interest, and anxiety for the task. Asian American students were found to display on average higher levels of fear of failure, performance-avoidance goals, anxiety, and math performance than Anglo American students. [28]

A study, which is very similar to this, was accomplished in 2004. The primary purpose of that study is to assess the degree to which SAT scores, high-school GPA and class rank predict success in college. Data collected from students enrolled in several sections of Principles of Economics at the University of South Carolina in 2000 and 2001 are used to study the relation between college GPA (the dependent variable) and high-school rank, high-school GPA, and SAT scores (the key independent variables). It is also investigated whether there are race–sex differences in the likelihood of success in college. According to the study nonwhites are less likely than whites, and males are less likely than females, to achieve the 3.0 GPA in college required maintaining their scholarships. [29]

In another study the relationship between emotional intelligence and academic achievement in high school was examined. Students (667) attending a high school in Huntsville, Alabama completed the Emotional Quotient Inventory (EQ-i:YV). At the end of the academic year the EQ-i:YV data was matched with students academic records for the year. When EQ-i:YV variables were compared in groups who had achieved very different levels of academic success (highly successful students, moderately successful, and less successful based on grade-point-average for the year), academic success was strongly associated with several dimensions of emotional intelligence. Results are discussed in the context of the importance of emotional and social competency on academic achievement. When the relationship between academic success and emotional intelligence was examined using the total sample, overall EI was found to be a significant predictor of academic success. [30]

In a study, path analysis was used to test predictions of a model explaining the impact of students' perceptions of classroom structures (tasks, autonomy support and mastery and evaluation) on their self-efficacy, perceptions of the instrumentality of class work, and their achievement goals in a particular classroom setting. Additionally, the impact of self-efficacy, instrumentality, and goals on students' cognitive engagement and achievement was tested. There were 220 high school students who completed a series of questionnaires over a three-month period in their English classes. Data strongly supported the model demonstrating that student perceptions of classroom structures are important for their motivation. Also supported was the importance of perceiving the current class work as being instrumental for future success.[31]

In Australia, one of the most noticeable changes to the university student profile over the last decade is the increasing number of mature-age students. This study examined the relationship between previous academic performance, psychological characteristics of the student, learning strategies, and the first year academic performance of school leavers and mature-age students through structural equation modeling. A total of 1193 first year university students completed a questionnaire at the beginning of their first year of study, and provided consent for their academic results to be tracked over their first year at university. While the overall model of success appeared to fit for both groups, differences were identified in the order of importance of constructs affecting achievement and the strength of some relationships. The most obvious difference occurred in the relationship between previous academic performance, self-reported learning strategies, and achievement in first semester, with previous performance a more accurate predictor of school leavers' performance and self-reported learning strategies a more accurate predictor of mature-age students' performance. This study has implications for the university in terms of the targeting of academic support services for students of different ages. [32]

The present study was designed to investigate the chain of associations between parenting behavior and early adolescents' school success. Students' goal orientations and classroom behavior were hypothesized to mediate between parenting and school success. The sample consisted of 327 pre university- tracked pupils in their first year of secondary school. Results indicate that boys and girls shared the same pathway from maternal disciplinary strategies to school success mediated by the child's goal orientations and cognitive classroom engagement. Path analyses revealed moderate associations between parenting and goal orientations. Goal orientations were found to be moderately linked to classroom behavior dimensions conducive to school success. Although models for boys and girls differed slightly, overall results highlight the continuing relationship in early adolescence between parenting and pupils' beneficent academic behavior. The present study highlights several processes by which parents might shape their early adolescents' school success. [33]

A study is performed to determine if any existent preadmission academic or personal variables predict academic success in the first year of the Palmer College of

Chiropractic West (PCCW) program. One hundred ninety-two students at PCCW who had completed the first year of the program are the participants of the study. One-way analysis of variance and stepwise linear multiple regression methods are applied in the study. Student characteristics on entering PCCW may help predict student performance in the first academic year. A relatively strong and statistically significant prediction model for Year 1, GPA exists for PCCW. Used in conjunction with other available empirical data, this regression model may allow the institution to make more informed decisions when selecting students for admission. [34]

There has been an increasing show of interest in the different ways in which individuals approach and respond to challenges in academic settings, such as school and university. On the one hand, it has been assumed that students_ achievement motivation and related strategies influence their academic performance and related satisfaction. Fear of failure and showing a low level of effort and a high level of task-avoidance are likely to lead to failure, whereas optimism, trying hard, and concentrating on the task at hand increase the likelihood of success. On the other hand, it might be assumed that academic achievement and related feedback provide a basis for the kinds of achievement-related goals, beliefs, and strategies individuals turn to later on. Repeated failures and low achievement increase anxiety and lead to task-avoidance, whereas academic success leads to active ways of dealing with future academic challenges. [35]

The relationship between the Big Five personality traits, cognitive ability, and beliefs about intelligence was explored in a longitudinal study using a sample (N = 93) of British university students. These three sets of variables were used to predict academic performance (i.e., examination grades) as well as seminar performance (i.e., behavior in class, essay marks, and attendance record) aggregated over a 2-year period. This well-established questionnaire is a 240-item measure of the Big Five personality factors: Neuroticism, Extraversion, and Openness to Experience, Agreeableness, and Conscientiousness. Items involve questions about typical behaviors or reactions, which are answered on a five-point Likert scale. Correlational analyses showed that personality (but not intelligence) was related to beliefs about intelligence (specifically entity vs. incremental beliefs). More conscientious participants were more likely to think that intelligence can be increased throughout the life span; whilst low conscientious individuals were more likely to believe that

intelligence is stable. However, these beliefs were not themselves significantly related to academic performance; only personality traits (Conscientiousness positively, Extraversion negatively) and gender were significantly correlated with academic performance. Further, following a series of hierarchical regression, it was shown that the Big Five personality traits are better predictors of academic performance than cognitive ability, beliefs about intelligence, and gender. When seminar performance indicators were regressed onto these variables, a similar pattern was obtained: Personality was the most powerful predictor of absenteeism, essay marks, and behavior in seminar classes (as rated by different tutors), with Conscientiousness being the most significant predictor. [36]

7.2 Recommendations for Following Studies

The fundamentals of this study are the questions that are asked in the questionnaire. So, making improvements on the questionnaire will result in better outputs of the data mining study. New questions should be included in the scope of the survey. According to the above mentioned studies, collecting different variables from the students can be suggested for further studies. New questions should be added to the questionnaire. For example, whether career choices of the students were self-directed or other-directed should be asked. Since, the motives of the students should affect their success levels; it should take part in the questionnaire. High school rank and university examination scores should be asked to the students. A question regarding emotional intelligence should take part in the questionnaire. Student perceptions of classroom activities should be asked.

