SEPTEMBER 2018

M.Sc. in Industrial Engineering

BEKİR POLAT

UNIVERSITY OF GAZIANTEP
GRADUATE SCHOOL OF
NATURAL & APPLIED SCIENCES

ESTIMATION OF SUCCESS OF ENTREPRENEURSHIP PROJECTS WITH DATA MINING

M.Sc. THESIS
IN
INDUSTRIAL ENGINEERING

BY
BEKİR POLAT
SEPTEMBER 2018

Estimation of Success of Entrepreneurship Projects with Data Mining

M.Sc. Thesis

in

Industrial Engineering

University of Gaziantep

Supervisor

Asst. Prof. Dr. Alptekin DURMUŞOĞLU

by

Bekir POLAT

September 2018

REPUBLIC OF TURKEY
UNIVERSITY OF GAZİANTEP
GRADUATE SCHOOL OF NATURAL & APPLIED SCIENCES
DEPARTMENT OF INDUSTRIAL ENGINEERING

Name of the thesis: Estimation of Success of Entrepreneurship Projects with Data Mining

Name of the student: Bekir POLAT

Exam date: September 14, 2018

Approval of the Graduate School of Natural and Applied Sciences

Prof. Dr. Ahmet Necmeddin YAZICI
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

Prof. Dr. Serap ULUSAM SEÇKİNER
Head of Department

This is to certify that we have read this thesis and that in our consensus opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. Alptekin DURMUŞOĞLU
Supervisor

Examining Committee Members:	Signature
Asst. Prof. Dr. Alptekin DURMUŞOĞLU	……
Asst. Prof. Dr. Zeynep Didem UNUTMAZ DURMUŞOĞLU	……
Asst. Prof. Dr. Yunus EROĞLU	……

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Bekir POLAT

ABSTRACT

ESTIMATION OF SUCCESS OF ENTREPRENEURSHIP PROJECTS WITH DATA MINING

POLAT, Bekir
M.Sc. in Industrial Engineering
Supervisor: Asst. Prof. Dr. Alptekin DURMUSOGLU
September 2018
59 pages

Small and medium-sized enterprises (SMEs) have an important place in the economy due to the fact that 99.8% of businesses in Turkey are SMEs. It is important to survive for SMEs, especially newly founded enterprises. In order to help SMEs survive, KOSGEB provides the entrepreneurs with 3 year-support. However, the supported entrepreneurship projects still fail and cause to the waste of allocated resources for these projects. This study aimed to prevent waste of resource and to estimate the success and failure of proposed entrepreneurship projects with data mining algorithms. Thereby, the accuracy of the estimates increased and decisions about the projects were based on a scientific approach. As data of the study, the projects evaluated by KOSGEB Gaziantep Directorate between 2012-2014 were analyzed by taking some features such as age, gender, experience, education, partnership structure, market, location, sector, personnel, and capital into consideration. As a result of the analysis of the data, it has been examined whether entrepreneurial projects were successful or not. The data obtained from the entrepreneurship projects were pre-processed and adapted to WEKA 3.9.2 software. The dataset was classified using 10-fold cross-validation with C4.5, Naive Bayes, Logistic Regression, Random Forest and Support Vector algorithms. The results of the classification were compared and the C4.5 algorithm was found as the most successful algorithm with 70.75% prediction accuracy. In consequence of the C4.5 algorithm, the features affecting the tree were found as capital, partner, location, and age, respectively. The features that did not affect the tree were gender, education, market, sector, and personnel.

Key words: Entrepreneurship, SME, Data Mining, Classification

ÖZET

GİRİŞİMCİLİK PROJELERİNİN BAŞARISININ VERİ MADENCİLİĞİ ALGORİTMALARI İLE TAHMİNİ

Bekir POLAT
Yüksek Lisans Tezi, Endüstri Mühendisliği
Tez Yöneticisi: Dr. Öğr. Üyesi Alptekin DURMUŞOĞLU
Eylül 2018
59 sayfa

Türkiye'deki işletmelerin % 99,8'i küçük ve orta ölçekli işletmeler (KOBİ) olduğu için ekonomide önemli bir yere sahiptirler. KOBİ'lerin, özellikle de yeni kurulan KOBİ'lerin hayatta kalması önemlidir. Girişimcilerin başarısı için KOSGEB 3 yıl süreli destek vermektedir. Bununla birlikte, desteklenen girişimcilik projeleri hala başarısız olmakta ve bu projeler için ayrılan kaynağın israfına neden olmaktadır. Bu çalışma, veri madenciliği algoritmaları ile girişimcilik projelerinin başarı ve başarısızlık durumlarını tahmin etmeyi amaçlamaktadır. Böylece, tahmin doğruluğu artacak ve projeler hakkındaki kararlar bilimsel yaklaşıma dayandırılacaktır. Çalışmada kullanımak üzere 2012-2014 yılları arasında KOSGEB Gaziantep Müdürlüğü tarafından değerlendirilen projeler; yaş, cinsiyet, deneyim, eğitim, ortaklık yapısı, pazar, yer, sektör, personel ve sermaye özelliklerine göre analiz edilmiştir. Bu özelliklerin bir sonucu olarak, girişimci projelerinin başarılı olup olmadıklarına bakılmıştır. Girişimcilik projelerinden elde edilen veriler, ön işleme tabi tutularak WEKA 3.9.2 yazılımına uyarlanmıştır. Veriler C4.5, Naive Bayes, Logistic Regression, Random Forest ve Support Vector algoritmaları 10-katlamalı çarpraz doğrulama yöntemi kullanılarak sınıflandırılmıştır. Sınıflandırma sonuçları kıyaslanmıştır ve en başarılı tahmini yapan algoritma %70,75 doğruluk ile C4.5 algoritması olmuştur. C4.5 algoritması sonucunda ağacı etkileyen özellikler sırasıyla sermaye, ortak, konum ve yaş olarak bulunmuştur. Ağacı etkilemeyen özellikler ise cinsiyet, eğitim, pazar, sektör ve personeldir.

Anahtar Kelimeler: Girişimcilik, KOBİ, Veri Madenciliği, Sınıflandırma

ACKNOWLEDGEMENTS

I would like to express my deepest respect and most sincere gratitude to my supervisor, Asst. Prof. Dr. Alptekin DURMUŞOĞLU, for his guidance and encouragement at all stages of my work. His constructive criticism and comments from the initial conception to the end of this work is highly appreciated.

I would like to thank my colleagues KOSGEB Experts Mustafa ŞİMŞEK, Muhammed PAKSOY and Selim ÇÖREKCİOĞLU for their help in collecting the data.

Finally, I would like to serve my gratitude to examining committee members spending their valuable time for attending my M.Sc. qualification.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

KOSGEB	Küçük ve Orta Ölçekli İşletmeleri Geliştirme ve Destekleme İdaresi
EBIT	Earnings before interest and taxes
CART	Classification and Regression Trees
MDA	Multiple Discriminant Analysis
KNN	K-Nearest Neighbors algorithm
SVM	Support Vector Machine
SMEs	Small to Medium Enterprises
CHAID	Chi-squared Automatic Interaction Detector
KDD	Knowledge Discovery in Database and Data Mining
OLAP	Online Analytical Processing
RDBMS	Relational Database Management System
SQL	Structured Query Language
CRISP-DM	Cross-Industry Standard Process for Data Mining
ID3	Iterative Dichotomiser
ROC	Receiver Operating Characteristic
CRM	Customer Relationship Management
WEKA	Waikato Environment for Knowledge Analysis
R & D	Research and Development
ISGEM	İş Geliştirme Merkezi (Business Development Center)

CHAPTER 1

INTRODUCTION

Nowadays, with the development of technology and information systems in government institutions, marketing, banking, telecommunication, health services, many data can be stored and easily accessed. Developing computer technology and databases increased the amount of data that can be stored and processed, but made it difficult to control. As the amount of data increases, it becomes impossible for people to understand this data and obtain useful information when no tools are used. As a solution to this situation, the concept of "Data Mining" was developed.

Data mining can be defined as the process of analyzing data from many different perspectives and obtaining meaningful information from that data. By means of data mining, valuable information can be reached by discovering meaningful and new relationships among many data. Data mining techniques, which have been developed to detect meaningful and useful patterns, have become a necessary tool because analyzing information collected and stored in computer environment is not possible with classical inquiry methods or simple statistics.

This study was prepared to solve the problem of resource waste caused by unsuccessful attempts of entrepreneurs who supported by KOSGEB. The effects of failure were explored in order to prevent the waste of limited state resources and to prevent unsuccessful projects being supported. Data mining classification algorithms were used to make the state support more effective by examining the features that affect the success or failure of entrepreneurs in the previous studies carried out in the field. The data belonging to projects benefiting from KOSGEB entrepreneurship support were examined, features that affect the failure were identified and data mining classification algorithms were used to predict success or failure. The first chapter of the study included what the problem was and what methods were used to

solve the problem. Atthe end of the first chapter, the literature review was given. In the second chapter, the definition and related concepts of data mining, data mining application areas, data mining application processes and methods and algorithms used in the work were explained. In the third chapter, KOSGEB Entrepreneurship Support and its stages were explained. In the fourth chapter, the data was explained in detail and the used algorithms were applied to the dataset. In the conclusion section, the information obtained from the algorithms were interpreted, the features affecting the success were revealed, and the algorithm that predicts with the best accuracy was selected by comparing the success and the accuracy of the used algorithms.

1.1 Literature Survey

In this part of the thesis, the studies about the previous company failures have been discussed. When we look at studies in the literature, financial ratios were used in the first studies in the analysis of successful or unsuccessful firms. In first studies, estimations were made with classical statistical methods. The sufficient information obtained from large-scale corporations was made it easier to make predictions with classical methods. But, using these methods to newly established and the small-medium enterprises successful result was not achieved because of lack of financial information. For this reason, non-financial methods were begun to be used in the analysis of newly established and small-medium enterprises. Lussier method is the pioneer of non-financial methods, discriminant analysis, and logistic regression were used as models in this method. Along with the widespread use of the algorithms, data mining algorithms were begun to use classification data obtained from both financial methods and non-financial methods.

Beaver’s study [1] was based on financial ratios and used Moody's Industrial Manual, including data from companies in many different industries. Selected data was compared with 30 different financial ratios. Companies failing to pay their debts were determined as unsuccessful. 79 of the selected companies failed while 1200 did not. The best accuracy rate was the capital / debt ratio and was estimated to be 90% successful. The data was analyzed by Likelihood Ratios and the Dichotomous Classification Test.

Altman [2] analyzed the financial ratios of 66 production businesses by using multiple discriminant analysis method and developed the Z Score model which was formed by five ratios which can be used in predicting financial failure. The model was shown in equation (1.1).

$$Z \text{ Score model} = 0.012X1 + 0.014X2 + 0.033X3 + 0.006X4 + 0.999X5. \quad (1.1)$$

Where:

X1: Working capital / Total assets,

X2: Retained earnings / Total Assets,

X3: EBIT / Total assets,

X4: Shareholders' Equity / Total debts book value,

X5: Sales / Total assets.

Altman [2] classified businesses with a Z Score greater than 2.99 as "Non-Bankrupt". There was no financial risk for businesses in this region. Enterprises with a Z score of 1.81 to 2.99 were classified in the "gray area". Though the risk of financial failure for businesses in this area was not very high, investment needed to be cautious. While those with a Z Score of less than 1.81 were identified as having high risk of financial failure. This work [2] by Altman showed a successful classification performance of 95% a year ago and 72% two years ago in predicting financial failure.

In another study, by Izan [3], 103 firms in different sectors in Australia were classified as fail or non- fail with the help of discriminant model with 5 variables. The five variables in the classification were determined as Earnings before interest and taxes / Tangible total assets, Earning before interest and taxes / Interest payment, Current assets / Current liabilities, Borrowings / Shareholder Funds, Market value of equity / Total liabilities. In the analysis, 53 companies failed, 50 companies were successful. The classification results in 91.9% accuracy.