Some of the current questions should be omitted from the questionnaire. Father working status, students' nationality, marital status, child ownership, expected graduation GPA variables are not used in the data mining study.

The target population of the study is second, third and fourth class students of the Industrial and Management Engineering students. The total number of target population is 570 students. 438 students out of 570 students are involved to the survey. It is a fact that 76% of all the population are taken part in the study. It is a very sufficient ratio. But, it can be concluded that the higher the number of involved students yield in the better results of the data mining study. So, participant of more number of students should be provided for the questionnaire.

REFERENCES

[1] **Ransdell, S.,** 2001. Predicting college success: the importance of ability and non-cognitive variables, *International Journal of Educational Research*, **35**, 357 – 364.

[2**] Nurm ,J., and Aunola K.,** 2001. How does academic achievement come about: cross-cultural and methodological notes, *International Journal of Educational Research*, **35**, 403–409.

[3] **Chye, K. and Gerry, K.**, 2002. Data mining and customer relations in the banking industry, *Singapore Management Review*, **24**, 1-27

[4] **Berry, M. J. A. and Linoff, G.S.,** 2000. Mastering Data Mining: The Art and Science of Customer Relationship Management, John Wiley & Sons, New York.

[5] **Bashein, B.J., and Markus, M.L.,** 2000. Data Warehouses, More Than Just Mining, Financial Executives Resources Foundation, USA.

[6] http://www.data-mining-software.com/data_mining_history.htm, 2005.

[7] http://eot.apac.edu.au/modules.php?name=Content&pa=showpage&pid=5, 2005.

[8] **Gianotti, F. and Pedreschi, D.,** 2000. Knowledge Discovery & Data Mining, Pisa KDD Lab, http://www-kdd.di.unipi.it.

[9] **Newing, R.,** 1996. Data Mining, *Management Accounting: Magazine for Chartered Management Accountants*, **74**, 34-38

[10] **Brachman, J., Khabaza, T., Kloesgen, W., and Simoudis, E.,** 1996. Mining Business Databases, Communications of the ACM. Vol **39**, no.11, 42-48

[11] **Bayram, E.,** 2001. Customer Segmentation and Churn Modeling In Wireless Communications, *Master of Science Thesis*, Boğaziçi University Institute for Graduate Studies in Science and Engineering.

[12] **Büyükakın, A.,** 2005. Data Mining with Fuzzy Logic, *Master of Science Thesis*, İstanbul Technical University Science Institute.

[13] **Bozdogan, H.,** 2004. Statistical Data Mining and Knowledge Discovery, Chapman-Hall, Florida.

[14] **Chan, C. and Lewis, B.,** 2002. A Basic Premier on Data Mining, Information Systems Implementations, Morgan Kaufmann, San Diego.Management, 56-60

[15] **Han, J., & Kamber, M**, 2001. Data mining: Concepts and techniques, Morgan Kaufmann, SanFrancisco.

[16] **Chen, Y., Hsu , C., Chou, S**., 2003. Constructing a multi-valued and multi-labeled decision tree, *Expert Systems with Applications*, **25**, 199–209.

[17] **Güvenç, E.,** 2001. Student Performance Assessment In Higher Education Using Data Mining, *Master of Science Thesis*, Boğaziçi University Institute for Graduate Studies in Science and Engineering.

[18] **Baragoin, C., Andersen, C.M., Bayerl, S. and Schommer, C.,** 2001. Mining Your Own Business in Banking Using DB2 Intelligent Miner for Data, International Business Machines Corporation, California

[19] **Çerkez, H.S.,** 2003. Business Intelligence and Data Mining in Customer Relationship Management, *Master of Science Thesis*, İstanbul Technical University Science Institute.

[20] **Yang, C., Prasher, S.O., Enright, P., and Burgess, M.,** 2003. Application of decision tree technology for image classification using remote sensing data, *Agricultural Systems*, **76**,1101–1117

[21] **Walsh, S., Potts, W., and Wielenga, D.,** 2002. Applying Data Mining Techniques Using Enterprise Miner, SAS Institute Inc., Cary.

[22] **Padraic, G.,** 1999. Trees for Predictive Modeling, SAS Institute Inc., Neville.

[23] **Potts, W.,** 2001. Decision Tree Modeling, SAS Institute Inc., Neville.

[24] **Berson, A. and Smith, S. J.,** 1997. Data Warehousing, Data Mining and OLAP, McGraw-Hill, New York

[25] **Witten, H.I., and Frank, E.,** 2000. Data Mining Practical Machine Learning Tools and Techniques with Java, Morgan Kaufmann, San Francisco.

[26] **Pyle, D.,** 2003. Business Modeling and Data Mining, Morgan Kaufmann, San Francisco.

[27] **Rehfussö, M.C. and Borges, N.J.,** 2005. Using narratives to explore other-directed occupational choice and academic success, *Journal of Vocational Behavior*, Article in Press.

[28] **Zusho, A., Pintrich, P. R., and Cortina, K.,S.,** 2005. Motives, goals, and adaptive patterns of performance in Asian American and Anglo American students, *Learning and Individual Differences*, **15**, 141-158.

[29] **Cohn, E., Cohn, S., Balch, D.C. and Bradley J.,** 2004. Determinants of Undergraduate GPAs: SAT Scores, High-School GPA and High-School Rank, *Economics of Education Review*, **23**, 577–586.

[30] **Parker, J. D.A. , Creque R. E. and Barnhart , D. L.,** 2004. Academic achievement in high school: does emotional intelligence matter?, *Personality and Individual Differences*, **37**, 1321–1330.

[31] **Greene, B. A., Miller, R. B. and Crowson, H. M.,** 2004. Predicting high school students' cognitive engagement and achievement: Contributions of classroom perceptions and motivation, *Contemporary Educational Psychology*, **29**, 462–482.

[32] **McKenzie, K.and Gow, K.,** 2004. Exploring The First Year Academic Achievement Of School Leavers And Mature-Age Students Through Structural *Equation Modelling, Learning and Individual Differences*, **14**, 107–123.

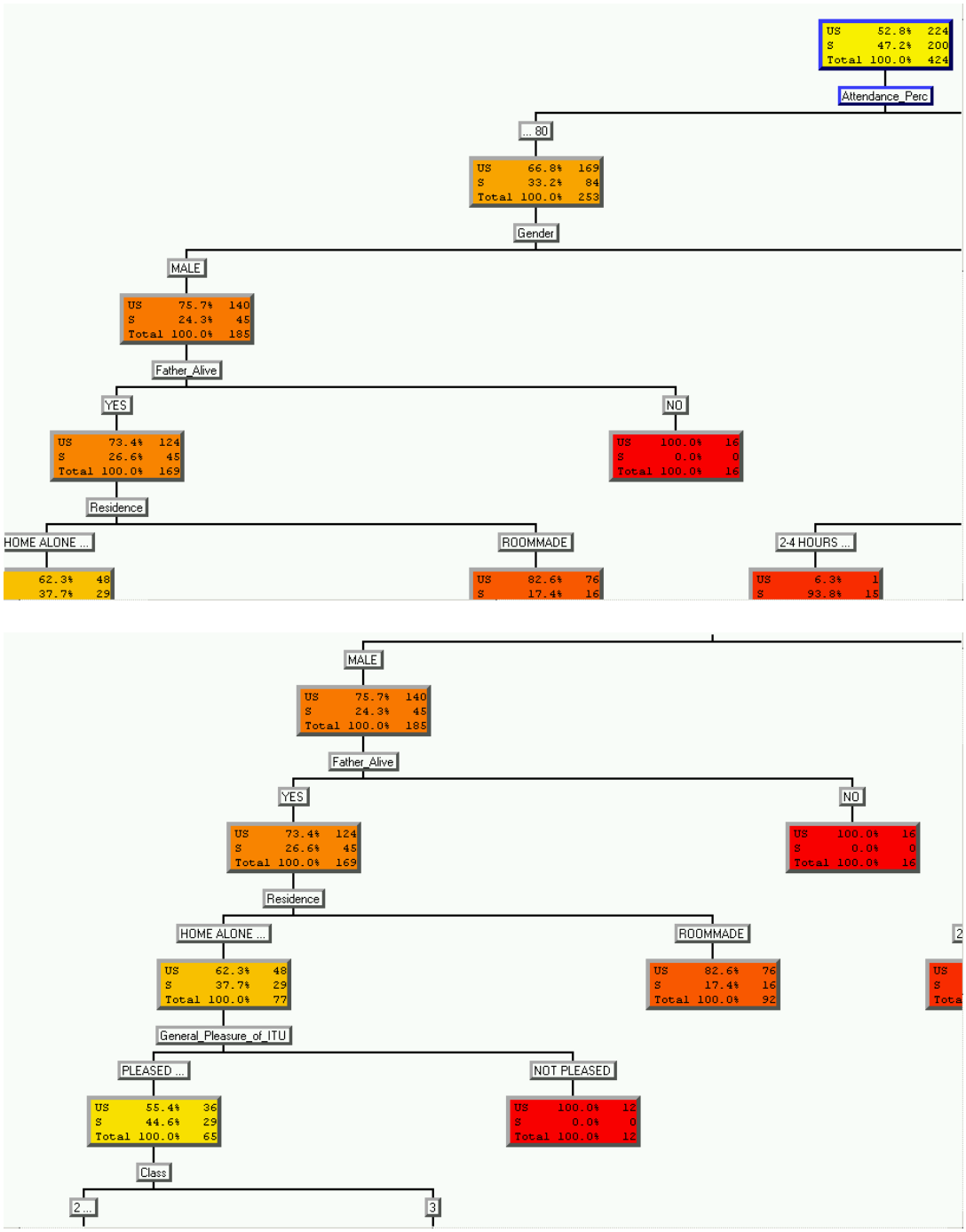
[33] **Bruyn, E. H., Dekovic , M. and Meijnen, G. W.,** 2003. Parenting, Goal Orientations, Classroom Behavior, And School Success In Early Adolescence, *Applied Developmental Psychology*, **24**, 393–412.

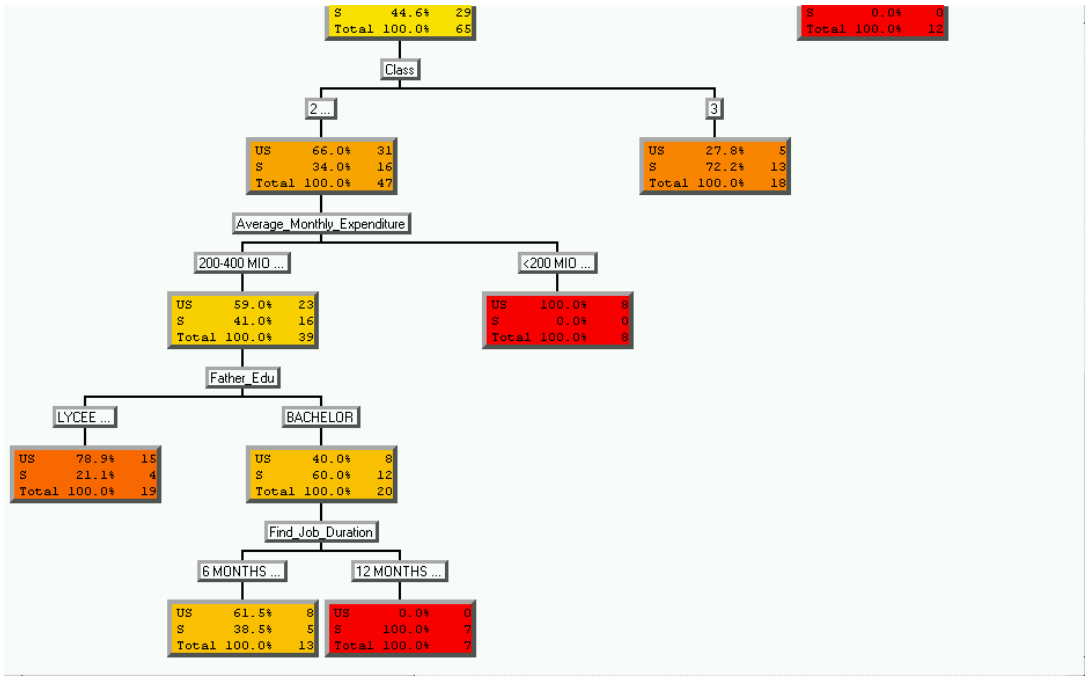
[34] **Green, B. N., Johnson, C. D., and McCarthy, K.,** 2003. Predicting Academic Success In The First Year Of Chiropractic College, *Journal of Manipulative and Physiological Therapeutics*, **26**, 40-46.