In one another study by Duchesneau and Gartner [4] in the 90's, when non-financial studies began, data were generated by quantitative and qualitative data of 26 small

businesses engaged in the distribution of fresh fruit juice in the United Stated. Univariate analysis of variation and correlations were applied to the obtained data and the results were interpreted. According to the obtained results, the characteristics of the successful firms were as follows: have entrepreneurial parents, have a broad range of previous managerial experience and seek to reduce risk.

Cooper and colleagues [5] followed 2994 companies in the United States for three years to be able to indicate success or failure. The entrepreneurial capability of enterprises, relevant knowledge bases and expertise and financial capital data sets were formed. Data in the data set was obtained by surveys, short-response postcards and post-office return mail methods. After the enterprises with missing data were removed from the database, there were 1053 firms left for the study. The businesses were grouped in three different ways: failed, survived with low growth, and grew substantially. The logistic regression model was applied to the database and the x^2 test was applied for the accuracy of the results. According to the results obtained, the p-values of businesses that survived were 0.98 and businesses that failed, grew substantially were 0.99.

In his studies, Lussier [6-9], the empirical survey was conducted to examine the entrepreneurs' experiences. Lussier created the data set of the model known as the S / F model. The features included in the model were age, partners, advisors, planning, education, minority business ownership, staffing, parents owned business, record keeping and financial control, capital and product service timing. Using the discriminant analysis and logistic regression in his model, he investigated whether the entrepreneurs were successful or not. The model had a different prediction level of accuracy between 70%- 85%.

In another study of Lussier [10], dataset was prepared by taking information from the human resources of Republic of Croatia companies. Successes or failures of the firms were estimated by logistic regression method. Instead of 15 different parameters in previous studies, in the model known as Lussier model, only planning, professional advisors, education and staffing were taken into consideration. Reduced data, normal Lussier model and the accuracy of classification was compared. As a result of logistic regression, the data was correctly classified as full model 72.32% and reduced model 71.79%.

In a study [11], failure estimates were made by using data mining methods on the database prepared using 1133 firms' 20-year data in the UK. Firstly, discriminant analysis, logistic regression, neural networks and C5.0 classifications were used. The models were then hybridized with ANOVA statistical model. Obtained accuracy rates were compared and it was observed that hybrid modeling increased the success rate. It was the best results in normal data mining methods decision trees and neural networks.

Mehralizadeh and Sajady [12] investigated internal and external factors that affected success or failure of entrepreneurs with a survey study. In the study, 51 entrepreneurs in Ahvaz city were chosen randomly. The results of the questionnaires were interpreted by establishing test with SPSS program. The missing aspects of unsuccessful entrepreneurs were found to be weaker technical skills, financial issues, planning and organizing of their business, economic issues, informal issues, weak managing conceptual skills, personnel skills, education and low training, and weak human relations. The features that successful entrepreneurs possess were suitable managing technical skills, selecting appropriate personnel with relevant skills, education and paying more attention to personnel training, application of management conceptual skills, financial issues, better human relation, recognize the economic situation, planning and organizing of their business and informal issues.

Nguyen [13] was made to estimate the failures of businesses in the Corporate Scorecard Group database in Australia between 1988-2002 using multi-layer neural networks, probabilistic neural networks, and logistic regression models. The data set was analyzed using ANGOSS Studio software and it was observed that the neural network models were more successful than the classical statistical methods.

In a study [14], historical data were combined with hybrid model, C4.5 and genetic algorithm when estimating credit score. German and Australian credit data sets were used to evaluate different credit scoring models. The classification accuracy of the hybrid model was higher than that of C4.5.

The decision tree algorithm was tried to predict the success or failure of the enterprises in place of the discriminant and logistic regression analysis which are frequently used in the business failure prediction model. As a result of the study, the

CART and RPA decision tree methods created very parallel outcomes and were the best complete forecasters [15].

In a study [16], instead of classical statistical methods such as multiple discriminant analysis (MDA) and logistics regression, kNN, SVM, CART and C4.5 algorithms were used in predicting firm failure. The success of these algorithms, which are the most preferred data mining algorithms, was compared. The CART and C4.5 algorithms were used for the first time in this work in predicting firm failure. The financial ratios of 153 Chinese companies listed on the Shenzhen Stock Exchange and Shanghai Stock Exchange were used as data base. The results were compared using two-tailed paired-sample t test and the accuracy rank was found as CART> SVM> kNN> MDAFS-CART> MDA> Logit.

Yap and colleagues [17] classified home loans, small business loans and insurance applications as bad risk or good risk. In this study, the logistic regression and decision tree algorithm were used to compute the credit score model and compare the results. The classification error rates for credit logistic regression and decision tree were 28.8% and 28.1%, respectively. classified home loans, small business loans and insurance applications as bad risk or good risk. In this study, the logistic regression and decision tree algorithm were used to compute the credit score model and compare the results. The classification error rates for credit logistic regression and decision tree were 28.8% and 28.1%, respectively.

Lussier S / F model was applied by collecting data from 234 small enterprises in Chile. Logistic regression model was applied to the Lussier model and the results obtained from this study were used to compare the accuracy of this model with those of the United State and Croatia. The correct classification rate in Chile was 63.2%, similar to that of United State (69%). The accuracy rate in Croatian was 72%, which was attributed to the fact that there were more successful firms [18].

Marom and Lussier [19] examined 205 SMEs in Israel. Of these, 101 were unsuccessful and 104 were successful. The characteristics of Israeli SMEs were researched and analyzed in terms of Capital, Record Keeping and Financial Control, Industry Experience, Management Experience, Planning, Professional Advisors, Education, Staffing, Product / Service Timing, Economic Timing, Age, Partners,

Parents, Minority, and all these were tested by regression analysis. In the result of the model, 5 parameters affecting the result appeared. Entrepreneur who had adequate capital, good record keeping and financial controls, developed plans and used professional advisors had a meaningfully better chance of achievement in Israel. Age was similarly substantial for starting a business.

In a study [20] was conducted to investigate whether C5.0 and CHAID decision tree algorithms could be used to predict the financial failure and / or success of the manufacturing company. According to the results, the classification of the C5.0 model was 90.97% for the training set and 87.5% for the test set. The classification of the CHAID model was 83.03% for the training set and 82.5% for the test set. Thus, classifications made by C5.0 and CHAID algorithms can be considered effective.

Hyder and Lussier [21] investigated the features that influenced the success or failure of entrepreneurs in Pakistan. 143 entrepreneur survey studies were conducted and data of 15 features in the Lussier model were collected. The obtained data was interpreted using logistics regression statistical analysis. As a result of the study, it was found that businesses with adequate capital, developing business plans, proper staffing and partners had a higher chance of success.

CHAPTER 2

DATA MINING

2.1 Definition of Data Mining

Data mining has many different definitions in the literature. Data mining is a collection of improved methods for bringing both understandable and usable forms of data together for the data owner, by discovering data that has unexpected / unknown relationships between them with potentially useful results. In general, data mining is the process of transforming data into meaningful information by analyzing the data available.

The increased use of the database has necessitated the emergence of new technologies to digest large volumes of generated data [22]. It is also important to ensure that data becomes meaningful information as long as recorded and managed. Traditional methods and human analyzes are insufficient to be able to make large data meaningful. Data mining and Knowledge Discovery in Database (KDD) are the processes that produce methodologies and tools to make large data meaningful [23].

Wisely analyzed data is a prized source. Data mining is about answering complications by analyzing the data in the database. As the world grows in mass, data mining is our only hope to unveil the underlying patterns [24].

Taking out of unseen guessing data from huge databases is called data mining. Data mining is a powerful technology that will advantage users to emphasize the most essential data in their data warehouses. Data mining tools anticipate upcoming tendencies and behaviors, helping users make practical, knowledge-driven results [25, 26].

The procedure of determining amazing patterns and information from huge volumes of data is called data mining. Many people outlook data mining as an important stage

in the information discovery process, while others treat data mining in another popular term, verbally discovery of information, or synonymously with KDD [27]. The progression of realizing understanding, remarkable, and original patterns, as well as imaginative, reasonable, and analytical models from huge data is data mining [28].

Data mining is the study of gathering, cleaning, treating, analyzing, and acquisition valuable insight from data. An extensive disparity occurs in terms of problems areas, claims, formulations, and data presentations that are faced in actual applications. Consequently, data mining is a wide umbrella term that is used to define these dissimilar features of data processing [29].

The advance of Information Technology has created huge volume of databases and vast data in countless parts. The investigation in databases and information technology has specified increase to a method to collection and operate this valuable data for additional decision making. Data mining is a method of taking out of valuable information and forms from vast data [29].

2.2 History of Data Mining

Data mining is a multi-disciplinary field that helps as a bridge concerning countless technical fields. There is relationship between data mining and other disciplines such as Database Technology, Statistics, Artificial Intelligence, Machine Learning, Pattern Recognition and Data Visualization. Looking at the history of data mining, it is necessary to look at the background process with other disciplines as data mining is connected to many disciplines.

Data mining is rooted in classical statistics, artificial intelligence and machine learning. The statisticians use estimation algorithms to make valid estimates by exploring the patterns and relationships within the data with the aid of analysis tools from large-scale databases. Statisticians have been trying to find relationships between patterns of data for a long time, manually. Data mining has evolved as a process by which statisticians can automatically do what they do. Machine learning, the backbone of artificial intelligence work, is the basis of data mining. Experts working in the field of machine learning have come up with predictive algorithms. It is seen that these algorithms try to produce solutions by making inferences.

Figure 2.1 Relation of Data Mining to Other Disciplines

The concept of data mining has been intensively talked about in the 1990s, but it actually dates back to the 1700s. Basic statistical methods are used to analyze data. For this reason, it is necessary to investigate the bases of data mining in classical statistical methods.

Statistics has become a method of providing services for the evaluation and analysis of data from the past to the present day. The process starting with classical statistical methods forms the basis of data mining. Data mining and statistics are included in a rigorous study of the process of extracting, analyzing and using data. In 1763 Thomas Bayes published a journal called Bayes' theorem [30]. This work is essential to data mining and probability, as it lets considerate of compound actualities based on predicted possibilities. In 1805 Adrien-Marie Legendre and Carl Friedrich Gauss applied Regression, which aims to predict the relationship between variables [31]. Regression has become one of the most important tools in data mining. With the computers' existence, it became possible to conduct possible statistical surveys that were not possible before, and after the 1990s, statistics and data mining is carried a common platform.

Alan Turing [32] published an article [32], On Computable Numbers, 1936, and presented a universal machine idea that allowed computers to process data in large quantities like today's computers. So the data was analyzed faster and interpreted faster. As computers began to develop, concepts of artificial intelligence and machine learning began to be discussed. In 1943, Warren McCulloch and Walter Pitts were the first people to generate a theoretical model of the neural network [33]. They have described a network idea in a paper titled Logical Account of Neural Activity, and found that the data could be analyzed by computer and artificial intelligence algorithms.

The concept of data mining was called data scanning and data fishering in the 1960s. It was assumed that the required information could be obtained when the necessary inquiry was made with the help of the computer. Similar to the tree data structure, a hierarchical data model based on the existence of a parent record for each record and the presence of many child records was also introduced in the 1960s.

The relational database management system, which is a system that kept the data in rows and columns on tables, and provided data consistency, emerged towards the end of the 1970s. The rows in the table related to the rows in other tables and provided a link between the data. In 1989, the term "Knowledge Discovery in Databases" (KDD) was coined by Gregory Piatetsky-Shapiro. At the same time, he established the first workshop called KDD.

In the 1990s, it was started to think about how to extract useful information from dataset that increased exponentially. So, Online Analytical Processing (OLAP), a tools used to consolidate huge business databases and maintenance the decision support system, emerged in the 1990s. OLAP databases were divided into one or more clips and each cube was arranged and designed to fit the format of the data retrieval and analysis required by a cube manager so that you could generate and maintain PivotTable reports and PivotChart reports that you need.

In 2000s, data mining was constantly evolving and started to be applied to almost all areas. Data mining started to be used by computer engineers. It was accepted that this new concept was also evaluated by algorithmic computer modules rather than

traditional statistical methods. The historical development of data mining is shown in Table 2.1 [27].