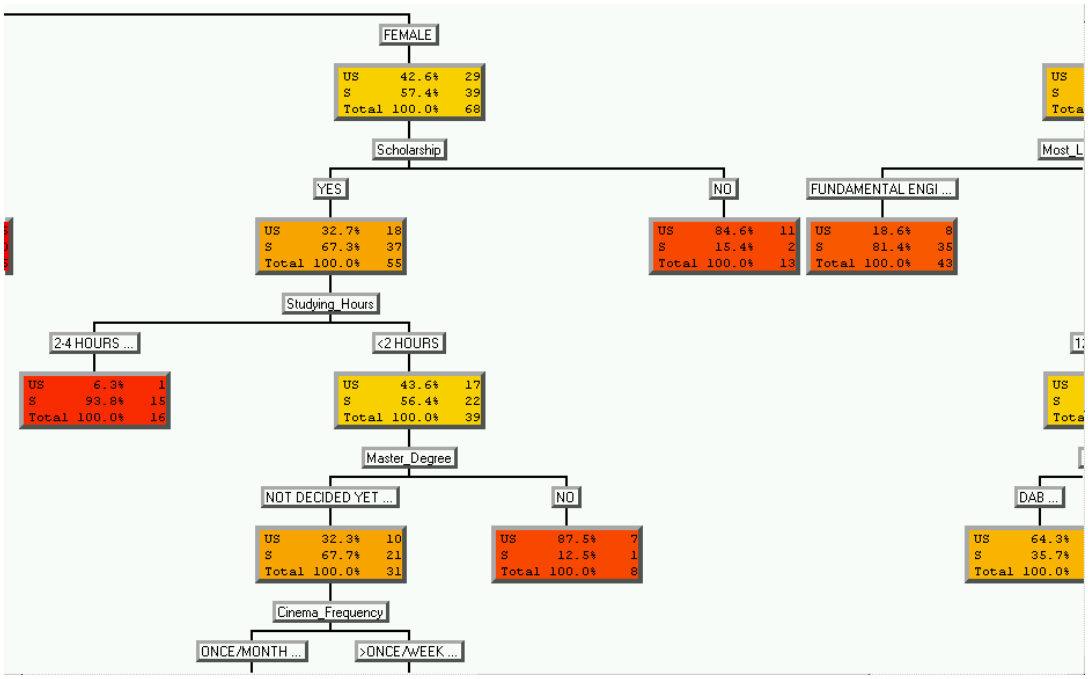
[35] **Nurmi , J. E., Aunola, K. , Katariina, and Lindroos, M.,** 2003. The role of success expectation and task-avoidance in cademic performance and satisfaction: Three studies on antecedents, consequences and correlates, *Contemporary Educational Psychology*, **28**, 59–90.

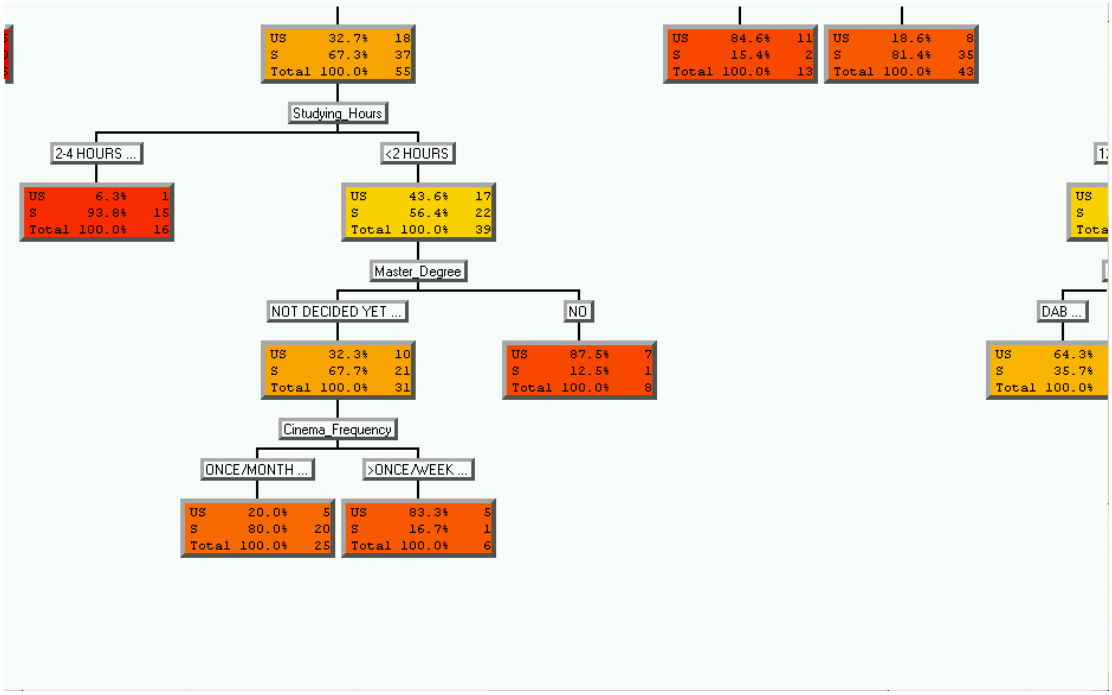
[36] **Furnham, A., Chamorro-Premuzic, T.and McDougall, F.,** 2003. Personality, cognitive ability, and beliefs about intelligence as predictors of academic performance, *Learning and Individual Differences*, **14**, 49–66.

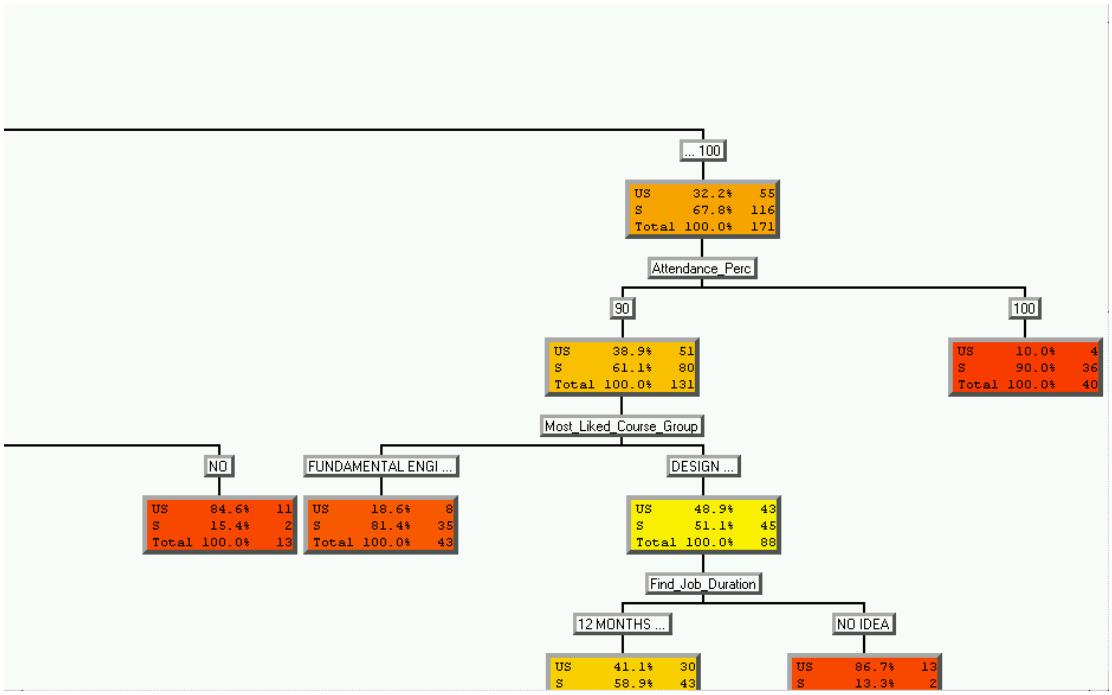
APPENDIXA : THE RESULTING DECISION TREE OF THE MODEL

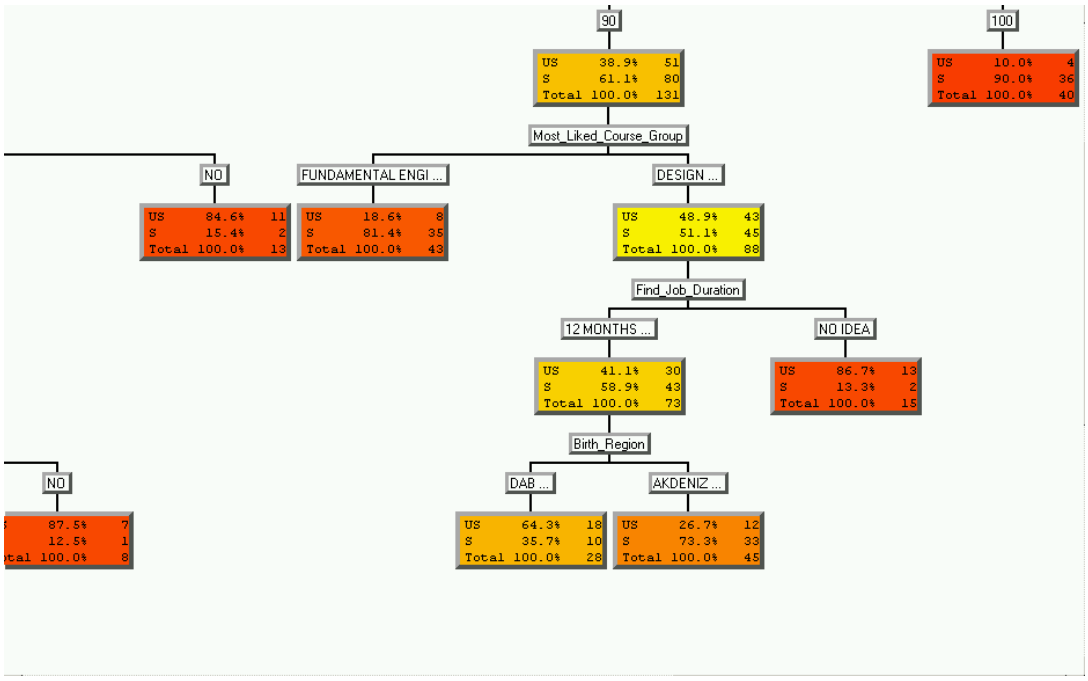



S 44.6% 29
Total 100.0% 65
S 0.0% 0
Total 100.0% 12
Class
2 ...
3
US 66.0% 31
S 34.0% 16
Total 100.0% 47
US 27.8% 5
S 72.2% 13
Total 100.0% 18
Average_Monthly_Expenditure
200-400 MIO ...
<200 MIO ...
US 59.0% 23
S 41.0% 16
Total 100.0% 39
US 100.0% 8
S 0.0% 0
Total 100.0% 8
Father_Edu
LYCEE ...
BACHELOR
US 78.9% 15
S 21.1% 4
Total 100.0% 19
US 40.0% 8
S 60.0% 12
Total 100.0% 20
Find_Job_Duration
6 MONTHS ...
12 MONTHS ...
US 61.5% 8
S 38.5% 5
Total 100.0% 13
US 0.0% 0
S 100.0% 7
Total 100.0% 7


FEMALE
US 42.6% 29
S 57.4% 39
Total 100.0% 68
Scholarship
YES
NO
FUNDAMENTAL ENGI ...
US 32.7% 18
S 67.3% 37
Total 100.0% 55
US 84.6% 11
S 15.4% 2
Total 100.0% 13
US 18.6% 8
S 81.4% 35
Total 100.0% 43
Studying_Hours
2-4 HOURS ...
<2 HOURS
US 6.3% 1
S 93.8% 15
Total 100.0% 16
US 43.6% 17
S 56.4% 22
Total 100.0% 39
Master_Degree
NOT DECIDED YET ...
NO
DAB ...
US 32.3% 10
S 67.7% 21
Total 100.0% 31
US 87.5% 7
S 12.5% 1
Total 100.0% 8
US 64.3%
S 35.7%
Total 100.0%
Cinema_Frequency
ONCE/MONTH ...
>ONCE/WEEK ...