Table 2.1 Steps in the Evolution of Data Mining [51]

Evolutionary Step	**Business Question**	**Enabling Technologies**	**Product Providers**	**Characteristics**
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Warehousing &Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry)	Prospective, proactive information delivery

2.3 Data Mining Processes

Data mining is a process between the data stacks, decomposing the samples in the information discovery process and making the next step ready are also part of this process. In order to implement data mining methods, data held in data warehouses or databases must pass through certain stages.

There is a Cross-Industry Standard Process for Data Mining (CRISP-DM) commonly used by business supporters [34]. This model involves of six stages in- tended as a cyclical procedure [34-37]. CRISP-DM model was developed in 1996 by analysts representing DaimlerChrysler AG, SPSS, NCR and OHRA. CRISP provides an unpatented and freely available standard process for modifying data mining processes and standards into the general problem solving strategy of a business or research unit. The life cycle of the data mining project contains of 6 stages. Stages do not follow a strict order. There is always movement back and forth between phases. Mobility depends on which phase or phase of work is to be done at the end of each phase. Arrows represent the most frequently occurring dependencies between phases.

2.3.1 Business Understanding

This phase, which can also be called as the stage of understanding the research problem, constitutes the first step of the data mining process. Precisely expressing the objectives and needs of the project as a whole includes the process of setting goals and constraints in terms of data mining formulations and establishing a preliminary strategy in this direction. The aim should be focused on business problem and must be expressed clearly. How to measure the success rates of the results obtained at the end of the study should be defined at this step. In this phase, it is essential to know how to assess and use the outcomes.

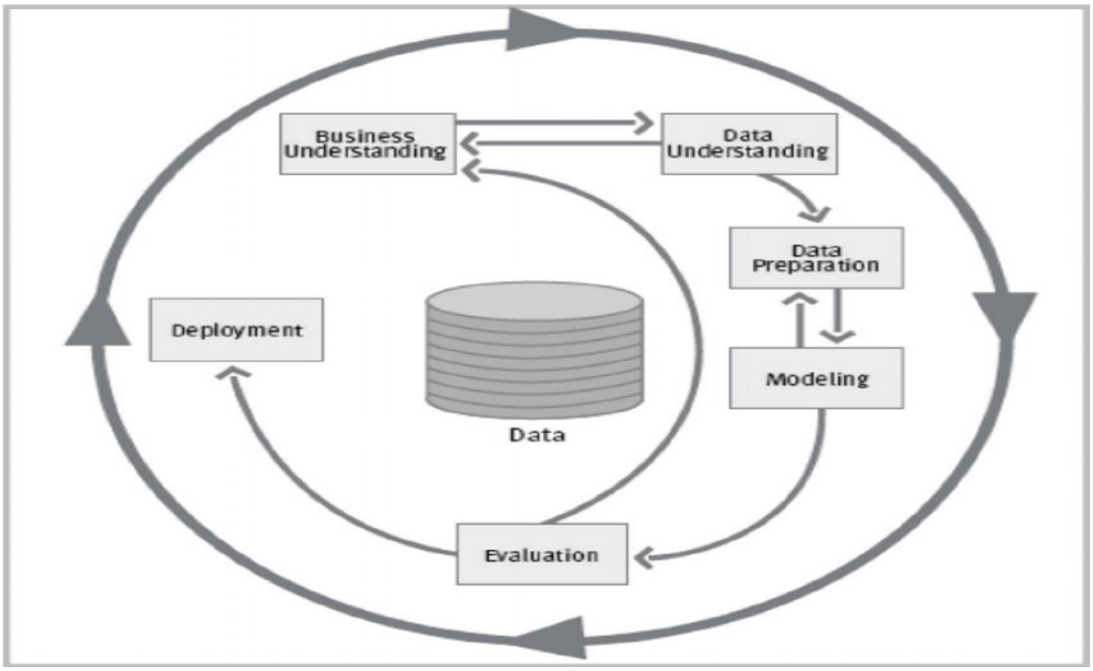

Figure 2.2 Phases of the CRISP-DM reference model [27]

2.3.2 Data Understanding

The data understanding phase begins with the data collection step. The data can be acquired from interior and exterior sources. Internal resources are the database of the business, such as customer records, purchases. External sources are data obtained from outside the enterprise such as population census, research companies, and databases. Then the process continues with the identification and the evaluation of the data quality data. If the analyst does not have enough knowledge about the success of the project, somebody who has the knowledge must be asked for help. The aim is to acquire familiarity with the data to be used in the study. The data identification process and the research problem identification are involved as sub-processes. As you understand the problem, it becomes possible to look at different data and gain different perspectives about the problem.

2.3.3 Data Preperation

This phase is a very intensive phase of labor that involves all the arrangements to be processed from the first raw to final. Determination of appropriate conditions and variables for analysis includes processes such as transformation and cleaning of data. The pre-processing phase of data is the most time-consuming phase in the data mining process, usually due to the size of the data sets. Compared with the other phases of CRISP-DM, it is stated that the most effort and time (80%) are spent in the preparation phase of the data.

- **Collection of data:** The data, records and variables that start to be collected during the understanding of the data are selected. It is also necessary to pay attention to the technical limitations that exist at the same time with the quality of the data and the targets. The number of used records is another important point. The number of enrollments should be assessed to ensure that the work is not missing or vice versa. Records from different data sources are integrated at this stage. Obtaining data from different sources leads to differences in data formats. It is necessary to evaluate the harmony of the data and eliminate the existing incompatibilities by treating it carefully in order to prevent the results from distorting the situation in the future.
- **Cleaning of Data:** It is the phase of improving the data quality by removing erroneous and inconsistent data. It is preferable to remove the data resulting from

incorrect entries or exceptional cases in order to prevent the resultant distortion. In some cases, it may be preferable to proceed with a sample that can represent the entire population in the name of preventing data contamination. In some cases, different models can be applied in order to complete the missing data, instead of directly extracting the missing data.

- **Conversion of Data:** It is the process of converting the given data at specific intervals. Since a variable has a very high value, low value variants are used to alleviate potential deviations that can occur with dominance.
- **Reduction of Data:** When applying data mining, large data sets are usually preferred, but this also causes problems. Especially in some applications, the number of variables and states is considerably large, and the applicant's state and variable numbers have to be reduced to manageable numbers. With resolutions only for certain dimensions, it is possible to narrow the data by removing certain dimensions from the dataset, or by creating smaller datasets that represent it instead of larger datasets.

2.3.4 Modelling

The modeling phase is the stage in which the appropriate modelling techniques are selected and applied in the name of the application of several different methods for the problem if desired. Some methods may not be suitable for the type of data or problem, so if necessary, they can be returned to the previous stage of data preparation. Therefore, the steps of modelling and data preparation are repeated until the most appropriate model is achieved. It is difficult to decide the most appropriate technique before starting the model establishment studies. It is useful to make studies to find the most appropriate model by comparing different models and accuracy ratings. But even if the established model is fairly accurate, it should be known that it is not possible to guarantee that it truly modelled the real world.

2.3.5 Evaluation

It is the process of assessing whether the quality and efficiency of one or more models applied in the modelling phase and the model conforms to the objectives set in the first phase. The evaluation phase involves reviewing the applied data mining processes and deciding on the use of their results. It is important that the model is in line with the identified goals and should be assessed to see if it is an uncovered topic.

At the same time, it is also assessed at this stage which different data can be used in the future and whether the place of study is sufficient.

2.3.6 Deployment

Deployment phase, the information generated at the end of the process is used to solve the problem in the direction of the determined purpose. The results are assessed, interpreted and the strategies and decisions determined in the conclusion of the results obtained are applied in real life. The decision model can be either a direct application or a subset of another application. The aim of the model is to increase the known knowledge about the given data, but the resulting data must be organized and presented, so that the data can be used. At the same time, this phase includes the writing of the research report and the re-opening of the project. The reports are important for ensuring that the results of the projects are communicated and, if necessary, replicable by others. Changes in the systems over time will require monitoring of the established models and re-arrangement if necessary. Tracking is important to prevent the model from working using the wrong data.

In [27], data mining is defined as discovery of information. Similar to the CRISP-DM process, the information discovery process is shown in Figure 2.3. These processes are briefly described below:

- Data cleaning (to eliminate noise and inconsistent data)
- Data integration (to association numerous data sources)
- Data selection (to regain data related to the examination assignment from the database)
- Data transformation (to convert and merge data into forms suitable for mining by execution summary or combination processes)
- Data mining (an important procedure where intelligent methods are applied to extract data patterns)
- Pattern evaluation (to classify the accurately interesting patterns representing information based on conditions of interest)
- Knowledge presentation (where imagining and information illustration procedures are used to current mined knowledge to users)

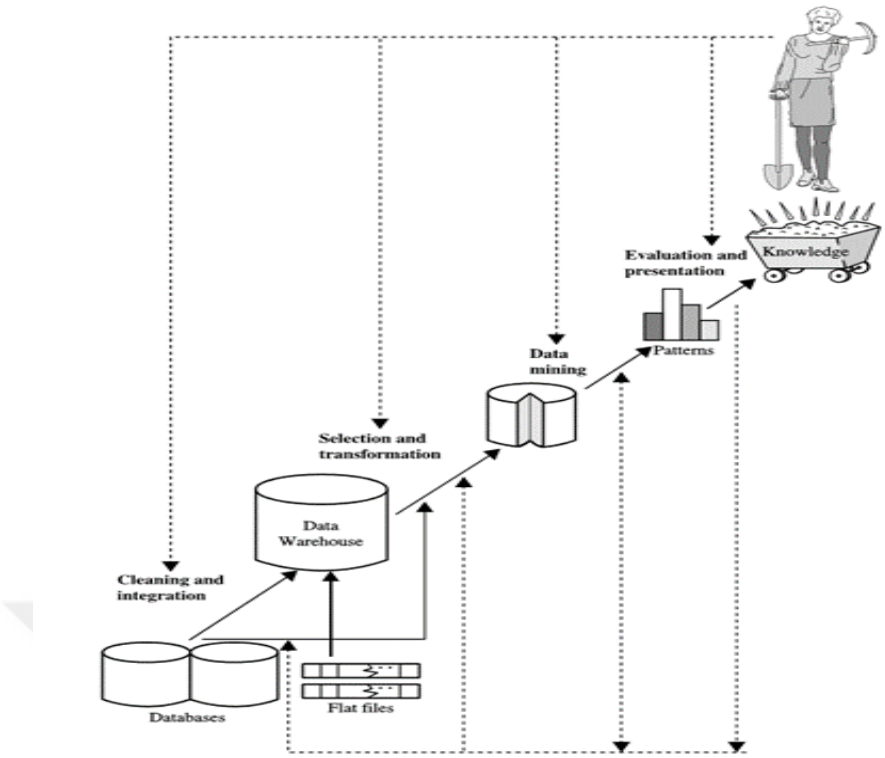

Figure 2.3 Data mining as a step in the process of knowledge discover [27]

2.4 Data Mining Methods

Data mining models can be divided into two groups as descriptive and predictive [23, 27, 38]. In predictive models, the goal is to reveal patterns within existing data, or to anticipate a future state of existing data. In predictive models, target groups are known from the beginning. The available data is analyzed and divided into known classes and a future situation is foreseen. Classification, Regression and Deviation Detection are predictive models. In the descriptive models, the patterns are defined from the given data and the results are displayed in order to take action. Clustering, Association Rules and Sequential Pattern Discovery are among the descriptive models.

2.4.1 Classification

Classification is called grouping of existing data into groups according to their common characteristics. Classification can be used in many different areas. It is possible to find examples of classification in many different contexts such as assessment of loan applications, determination of user behavior or diagnosis of illness. Classification is applied based on a learning algorithm. In other words,

classification is the process of determining the classes of existing data whose classes are unknown.

Classification is a two-step procedure, involving of a learning step (where a classification model is created) and a classification step (where the model is used to forecast class labels for given data) [23, 27].

The first step is to examine the qualities of the data in the database and to present a model suitable for the data sets. Some of the data in the database is selected randomly and used as training data. The remaining data is also used as test data. A classification model is obtained by applying an algorithm over the specified training data.

The second step is to test the rules by applying them to the test data. If the correctness of the model obtained in the test result is accepted, the model is applied on the other data.