US 32.7% 18
S 67.3% 37
Total 100.0% 55
Studying_Hours
2-4 HOURS ...
<2 HOURS
US 6.3% 1
S 93.8% 15
Total 100.0% 16
US 43.6% 17
S 56.4% 22
Total 100.0% 39
Master_Degree
NOT DECIDED YET ...
NO
US 32.3% 10
S 67.7% 21
Total 100.0% 31
US 87.5% 7
S 12.5% 1
Total 100.0% 8
Cinema_Frequency
ONCE/MONTH ...
>ONCE/WEEK ...
US 20.0% 5
S 80.0% 20
Total 100.0% 25
US 83.3% 5
S 16.7% 1
Total 100.0% 6
US 84.6% 11
S 15.4% 2
Total 100.0% 13
US 18.6% 8
S 81.4% 35
Total 100.0% 43
DAB ...
US 64.3%
S 35.7%
Total 100.0%


... 100
US 32.2% 55
S 67.8% 116
Total 100.0% 171
Attendance_Perc
90
100
US 38.9% 51
S 61.1% 80
Total 100.0% 131
US 10.0% 4
S 90.0% 36
Total 100.0% 40
Most_Liked_Course_Group
NO
FUNDAMENTAL ENGI ...
DESIGN ...
US 84.6% 11
S 15.4% 2
Total 100.0% 13
US 18.6% 8
S 81.4% 35
Total 100.0% 43
US 48.9% 43
S 51.1% 45
Total 100.0% 88
Find_Job_Duration
12 MONTHS ...
NO IDEA
US 41.1% 30
S 58.9% 43
US 86.7% 13
S 13.3% 2


90
US 38.9% 51
S 61.1% 80
Total 100.0% 131
100
US 10.0% 4
S 90.0% 36
Total 100.0% 40
Most_Liked_Course_Group
NO
US 84.6% 11
S 15.4% 2
Total 100.0% 13
FUNDAMENTAL ENGI ...
US 18.6% 8
S 81.4% 35
Total 100.0% 43
DESIGN ...
US 48.9% 43
S 51.1% 45
Total 100.0% 88
Find_Job_Duration
12 MONTHS ...
US 41.1% 30
S 58.9% 43
Total 100.0% 73
NO IDEA
US 86.7% 13
S 13.3% 2
Total 100.0% 15
Birth_Region
NO
87.5% 7
12.5% 1
tal 100.0% 8
DAB ...
US 64.3% 18
S 35.7% 10
Total 100.0% 28
AKDENIZ ...
US 26.7% 12
S 73.3% 33
Total 100.0% 45

APPENDIXB : THE RULES OF EACH NODE

IF Attendance_Perc EQUALS 100
THEN
NODE : 7
N : 40
US : 10.0%
S : 90.0%

IF Father_Alive EQUALS NO
AND Gender EQUALS MALE
AND Attendance_Perc IS ONE OF: 70 80
THEN
NODE : 9
N : 16
US : 100.0%
S : 0.0%

IF Scholarship EQUALS NO
AND Gender EQUALS FEMALE
AND Attendance_Perc IS ONE OF: 70 80
THEN
NODE : 11
N : 13
US : 84.6%
S : 15.4%

IF Most_Liked_Course_Group IS ONE OF: FUNDAMENTAL ENGI
FUNDAMENTAL SCIE
AND Attendance_Perc EQUALS 90
THEN
NODE : 12
N : 43
US : 18.6%
S : 81.4%

IF Residence EQUALS ROOMMADE
AND Father_Alive EQUALS YES
AND Gender EQUALS MALE
AND Attendance_Perc IS ONE OF: 70 80
THEN
NODE : 17
N : 92
US : 82.6%

S : 17.4%

IF Studying_Hours IS ONE OF: 2-4 HOURS >4 HOURS
AND Scholarship EQUALS YES
AND Gender EQUALS FEMALE
AND Attendance_Perc IS ONE OF: 70 80
THEN
NODE : 18
N : 16
US : 6.3%
S : 93.8%

IF Find_Job_Duration EQUALS NO IDEA
AND Most_Liked_Course_Group IS ONE OF: DESIGN HUMAN AND SOCIAL
AND Attendance_Perc EQUALS 90
THEN
NODE : 23
N : 15
US : 86.7%
S : 13.3%

IF General_Pleasure_of_ITU EQUALS NOT PLEASED
AND Residence IS ONE OF: HOME ALONE HOSTEL OWN FAMILY WITH A FAMILY
AND Father_Alive EQUALS YES
AND Gender EQUALS MALE
AND Attendance_Perc IS ONE OF: 70 80
THEN
NODE : 27
N : 12
US : 100.0%
S : 0.0%

IF Master_Degree EQUALS NO
AND Studying_Hours EQUALS <2 HOURS
AND Scholarship EQUALS YES
AND Gender EQUALS FEMALE
AND Attendance_Perc IS ONE OF: 70 80
THEN
NODE : 31
N : 8
US : 87.5%
S : 12.5%

IF Birth_Region IS ONE OF: DAB KARADENIZ OAB YURTDISI
AND Find_Job_Duration IS ONE OF: 12 MONTHS 6 MONTHS IMMEDIATELY
AND Most_Liked_Course_Group IS ONE OF: DESIGN HUMAN AND SOCIAL
AND Attendance_Perc EQUALS 90
THEN
NODE : 34