2.4.1.1 C4.5 Algorithm

The most popular and well-known decision tree algorithms that are widely used in data mining areas are CART (Classification and Regression Tree), ID3 (Iterative Dichotomiser) and C4.5 algorithms. ID3, C4.5, and CART assume a grasping methodology in which decision trees are created in a top-down repeated split-and-overcome manner. Most algorithms for decision tree stimulation also follow a top-down method, which starts with a training set of tuples and their related class labels. The training set is sequential separated into minor subgroups as the tree is being built [39].

In the late 1970s J. Ros Quinlan [40], a researcher based in Australia, developed an algorithm which named ID3 method. Later, Quinlan improved his work of ID3, and he named his new work C4.5 algorithms. C4.5 algorithms has become a standard method to which other supervised learning algorithms are compared.

C4.5 algorithm generates decision trees that can be used for classification. It uses information gain and also gain ratio as its attribute splitting criteria. It can allow data with categorical or numerical values. C4.5 algorithm has a capability to handle missing values [41].

As a decision criterion, entropy calculation, information gain and finally pruning take place. The more entropy measurements, the more uncertain and undecided the results are. For this reason, areas with the least entropy measure are used at the root of the decision tree. Shannon's formula [42] for finding the entropy measure is:

The probability value of a class if it has n classes and it is assumed that these class values are repeated by T was shown in equation (2.1);

$$P_i = \frac{C_i}{|T|} \tag{2.1}$$

C_i represents the number of class values belonging to a class. The entropy value of these classes was shown in equation (2.2), H (T):

$$-\Sigma_i (P_i \, x \log_2(P_i)) \tag{2.2}$$

The information gain IG (Y, T) was shown in equation (2.3), to be obtained as a result of division of T class values using Y attribute values.

$$IG\ (Y, T) = H(T) - \sum_{i=1}^{n} \frac{|T_i|}{|T|} H(T_i) \tag{2.3}$$

Separation information was shown in equation (2.4),;

$$SI\ (Y) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} \log_2 \left(\frac{|T_i|}{|T|}\right) \tag{2.4}$$

The rate of the seperation information gives us how much information is gained by the separation of the relevant qualification. In this way, profit information is calculated for each attribute and the tree structure is allocated according to the quality with the highest information gain.

Another important step in making a decision tree is the pruning process. Pruning can be done in two ways. The first type, called pre-pruning, is done to stop the tree from dividing at the time the tree reaches a size that can not grow anymore. In the second type called final pruning, it is done by subtracting the separated points occured after the tree is completely created [43].

2.4.1.2 Naive Bayes Classifier

In [44] describe a simple approach as known Naïve Bayes Classifier. The Naïve Bayes classification aims to define the class or category of data presented to the system by a series of calculations defined according to probability principles. In the Naïve Bayes classification, the system is presented with a specific set of data. There must be a class / category of the data presented for teaching. With the probabilistic operations on the taught data, the new test data obtainable to the system is functioned allowing to the earlier obtained probability values and it is tried not to define which group of test data is given.

p (x | C_j): probability that x is an instance of class j, P (C_j): initial probability of class j and p (x): probability of any instance x.

The probability that a sample with class P (C_j | x): x is of class j (the last probability) was shown in equation (2.5).

$$P(x|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i)xP(x_2|C_i) \times \dots \times P(x_n|C_i) \quad (2.5)$$

For each x value, probability calculations in equation (2.5) are made and it is found which class the x value belongs to.

2.4.1.3 Logistic Regression

The goal of the logistic regression method is to find the simplest model that can predict the result of the dependent variable. Logistic regression is used to define data and to describe the connection between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level autonomous variables.

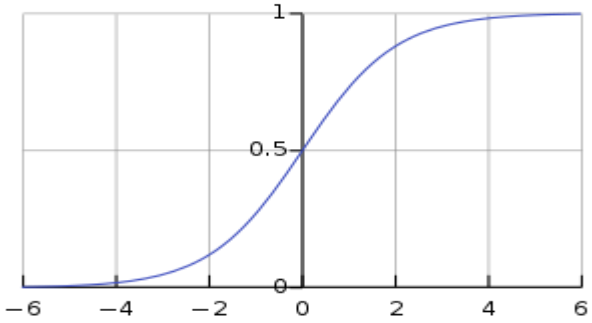

Figure 2.4 The standard logistic function [29]

Generally, the occurrence of the event is indicated by 1, and if it is not realized, it is indicated by 0. The logistic function has a range of 0 to 1, shown in Figure 2.4. The logistic regression formula is given below:

Probability that Y = 1, which is referred as p, the probability that Y is 0 is 1 – p.

$B_0 + B_1X$ is equation for the regression line. Probability that Y = 1 was shown in equation (2.6)

$$p = \frac{e^{B_0+B_1x}}{1 + e^{B_0+B_1x}} \tag{2.6}$$

2.4.1.4 Random Forest

The use of random data increases the accuracy rate while the bagging method is selected in the Random Forest algorithm. The features are randomly selected to create the tree, and the resultant tree showed the most likely voted tree assigned as the predictable class. To select a random feature, a new training data set is created, first of all, by displacement from the actual data set. Then a tree is developed from the new training set using the random feature selection. Developed trees are not pruned [45].

The Random Forest algorithm uses the Gini index to calculate the nodes of the tree. The value with the smallest Gini index specifies the division position. A branch is successful when the Gini index is less than the Gini index of a parent node. A random sample is chosen for a given T training set, and this example belongs to the C_i class. In this case, the Gini index is expressed as in equation (2.7);

$$\sum_{j \neq i} (f(Ci,T)\,|T|)(f(Ci,T)|T|) \tag{2.7}$$

in equality (2.7) (f (C_i, T) \ | T |) indicates that the C_i instance of the selected class.

2.4.1.5 Support Vector Algorithm

The working principle of the support vector machine algorithm is based on the principle of predicting the most suitable decision function that can differentiate two classes from each other, in other words, the meaning of a hyperplane which can differentiate two classes from each other optimally [46].

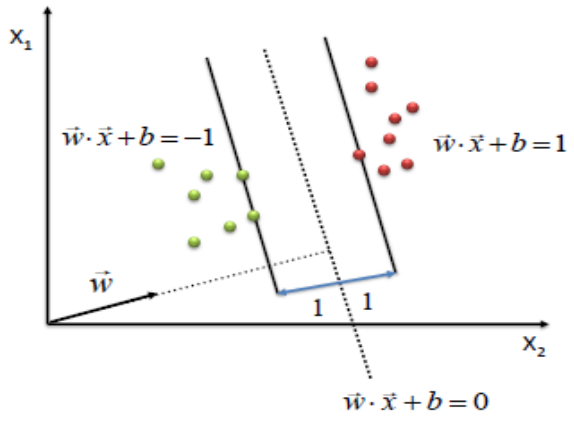

Figure 2.5 Support Vector Machine Hyperplane [46]

In order to determine the optimum hyperplane, this plane must be parallel and two hyperplanes must be defined to form the boundary (Figure 2.5). The points forming these hyperplanes are called support vectors and these planes are expressed as $wx_{i+}b=\pm1$.

As a result, the decision function for a two-class problem that can be separated is written as follows in equation (2.8);

$$f(x) = \text{sign}\left(\sum_{i=1}^{k} \lambda_i y_i (x \cdot x_i) + b\right) \tag{2.8}$$

2.4.2 Regression

In the regression analysis, it is essential to specify which events are affected when the observed event is evaluated. Any variable is a mathematical expression of the relationship between one or more different variables. The purpose of the regression is to create a model that can make the most accurate estimate of the relationship between inputs and outputs [24].

The feature that distinguishes the regression from classification is that the predicted dependent variable is a continuous numerical variable. In regression analysis, the result is called "dependent variable" and inputs are called "independent variable". When regression analysis is performed, variables included in the model are a dependent variable and one or more independent variables. Variables are countable or measurable. Univariate models are expressed as simple linear regression, multiple independent variable and multiple regression model. Regression analysis is divided into linear regression and nonlinear regression according to the equation of the created equation. A linear regression equation is expressed as in equation (2.9):

$$y = a + bx \tag{2.9}$$

The independent variable, x, represents y, and y represents classes. Thus, the y value is calculated by replacing the x value with the attribute value, i.e., the class to which x belongs is estimated. If the problem has more than one attribute, then the equation called multiple regression is shown as follows in equation (2.10):

$$y = a + b_1x_1 + b_2x_2 \tag{2.10}$$

Regression analysis is used in many areas. Regression analysis can be used to solve many different problems such as financial estimates, customer value, price evaluations, and drug reactions.

2.4.3 Deviation Detection

At the basis of the detection, the data set is divided into clusters and the data does not belong to these clusters are detected. In other words, anomalies are detected. In the anomaly analysis to be applied to the dataset, the dataset that does not have similar behavior characteristics is determined in the dataset that is set [47].

Deviation data is caused by errors during reading, recording, measuring, application, or calculation. For example, at the time program is entered, a person's age may be written as 445 instead of 44. Most of the algorithms of data mining are intended to minimize or completely remove the effect of unusual data.

Deviation detection analysis has a wide range of applications. It is used in fraud detection, for instance, detection of unusual use of credit cards and detection of fraud in telecommunication services. It is used to find unusual results in various medical treatments.

2.4.4 Clustering

Cluster analysis is the technique of separating a set of data items into subgroups. Each subgroup is a cluster, such that objects in a cluster are related to one another, however different from objects in other clusters. A set of clusters subsequent from a cluster analysis can be mentioned as a clustering. In this perspective, dissimilar clustering procedures may create unlike clustering on the similar dataset. The separating is not completed by individuals, but by the clustering algorithm. Therefore, clustering is beneficial in that it can guide to the finding of earlier unidentified groups in the data [27].

The biggest difference between classification models and clustering models is that they have not been previously defined. There is no predetermined class in the clustering model. The data contained in the database are grouped according to certain characteristics determined by the expert who sets up the model. Hence, homogeneous but heterogeneous clusters with different characteristics occur.

Clustering can sometimes be used to narrow or customize records in the database before other data mining methods are applied. The planned data mining method may be applied to only one of the clusters, or different clusters may need to be applied differently. Clustering models are applied especially when customer characteristics are specified, which is why it is widely used in the retail sector. With the clustering model, customers can be grouped according to their profitability, usage rate, or cluster of market potentials. It enables customers to communicate correctly with other customers who have similar characteristics and are expected to give common reactions.

2.4.5 Association Rules

Association rule is used to find the link between the data in the databases. Connection rules determine the connections between hosts and calculate the likelihood that the events are concurrent. Association rules establish links that are not

easy to distinguish between data. These links provide many benefits for companies, ranging from customer satisfaction to strategic decision-making [48].

The method is often used as a market basket analysis in the retail sector, as it aims to reveal relationships between data. In this respect, many rules are used in the retail sector, from placement of products on shelves, creation of advantageous product packages, preparation of catalogs, and determination of the area to be used. There are many applications of this method not only in retail sector, but it is also used for determining the product specifications in automotive sector, determining the products that should be close to storage.

2.4.6 Sequential Pattern Discovery

Sequence patterns, also called sequential time patterns, examine events occurring over time. Trends and cycles are determined by examining the existing records. Relations are established by examining the events in certain frequencies within certain time intervals. The fact that ordered sequences are based on events occurring over time distinguishes this algorithm from association rules [27].

For example, in the telecommunication sector, when the available records are examined, it seems that smartphone users do not have available internet packages or existing packages increase their capacity. With this relationship established, it may be possible to present to the persons who are found to have received smartphones a bid at a more reasonable price but to keep it connected with the commitment period.

Predicting customer behavior is also a great advantage in a highly competitive industry such as retail. With the development of barcode technology, companies have gathered quite a lot of data. Firms can analyze and correlate customer movements over time using the data which they can obtain via products such as loyalty cards, and turn them into an advantage for themselves by offering personalized campaigns. In the same way, with these records, it is possible to use sequential order to determine the behavior of the customer lost and to apply different ways to keep existing customers.

2.5 Model Performance Criterion

The different models used to estimate model performance are an error rate, precision, sensitivity, F-criterion and ROC Graphs [49]. The achievement of the model is associated with the number of samples that are properly categorized and the number of samples that are categorized mistakenly.

The performance evidence of the results completed in the test result can be said by the confusion matrix. In the confusion matrix, the rows symbolize the real numbers of the samples in the test set, and the columns symbolize the guess of the model [50].

Table 2.2 Confusion Matrix

		True Class	
		Class =1	Class =0
Hypothesized Class	Class =1	a	b
	Class =0	c	d

a: TP (True Pozitif) c: FP (False Pozitif)

b: FN (False Negatif) d: TN (True Negatif)

The most common and simple way used to measure model performance is the accuracy of the model. Accuracy was shown in equation (2.11) can be measured using one or more test data which are actually different from the training data set [51]. The correct number of classified samples (TP + TN) is the ratio of the total number of samples (TP + TN + FP + FN). The error rate was shown in equation (2.12) is 1 of this value. In other words, the number of misclassified samples (FP + FN) is the ratio of the total number of samples (TP + TN + FP + FN).

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2.11}$$

$$Eror\ Rate = \frac{FP + FN}{TP + FP + FN + TN} \tag{2.12}$$

Precision was shown in equation (2.13) is the ratio of the number of True Positive samples estimated as class 1 to the total number of samples estimated as class 1.

$$Precision = \frac{TP}{TP + FP} \tag{2.13}$$

Sensitivity was shown in equation (2.14) is the ratio of the number of correctly classified positive samples to the total number of positive samples.

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.14}$$

Specificity was shown in equation (2.15) is a testament to the ability of a test to distinguish what is really wrong.

$$Specificity = \frac{TN}{TN + FP} \tag{2.15}$$

Precision and sensitivity measures only are not enough to create a significant assessment. Estimating both measures collected gives extra precise results. The f-criterion is defined for this. The F-measurement was shown in equation (2.16) is the harmonic mean of precision and sensitivity.

$$F - measurement = \frac{2 \times Sensitivity \times Precision}{Sensitivity + Precision} \tag{2.16}$$

The ROC curve has become the standard method for the reliability and general accuracy of model tests. ROC analysis investigates the success of the model in discriminating positive groups from negative and the independence of the results from the frequency of positive subjects in the population. ROC analysis is particularly useful in estimating models and evaluating other tests because it captures a threshold value between sensitivity and specificity in a continuous range.

ROC curve is wrong for true positive ratio (sensitivity) for different threshold values positive ratio (1-specificity) function.

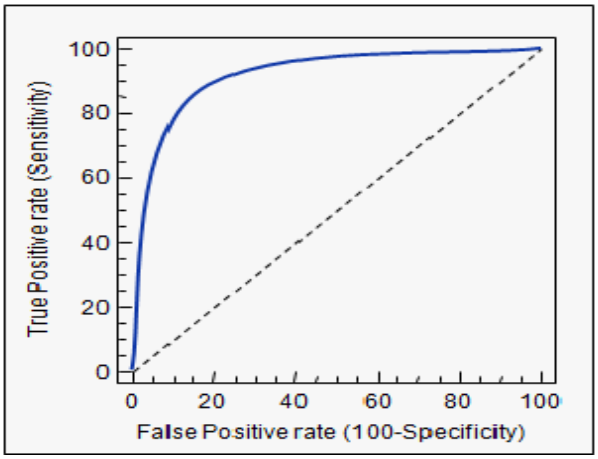

Figure 2.6 ROC Curve [51]

The total area under the ROC curve is the performance measure of the tests as it reflects the test performance at all possible threshold values. Since the AUC is a part of the area of the unit square, its rate will continuously be between 0 and 1.0. However, because chance predicting produces the diagonal line between (0,0) and (1, 1), which has an area of 0.5, no accurate classifier should have an AUC less than 0.5.

2.6 Data Mining Applications

Data mining can be originating in almost any field of application where data warehouses are built. As data mining develops, new and progressively novel applications occur for it. The applications of data mining are separated in the succeeding groups: Healthcare, Finance, Retail Industry, Telecommunication, Text Mining & Web Mining, Higher Education [52].

Information obtained as a result of data mining analysis is used for marketplace analysis, fraud detection, customer retention, production control and scientific discovery. The applications of data mining are shown in Figure 2.7 [27].

Figure 2.7 Common data mining application domains [27]

- **Marketing:** In today's increasingly intense marketing industry, customer retention and acquisition of new customers are used to determine customer buying habits, to increase market share and profit, and to determine the purpose and campaigns. Many data mining applications are used in the management of customer relationships, such as identifying the most profitable customers and determining the priority needs of these customers in order to strengthen their relationships.

- **Retail and Logistics:** Data mining is used for different purposes in retail and logistics. For instance, to forecast the sales of specific points in order to determine the right amount of inventory, to recognise the links between different products, to arrange products in the shop, to develop various promotions, or to determine the consumption levels of different product types in order to optimize transportation, sales, and techniques.

- **Banking and Insurance:** Credit card spending is used to identify customer groups, fraud and fraud detection, credit requirements and reimbursement. At the same time, data mining methods are being used to solve insurance-based problems, such as predicting which customers can buy, preventing wrong damages or determining fraud.

- **Telecommunication:** In many areas, such as identification of customer profiles, provision of appropriate campaigns and offers, and recovery of lost customers, data mining is being used assuming that keeping customers in a mature market, like the telecommunication sector, is much less costly than acquiring new customers.

- **Production:** Data mining is used to predict machine errors with the use of sensory data, to determine anomalies and generalizations in the production system to optimize the production capacity, and to discover different patterns in order to improve the quality of the product.

- **Medicine:** It is used to predict diseases, to take precautions, to provide early diagnosis of a disease, and for many other areas.

- **State and Defense:** Data mining methods can be performed more easily with the e-government application being implemented in our country. Data mining can be benefitted in many ways, for example, to enlighten suspect cases for security forces, to detect corruption related to tax or taxation, raising criminal charges and criminal profiling.

- **Tourism:** In order to determine optimal prices for services of many companies in the tourism sector, to estimate the demand at different locations, to allocate limited resources to the right places, to determine the customers with high profitability, data mining is used in such subjects as providing services.

- **Entertainment:** Data mining methods are used in resource planning to determine the most popular programs at the time of watching television and to find out the most profitable advertising, to take investment decisions by anticipating the success of films and series, to plan activities and to estimate demand.

- **Science and Engineering**: Today, scientific data is produced in high quantities in the process of simulating and analyzing systems in a laboratory or computer environment. Data mining provides a very convenient platform to understand the obtained data.

- **Customer Relationship Management (CRM):** Qualified information is extracted from the textual information obtained from all customers' access points such as e-mail, transaction, call center and questionnaire. This qualified information is used to estimate the customer's abandonment and cross-selling.

2.7 Data Mining with WEKA

Weka is an open source data mining program developed on the java platform. It was developed by Waikato University in New Zealand. Weka is a gathering of machine learning algorithms for data mining jobs. The algorithms can either be implemented straight to a dataset or called from your own Java code. Weka has tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also compatible for emerging new machine learning schemes [53].

WEKA is software that hosts machine learning algorithms and data preprocessing tools. It was designed to implement data mining methods on the database in an easy and flexible way. Experimental data mining provides comprehensive support for the whole process; data preparation, statistical evaluation of the learning scheme, and visualization and learning results [24].

2.7.1 WEKA Explorer

WEKA has several user interfaces. WEKA's basic graphical user interface is Explorer. Explorer, a panel-based interface, contains 6 panels corresponding to the data mining techniques supported by WEKA.

- **Preprocess:** The first panel. From this panel, the data set is selected and arranged in various ways. There are data editing tools that are expressed here as filters. Data can be uploaded in 3 ways; from file, database or URL. Supported data formats include; CSV, LibSVM.
- **Classify:** The second panel is the panel where the classification or regression algorithms are located. The reason for the paneling classification is that the continuous classes of regression techniques are seen as predictors. The panel performs independent cross-validation on the data set prepared in the data preparation panel with the selected learning algorithm to determine the predictive performance. It also shows the textual representation of the data set. The panel also provides a graphical representation of the model or decision trees if the conditions for the data are available. It also provides visualization of prediction errors in scatter graphs and evaluation of curves such as ROC. The model can also be permanently registered and reloaded in this panel.

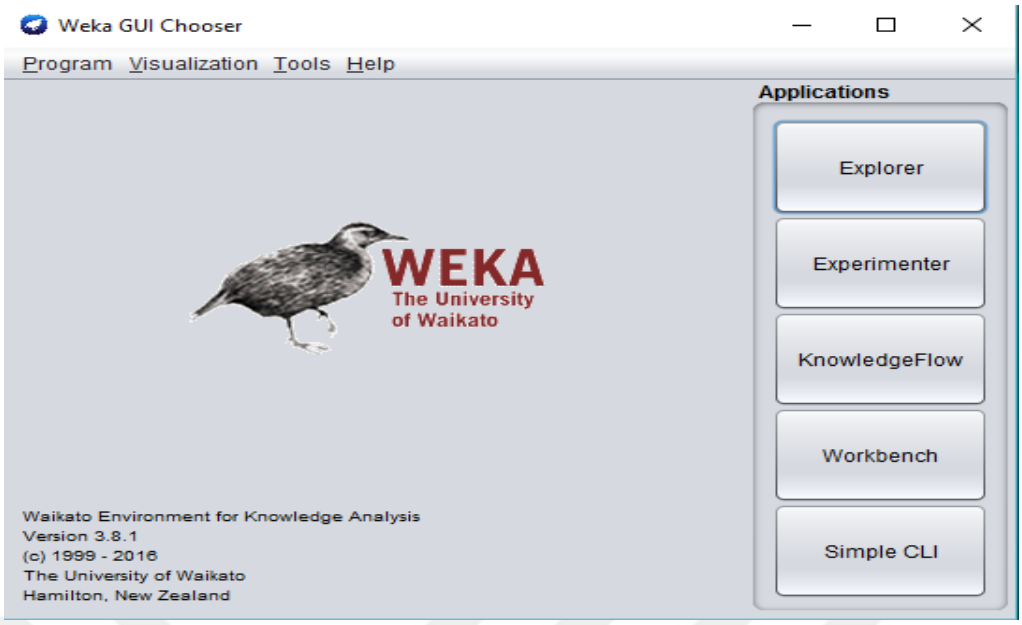

Figure 2.8 A screenshoot of WEKA Application Menu

- **Cluster:** Shows the data set loaded into the data preparation panel. When the clustering algorithms are performed, the WEKA shows how many clusters it has and the number of instances in each cluster. The panel provides simple statistics to evaluate clustering performance. If the data is appropriate, it is possible to visualize the clustering structure. The model can also be permanently registered.
- **Associate:** This panel, which learns and evaluates the cohesion rules on data, is easier to use than the clustering and classification panels. WEKA has 6 algorithms for coexistence rules.

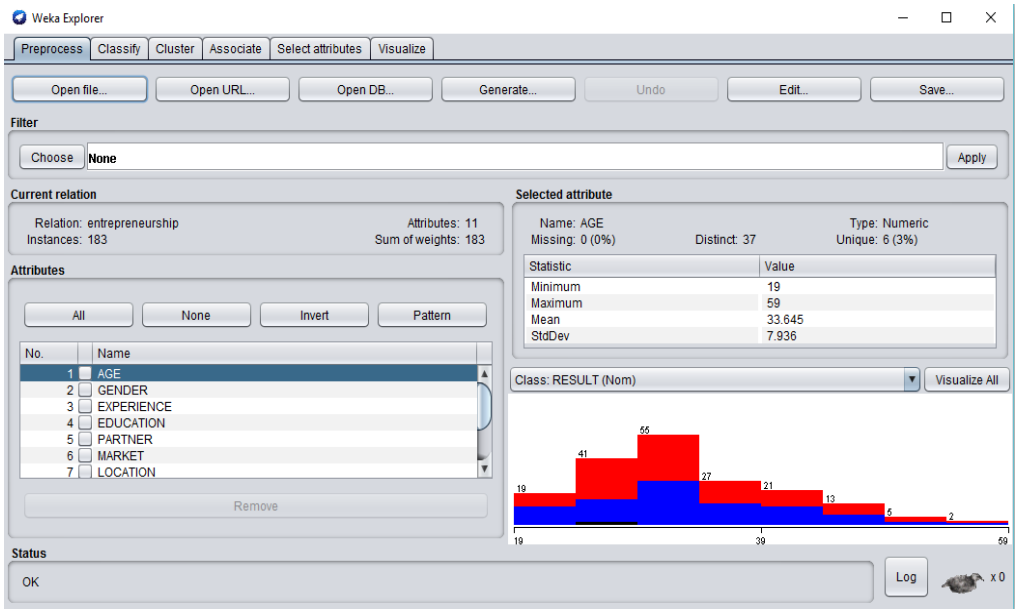

Figure 2.9 A screenshoot of WEKA Explorer Menu

- **Select attributes:** This panel provides access to a wide range of algorithms and evaluation criteria to determine the most important attributes within the data set. In this way, it is possible to combine different evaluation criteria with different search methods and to construct a wide variety of possible candidate techniques. The choice of qualification can be achieved by using the full training data set or by using the independent validation test.
- **Visualize:** Visualization helps visualize the data set. The point to note here is that the data set itself is visualized in this panel, not the results of the classification or clustering model. It shows scatter graph of all attribute pairs over a two dimensional matrix.