N : 28
US : 64.3%
S : 35.7%

IF Birth_Region IS ONE OF: AKDENIZ EGE MARMARA
AND Find_Job_Duration IS ONE OF: 12 MONTHS 6 MONTHS IMMEDIATELY
AND Most_Liked_Course_Group IS ONE OF: DESIGN HUMAN AND SOCIAL
AND Attendance_Perc EQUALS 90
THEN
NODE : 35
N : 45
US : 26.7%
S : 73.3%

IF Class EQUALS 3
AND General_Pleasure_of_ITU IS ONE OF: PLEASED VERY PLEASED
AND Residence IS ONE OF: HOME ALONE HOSTEL OWN FAMILY WITH A FAMILY
AND Father_Alive EQUALS YES
AND Gender EQUALS MALE
AND Attendance_Perc IS ONE OF: 70 80
THEN
NODE : 37
N : 18
US : 27.8%
S : 72.2%

IF Cinema_Frequency IS ONE OF: ONCE/MONTH ONCE/WEEK
AND Master_Degree IS ONE OF: NOT DECIDED YET YES
AND Studying_Hours EQUALS <2 HOURS
AND Scholarship EQUALS YES
AND Gender EQUALS FEMALE
AND Attendance_Perc IS ONE OF: 70 80
THEN
NODE : 42
N : 25
US : 20.0%
S : 80.0%

IF Cinema_Frequency IS ONE OF: >ONCE/WEEK NEVER
AND Master_Degree IS ONE OF: NOT DECIDED YET YES
AND Studying_Hours EQUALS <2 HOURS
AND Scholarship EQUALS YES
AND Gender EQUALS FEMALE
AND Attendance_Perc IS ONE OF: 70 80
THEN
NODE : 43
N : 6
US : 83.3%
S : 16.7%

IF Average_Monthly_Expenditure IS ONE OF: <200 MIO >600 MIO
AND Class IS ONE OF: 2 4
AND General_Pleasure_of_ITU IS ONE OF: PLEASED VERY PLEASED
AND Residence IS ONE OF: HOME ALONE HOSTEL OWN FAMILY WITH A FAMILY
AND Father_Alive EQUALS YES
AND Gender EQUALS MALE
AND Attendance_Perc IS ONE OF: 70 80
THEN
 NODE   :   49
 N   :   8
 US   : 100.0%
 S   :   0.0%

IF Father_Edu IS ONE OF: LYCEE MASTER PRIMARY SECONDARY ÖNLISANS
AND Average_Monthly_Expenditure IS ONE OF: 200-400 MIO 400-600 MIO
AND Class IS ONE OF: 2 4
AND General_Pleasure_of_ITU IS ONE OF: PLEASED VERY PLEASED
AND Residence IS ONE OF: HOME ALONE HOSTEL OWN FAMILY WITH A FAMILY
AND Father_Alive EQUALS YES
AND Gender EQUALS MALE
AND Attendance_Perc IS ONE OF: 70 80
THEN
 NODE   :   60
 N   :   19
 US   : 78.9%
 S   : 21.1%

IF Find_Job_Duration IS ONE OF: 6 MONTHS IMMEDIATELY
AND Father_Edu EQUALS BACHELOR
AND Average_Monthly_Expenditure IS ONE OF: 200-400 MIO 400-600 MIO
AND Class IS ONE OF: 2 4
AND General_Pleasure_of_ITU IS ONE OF: PLEASED VERY PLEASED
AND Residence IS ONE OF: HOME ALONE HOSTEL OWN FAMILY WITH A FAMILY
AND Father_Alive EQUALS YES
AND Gender EQUALS MALE
AND Attendance_Perc IS ONE OF: 70 80
THEN
 NODE   :   66
 N   :   13
 US   : 61.5%
 S   : 38.5%

IF Find_Job_Duration IS ONE OF: 12 MONTHS NO IDEA
AND Father_Edu EQUALS BACHELOR
AND Average_Monthly_Expenditure IS ONE OF: 200-400 MIO 400-600 MIO

AND Class IS ONE OF: 2 4
AND General_Pleasure_of_ITU IS ONE OF: PLEASED VERY PLEASED
AND Residence IS ONE OF: HOME ALONE HOSTEL OWN FAMILY WITH A FAMILY
AND Father_Alive EQUALS YES
AND Gender EQUALS MALE
AND Attendance_Perc IS ONE OF: 70 80
THEN
NODE : 67
N : 7
US : 0.0%
S : 100.0%

APPENDIXC : ORIGINAL QUESTIONNAIRE

Öğrenci Numarası : .....

Bölüm : .....

Sınıf : .....

DEMOGRAFİK SORULAR

1. Doğum Tarihi (Yıl- Ay-Gün formatında yazın lütfen)

   -- -- \ -- -- \ -- --

2. Cinsiyetiniz?

   Bayan ☐ Bay ☐

3. Medeni durumunuz?

   Evli ☐ Bekar ☐

4. Çocuk sahibi misiniz?

   Evet ☐ Hayır ☐

5. Anneniz hayatta mı?

   Evet ☐ Hayır ☐

6. Babanız hayatta mı?

   Evet ☐ Hayır ☐

7. Doğum Yeriniz il ismi ve il plaka kodu (Örn : İstanbul / 34)

   ---------- / -- --

8. En uzun süre yaşadığınız il ismi ve il plaka kodu

---------- / -- --

9. Lise öğrenimi aldığınız il ismi ve il plaka kodu

---------- / -- --

10. Annenizin Eğitim Durumu

Tahsili yok	
İlk okul	
Orta okul	
Lise	
Önlisans	
Lisans	
Yüksek lisans	
Doktora	

11. Babanızın Eğitim Durumu

Tahsili yok	
İlk okul	
Orta okul	
Lise	
Önlisans	
Lisans	
Yüksek lisans	
Doktora	

12. Kendinizden büyük Erkek / Kız kardeş sayınız nedir?

Erkek ..... Kız.....

13. Kendinizden küçük Erkek / Kız kardeş sayınız nedir?

Erkek ..... Kız.....