2.7.1.1 WEKA Classify Test Options

The outcome of implementing the selected classifier will be tested along with the choices that are set by clicking in the Test options box. There are four test modes [54]:

- Use training set: The classifier is estimated on how well it foresees the class of the examples it was trained on.
- Supplied test set: The classifier is estimated on how well it foresees the class of a set of examples loaded from a file. Clicking the Set button gets a window permitting you to select the file to test on.
- Cross-validation: The classifier is estimated by cross-validation, using the number of folds that are nominated in the Folds text part. To test the success of the implemented method, the dataset is divided into training and test clusters. Once the cross-validation has divided the data set by K values, one of the part will be selected for testing and used for the rest of the training.
- Percentage split: The classifier is estimated on how well it foresees a certain percentage of the data which is detained out for testing. The quantity of data detained out depends on the value nominated in the % part.

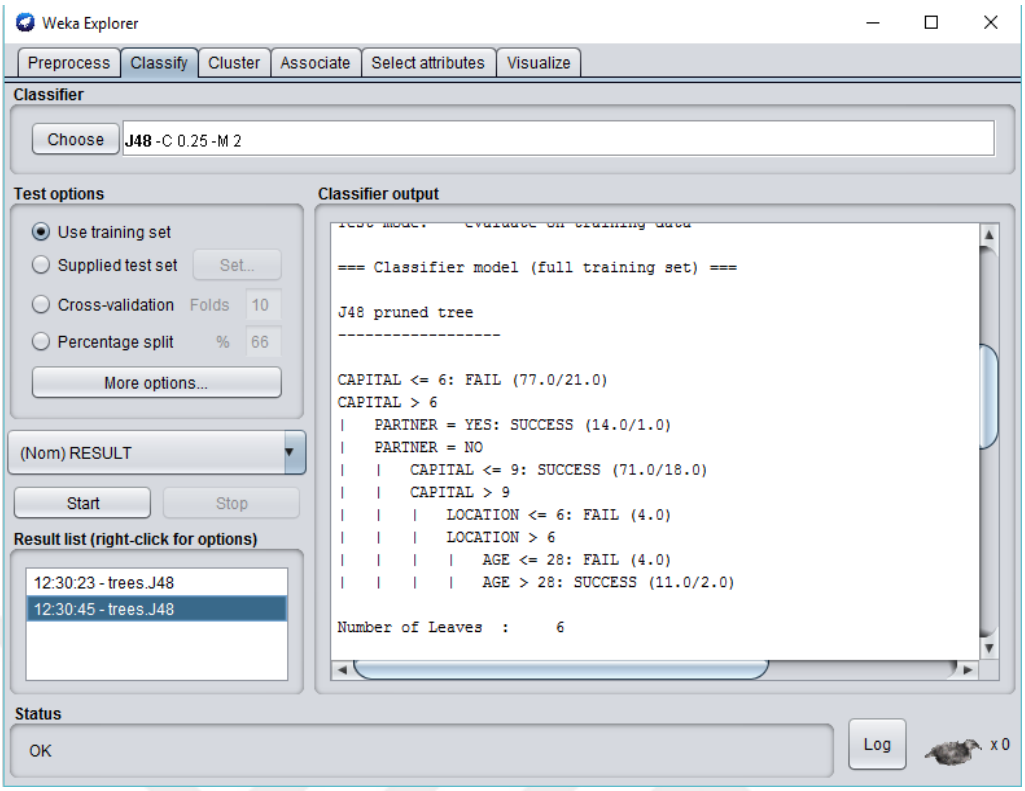

Figure 2.10 A screenshoot of Weka Test Options

2.7.2 WEKA Experimenter

In data mining applications, frequently developed algorithms show performance on more than one data set, or more than one algorithm compares performances on the same data set. Sometimes both situations can be seen in the same study. WEKA software can support this kind of work. This enables the designed tasks to be executed more efficiently, faster, and easier. Such studies can be performed with the "Experimenter" option of WEKA software (Figure 2.11). Work done can be saved to disk for re-use and configured or saved experiments can be run from the command line. On this screen, there are "Setup" where the algorithms to be studied can be selected and "Run" where the experiments can be executed and end of the executed work. Experiments can only be carried out on classification screens in this screen and are not suitable for clustering studies.

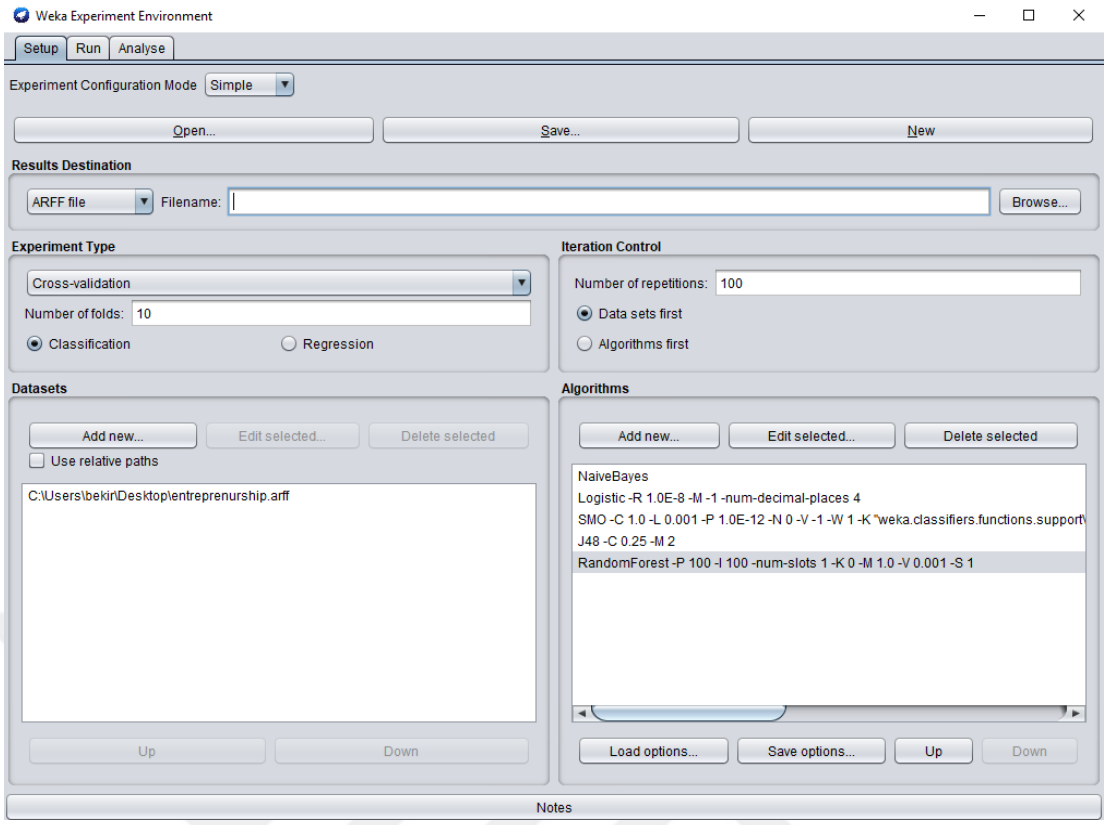

Figure 2.11 A screenshoot of Weka Experimenter

2.7.3 WEKA Knowledge Flow

"Knowledge Flow" is an alternative working environment that is developed in order to design the works that can be performed in the "Explorer" with visual drag-and-drop logic (Figure 2.12). The "Knowledge Flow" data, which collectively processes "Explorer" data on the screen which cannot be realized on the "Explorer" screen, has the processing feature to process gradually or collectively. This has "DataSources", "DataSinks", "Filters", "Classifiers", "Evaluation" tabs for large data for "Explorer". Intermediate operations to be performed are left to the "Knowledge Flow Layout" screen by drag-and-drop method, and operations are completed and then some extra features are executed. For example, this is a possible distress. On this screen and "Clusterers", "Visualization", right click on "Arff Loader" and the operation will be executed.

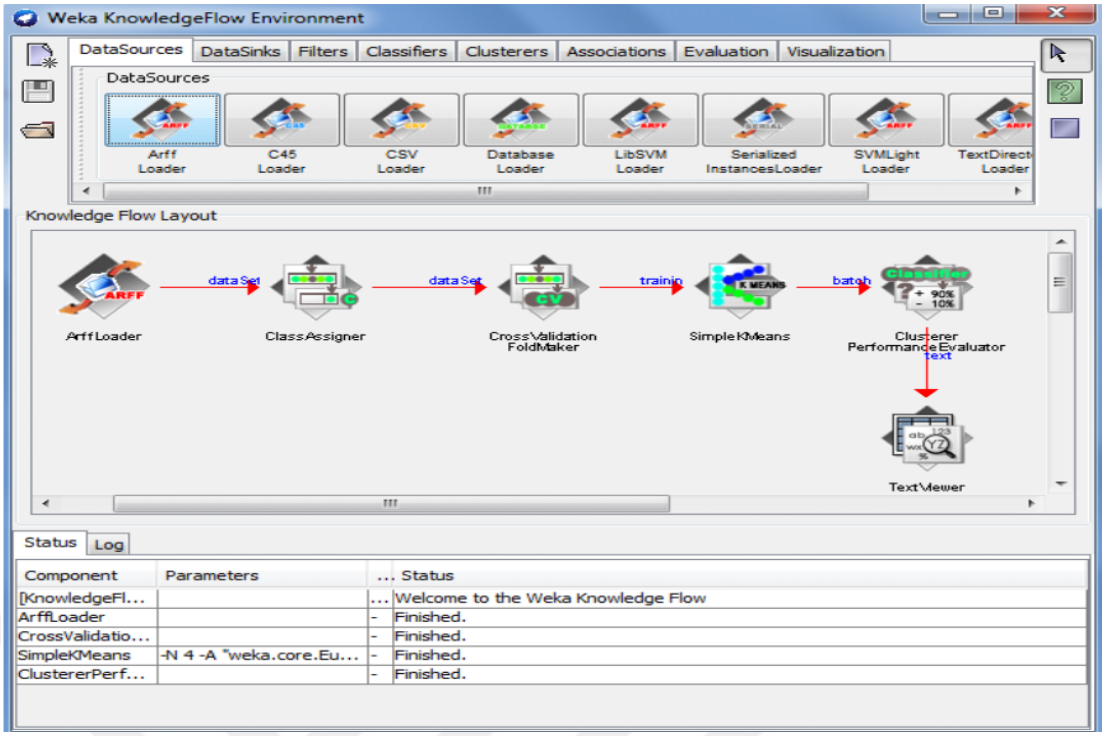

Figure 2.12 A screenshoot of Weka Knowledge Flow

CHAPTER 3

SMALL and MEDIUM ENTERPRISES DEVELOPMENT ORGANIZATION OF TURKEY (KOSGEB)

The Small and Medium Enterprises Development Organization was established on April 20, 1990 by Law No. 3624. KOSGEB is a private-budget public institution listed in Part B of Schedule II of the Public Financial Management and Control Law No. 5018, which is the "relevant" institution of the Ministry of Science, Industry and Technology [55].

KOSGEB has been established in order to increase the share and efficiency of small and medium sized enterprises (SMEs), to increase their competitiveness and their levels in meeting the economic and social needs of the country and to carry out integration in the industry in accordance with economic developments [56].

KOSGEB's strategic objectives are;

- To develop SMEs' production and management skills, to develop innovative and high value added products / services and to increase their competitiveness in the global market,
- To promote entrepreneurship culture, to develop entrepreneurship and to ensure the sustainability of entrepreneurship.

The aims set in line with these strategic objectives are:

- Improvement of R & D innovation activities and conversion of economic value will be ensured.
- Access to finance will be facilitated

- SMEs' production and management skills, culture of making business associations and contribution to the increase of productivity levels in the economy will be improved.
- International opportunities will be ensured by improving the facilities and capabilities for SMEs to access foreign markets.
- Support for prioritized issues in national policy documents will be improvedEntrepreneurship cultural expansion will be provided. Entrepreneurs having high chances of survival, growth, and business creation potential will be supported.
- Special target groups will be supported.
- Development of mechanisms to support entrepreneurs will be provided.

3.1 Entrepreneurship Support Program

The purpose of the Entrepreneurship Support Program is to support, disseminate and establish successful entrepreneurship, which is a key factor in resolving economic development and employment problems [57].

Entrepreneurship Support Program consists of four subprograms: Applied Entrepreneurship Training, New Entrepreneurship Support, Business Development Center Support and Business Plan Award [57].

Applied Entrepreneurship Training aims at disseminating entrepreneurship cultures, introducing entrepreneurs to the business plan concept and establishing successful businesses. At the end of training, entrepreneur candidates are aimed to gain necessary knowledge and experience to prepare business plans for their own business ideas. The target volume of Applied Entrepreneurship Training is the real people wishing to set up their own business, and the training can be open to general attendance as well as to a specific target group such as young entrepreneurs, women entrepreneurs, and university students.

Entrepreneurs who are entitled to obtain a certificate and completed Applied Entrepreneurship Training and who are involved in the ISGEM (Business Development Center) can apply for new entrepreneurship support. In the context of New Entrepreneur Support, the foundations supported by the board are Business

Enterprise Support, Establishment Machinery, Equipment, Office Hardware and Software Support, Business Expense Support and Fixed Investment Support.

The support elements of the New Entrepreneur Support Program, support upper limit and support rates are shown in Table 3.1.

Table 3.1 New Entrepreneur Support Program Limit [57]

<table>
<tr><th colspan="3">ENTREPRENEURSHIP SUPPORT PROGRAM SUPPORT</th><th>SUPPORT UPPER LIMIT (TL)</th><th colspan="2">SUPPORT RATE (%)</th></tr>
<tr><td colspan="4">NEW ENTREPRENEUR SUPPORT</td><td>Region 1 and 2</td><td>3rd, 4th, 5th and 6th Regions</td></tr>
<tr><td colspan="3">Business Foundation Support</td><td colspan="3">2.000.-TL</td></tr>
<tr><td colspan="3">Machinery, Equipment, Software and Office Hardware Support</td><td>18.000*</td><td rowspan="4">%60

If the entrepreneur is a woman, a veteran, close to a martyr in the first instance or disabled
%80</td><td rowspan="4">%70

If the entrepreneur is a woman, a veteran, close to a martyr in the first instance or disabled
%90</td></tr>
<tr><td colspan="3">Business Expense Support</td><td>30.000*</td></tr>
<tr><td colspan="3">Fixed Investment Support</td><td>100.000</td></tr>
<tr><td colspan="4">BUSINESS DEVELOPMENT CENTER SUPPORT</td></tr>
<tr><td colspan="3"></td><td></td><td colspan="2">%60</td><td>%70</td></tr>
</table>

ISGEM Foundation Support	Building Renovation	500.000		
	Furniture Hardware	100.000		
	Personnel expenses	50.000		
ISGEM Business Support	Personnel expenses	100.000		
	Education-Counseling	50.000		
	Small Renovation	20.000		
	Promotion organization expenses, access expenses to collaboration networks	30.000		
WORK PLAN AWARD				
First Prize		25.000		
Second prize		20.000	**%100**	
Third Prize		15.000		

Business Development Center has been established and operated to provide businesses with some services such as business development coaching, access to support networks, access to financial resources, business facilities, office equipment, and office services. The Business Development Center is established by municipalities, higher education institutions, private administrations, incubators and professional organizations, either solely or jointly.

The Business Plan Award Competition is the awarding of successful candidates among the students who have taken the "Entrepreneurship" course in the training organized by the universities, by evaluating the business plans prepared in accordance with the form determined by KOSGEB. The students who rank first three are awarded with 25.000 TL, 20.000 TL, and 15.000 TL respectively, with the condition to register within 24 months.

3.2 Support Process

Entrepreneurship support consists of 4 stages as shown in table 3.2. Applied Entrepreneurship Trainings, in accordance with the general objective of dissemination of entrepreneurship culture in the country and establishment of successful enterprises; are organized to enable entrepreneurs to have knowledge and skills in business establishment and execution, to be aware of their own roles and responsibilities in this process and to acquire the knowledge and experience to prepare a business plan for their own business ideas. Applied Entrepreneurship Training includes in-class workshops and 32 (thirty-two) lesson hour workshops at least. Trainings given to entrepreneurs in Applied Entrepreneurship Training is the concept of entrepreneurship, and experience sharing related to the business idea development and creativity exercises, business concept, business functions, types, organizational forms, financial and legal responsibilities, business plan concept and items (market research, marketing plan, production plan, management plan, financial plan), workshop studies on business model and work plan. Applicative Entrepreneurship, the first stage of the support process, ends with certification given to entrepreneur candidates who have participated in the training.

After establishing the entrepreneurial business, entrepreneur registers online on KOSGEB website. The entrepreneur will register online and fill the Business Plan online and send it to the expert for review.

The expert evaluates the Business Plan by rating the entrepreneur according to the educational status, the sector in which it operates, its market, its competitor analysis, where it operates, and its financial position. Entrepreneurs whose Business Plan has been prepared completely are sent to the approval of the manager. The manager will send the Business Plan to the committee after the last checks.

Table 3.2 Process of Entrepreneurship Support Program

Training	Applied Entrepreneurship Training
	Certificate acquisition
Startup	Business Opening
	Application to KOSGEB
	Preparing a Business Plan
Assesment	Evaluation of Business Plan by Expert
	Sending Appropriate Business Plan to Managers
	Sending Business Plan to the Committee
	Decision of the Committee to Approve or Reject the Business Plan
Goverment Support	Buying Acceptance Materials
	Entrepreneur Visit and Identification
	Payment of Support Amount

The committee reviews the Business Plan according to the Application Evaluation Criteria and interviews the entrepreneur. Supports are given to entrepreneur candidates who meet the criteria and who are successful in the interview. Failed entrepreneur candidates are rejected.

Entrepreneurs who are decided to be supported are visited by the Expert after they have purchased the supported expenses. The support process will be realized by paying the expenses that are compatible with the support given by the committee.

CHAPTER 4

EXPERIMENTAL RESULTS

In this section, classification algorithm is applied to entrepreneurship database. The contents of the characteristics of the entrepreneurs in the database stated in the application are examined in detail. After applying the algorithms to the dataset, we made the comparison and found the algorithm that gave the best results.

4.1 Dataset Description

In this study, the characteristics of the business plan evaluated by experts were used as the data. The number of entrepreneurship projects examined by KOSGEB evaluation committee between 2012 and 2014 is 183.

- Age:

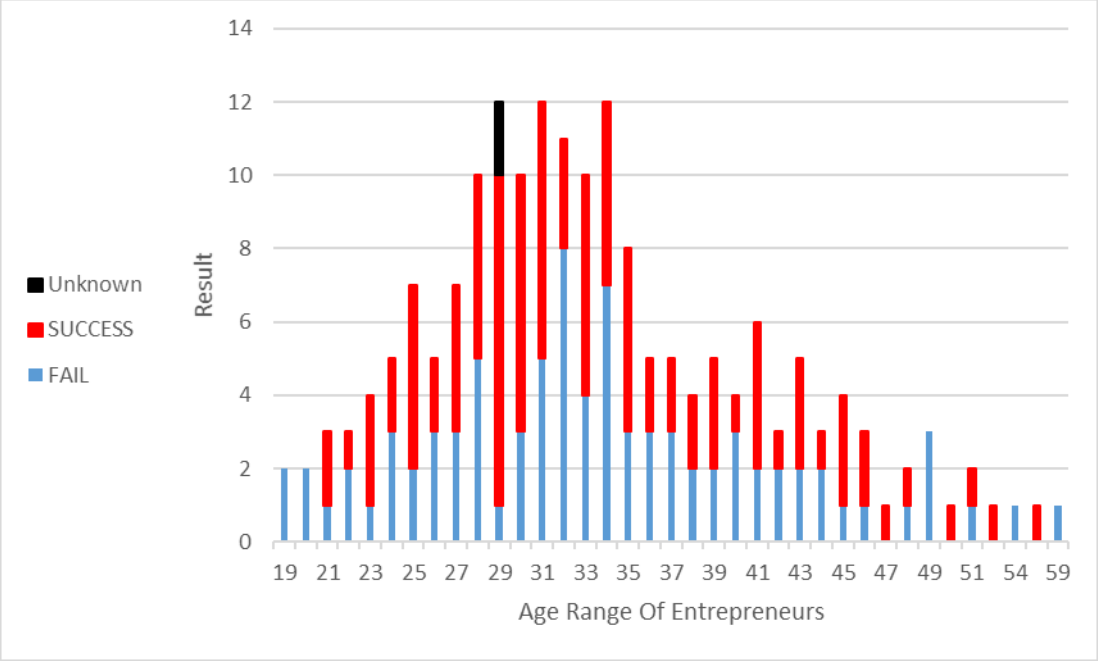

Figure 4.1 Age Graph of Entrepreneurs

The age range of entrepreneurs in the database is between 19 and 59, because entrepreneurs have to be older than 18 when they are supported. 52% of the entrepreneurs are between the ages of 24 and 34 because of the positive discrimination towards young entrepreneurs. Entrepreneurial ages entered in the business plan as day, month and year while the database was being prepared, were normalized to only age. The red color in the graph shows successful entrepreneurs who failed blue color.

- Gender:

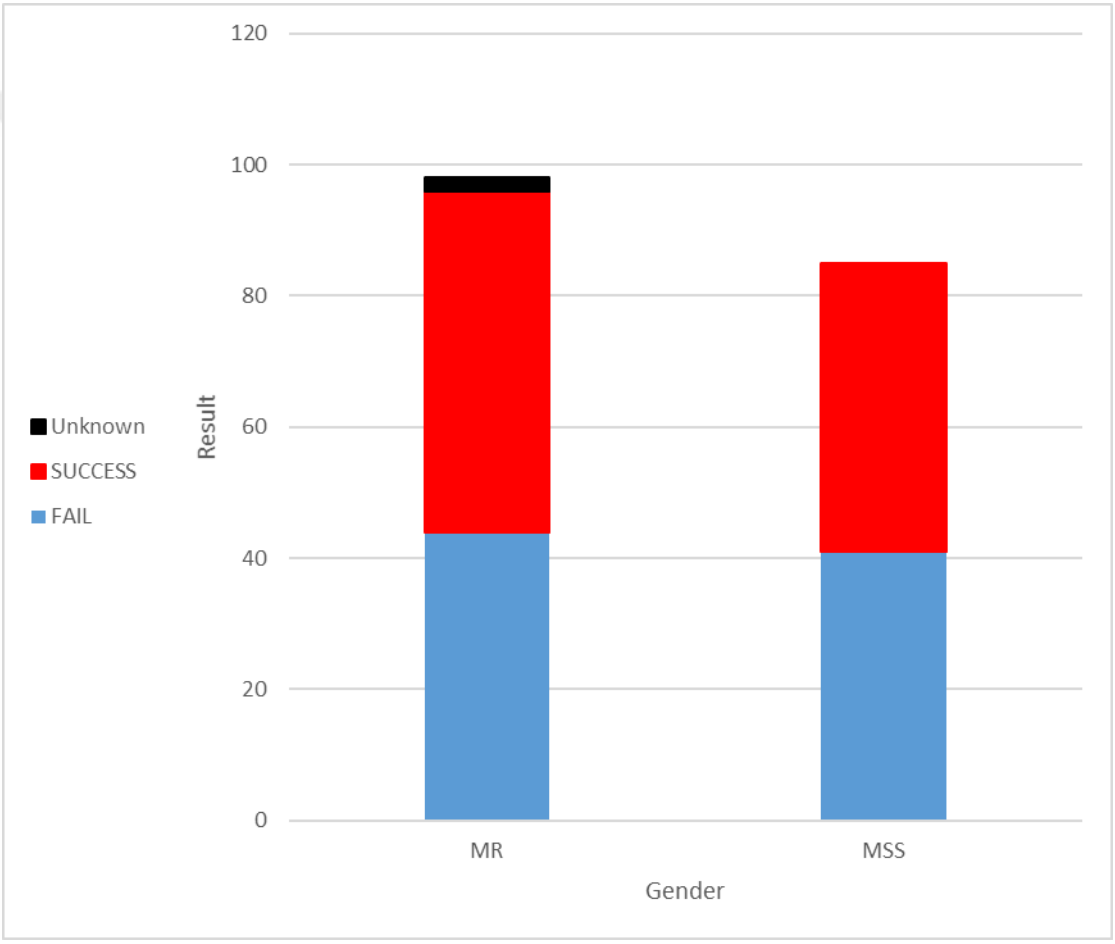

Figure 4.2 Gender Graph of Entrepreneurs

98 of the entrepreneurs are male and 85 are female. Entrepreneurial numbers are close to each other. Although the female entrepreneurs were given more support, they were behind the men in number of support applications. While preparing the database gender was selected by entrepreneur candidate who want to take