14. Türkiye Cumhuriyeti vatandaşı mısınız?

Evet ☐ Hayır ☐

15. Anneniz çalışıyor mu / çalıştı mı?

Evet ☐ Hayır ☐

16. Babanız çalışıyor mu / çalıştı mı?

Evet ☐ Hayır ☐

17. Anneniz ve Babanız beraberliği için şu anda hangisi geçerli? (Anne – Baba hayatta ise cevaplayın)

Evli	
Boşanmış	
Ayrı yaşıyorlar	

EĞİTİM

18. Mezun olduğunuz lise türü nedir?

Devlet lisesi	
Meslek lisesi	
Ticaret lisesi	
Anadolu lisesi	
Fen lisesi	
Özel lise	
Yurt dışı	
Devlet lisesi	

19. Lise mezuniyet ortalamanız nedir? (ÖRN : 4,30 (Sizin Lise Mezuniyet Ortalamanız) / 5,00 (Kullanılan skala) formatında yazın lütfen.

---------- / ---------

20. ÖSYM sınavına ilk girişinizde mi İTÜ'ye giriş hakkı kazandınız?

Evet ☐ Hayır ☐

21. Devam ettiğiniz bölüm ÖSYM tercih formunda kaçıncı sıradaydı?

AKEDEMİK

22. Okul kütüphanesinden kitap alıyor musunuz / faydalanıyor musunuz?

Evet ☐ Hayır ☐

23. Akademik olmayan Kitap / Dergi / Gazete okuma sıklığınız nedir?

Hiç ☐ Nadiren ☐ Bazen ☐ Çok sık ☐

24. Akademik (Mesleki) Kitap / Dergi okuyor musunuz? (Ders kitapları dışında)

Evet ☐ Hayır ☐

25. Eğitim dalınızla ilgili, herhangi bir konferans veya Sempozyuma katıldınız mı?

Evet ☐ Hayır ☐

26. Lisansüstü eğitime devam etmeyi düşünüyor musunuz?

Evet ☐ Hayır ☐ Karar vermedim ☐

27. Bir önceki yıl sonu Cumulative GPA iniz nedir?

---,---

28. Mezuniyet sırasındaki tahmini Cumulative GPA'iniz ne olabilir?

---,---

29. Okulda hazırladığınız proje / ödevleriniz eğitiminize katkıda bulunarak, sizi akademik ve iş hayatına hazırlıyor mu?

Evet ☐ Hayır ☐ Fikrim Yok ☐

30. Derslere genel olarak katılım sıklığınız için aşağıdakilerden hangi geçerli?

%70 ☐ %80 ☐ %90 ☐ %100 ☐

KİŞİSEL BİLGİLER

31. Part-time çalışıyor musunuz?

Evet ☐ Hayır ☐

32. Sinemaya ne sıklıkta gidiyorsunuz?

Hiç ☐ Ayda bir kez ☐ Haftada bir kez ☐ Hafta da bir kezden daha sık ☐

33. Spor yapıyor musunuz?

Evet ☐ Hayır ☐

34. Herhangi bir hobi / kişisel gelişim kursuna gidiyor musunuz?

Evet ☐ Hayır ☐

35. Ortalama günde kaç saat Internet kullanıyorsunuz?

İki saatten az ☐ İki – Dört saat arası ☐ Dört saatten fazla ☐

36. Ortalama günde kaç saat televizyon seyrediyorsunuz?

İki saatten az ☐ İki – Dört saat arası ☐ Dört saatten fazla ☐

37. Ortalama günde kaç saat ders çalışıyorsunuz?

İki saatten az ☐ İki – Dört saat arası ☐ Dört saatten fazla ☐

38. Okul dışında da okuldan arkadaşlarınızla görüşüyor musunuz?

Evet ☐ Hayır ☐

39. Okul kulüplerinde görev alıyor musunuz?

Evet ☐ Hayır ☐

40. İstanbul'da ikamet etme şekliniz aşağıdakilerden hangisidir?

Kendi aileniz ile	
Aile yanında	
Yurtta	
Evde (tek başına)	
Evde (ev arkadaşı ile)	

41. Okula ulaşımınız için aşağıdakilerden hangisi geçerlidir?

Özel araç ☐ Toplu Taşıma ☐ Yürüyerek ☐

42. Genel olarak, okula ulaşım süreniz için aşağıdakilerden hangisi geçerli ?

15 dakikadan az	
15 dakika – 30 dakika arası	
30 dakika – 1 saat arası	
Bir saatten fazla	

43. Kaldığınız yerdeki ders çalışma ortamınızdan memnun musunuz?

Memnun değilim ☐ Memnunum ☐ Çok memnunum ☐

44. Okul dışında da kullanabildiğiniz özel bilgisayar erişiminiz var mı?

Evet ☐ Hayır ☐

45. Ortalama aylık harcadığınız tutar?

< 200 YTL ☐ 201- 400 YTL arası ☐ 401 – 600 arası ☐ >600 YTL ☐

46. Burs alıyor musunuz?

Evet ☐ Hayır ☐

47. Harç kredisi alıyor musunuz?

Evet ☐ Hayır ☐

48. Mezun olunca istediğiniz zaman bir işe girebileceğinize inanıyor musunuz?

Evet ☐ Hayır ☐

49. Mezun olunca ne kadar süre içinde işe gireceğinizi tahmin ediyorsunuz?

Hemen ☐ İlk 6 ay içinde ☐ Bir yıl içinde ☐ Fikrim Yok ☐

50. İTÜ'de sosyal ortamı nasıl buluyorsunuz?

İyi ☐ Orta ☐ Kötü ☐

51. İTÜ'de en çok sevdiğiniz ders grubu aşağıdakilerden hangisidir?

Mesleki tasarım	
Temel mühendislik	
İnsan ve toplum bilimleri	
Temel bilimler	

52. Genel olarak İTÜ'de okumaktan memnuniyet dereceniz nedir?

Memnun değilim ☐ Memnunum ☐ Çok memnunum ☐

CURRICULUM VITAE

İnci Bölükbaşı was born on 14th of May, 1978 in İstanbul. She was grown up in Lüleburgaz, Kırklareli until university education. She has graduated from Lüleburgaz Anatolian High School in 1996, with CGPA of 5.00 out of 5.00.
She has attended Industrial Engineering Department of Marmara University in 1996. She has got undergraduate degree from Marmara Universtiy in 2000, with CGPA of 3.55 out of 4.00.
She has attended Engineering Management program of Istanbul Technical University in 2002. She has graduated from Istanbul Technical University at 2005.