government support, but some entrepreneurs were in the wrong choice. These incorrect entries have been fixed.

- Experience:

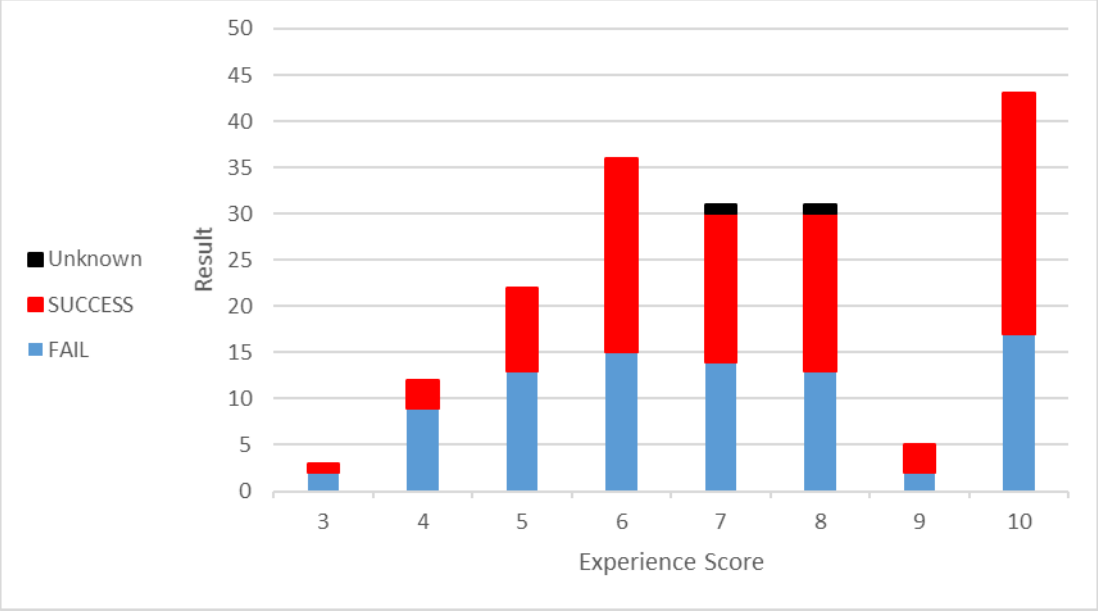

Figure 4.3 Experience Graph of Entrepreneurs

While experts evaluate the business plan, they score according to the entrepreneur's previous experience. The time they have previously worked in the business they want to establish is important for scoring.

Table 4.1 Experience Score Table

Experience	**Unit**	**Score**
Number of year of before experience	year	{low, medium, high} { <5, 5-7 ,>7}

If the entrepreneur had a business related trained, if he / she has a business-related certificate and has previously worked in the same business, the score is high. If the entrepreneur has no experience, the score is low. Figure 4.3 shows that the experience of the supported entrepreneurs is high.

- Education:

The level of education and knowledge of entrepreneurs are essential for scoring. For this reason, the educational status of entrepreneurs in the business plan is taken as data. Figure 4.4 shows the first bar as elementary school, the second bar as graduate, the third bar as high school, and the last bar as undergraduate. The level of education

of entrepreneurs is evenly distributed, but the ratio of graduate education is very low compared to others.

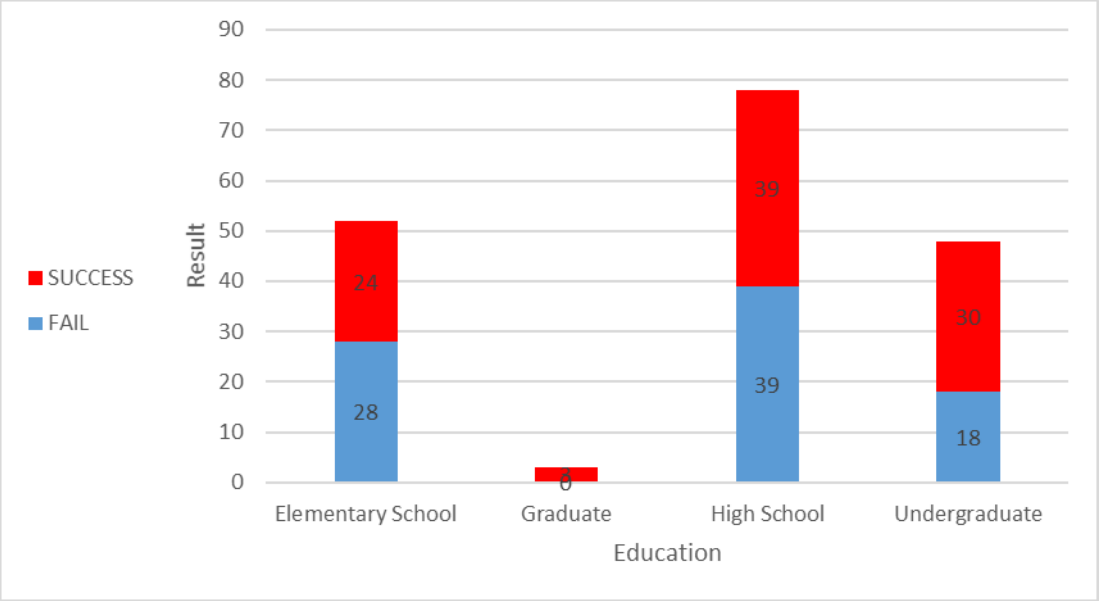

Figure 4.4 Education Graph of Entrepreneurs

- Partner:

When establishing a business, entrepreneurs can work alone or with a partner. As seen in Figure 4.5, there are 15 entrepreneurs who set up their operations with a partner. Entrepreneurship is influential in the scoring because it is more useful to cooperate with a partner who knows the business.

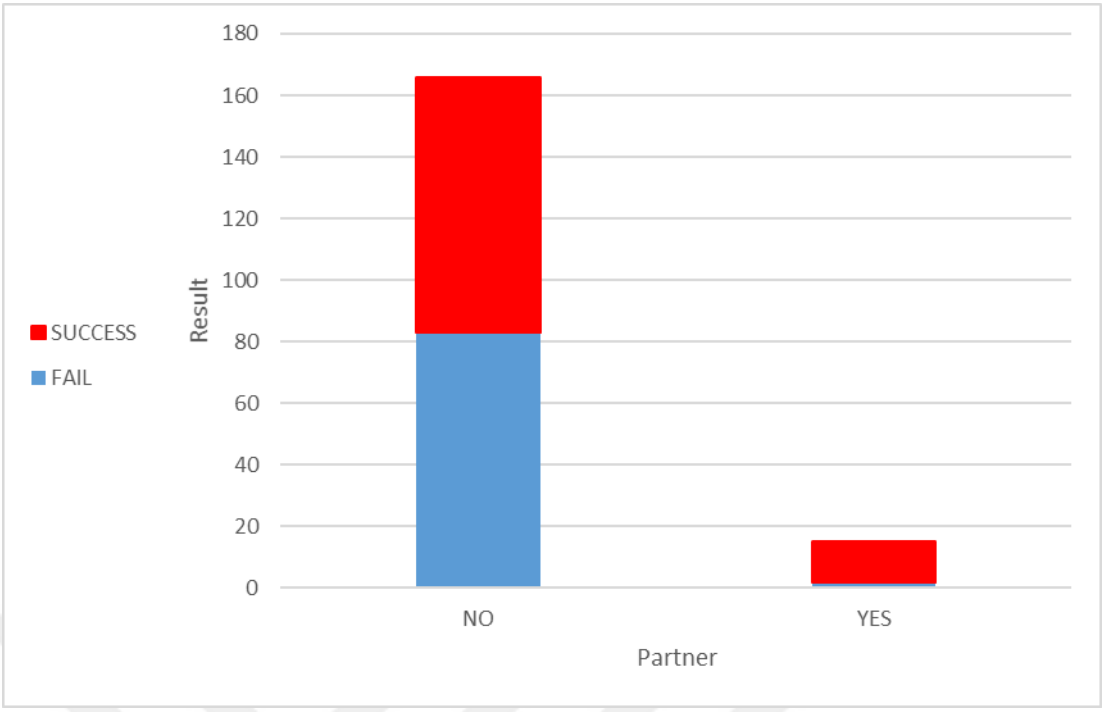

Figure 4.5 Partner Graph of Entrepreneurs

- Market:

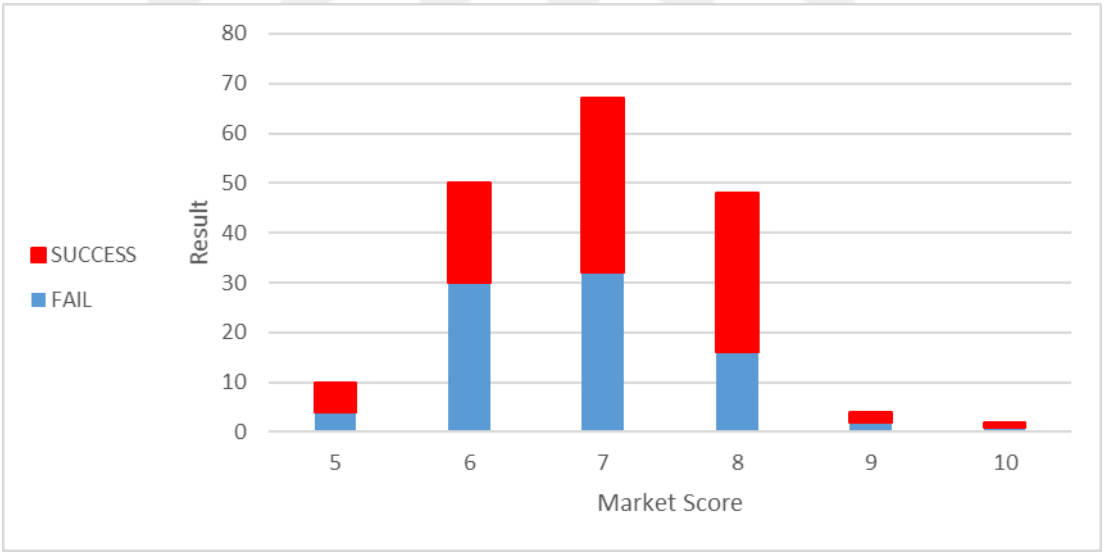

Figure 4.6 Market Graph of Entrepreneurs

The market for entrepreneurs is one of the most important criteria. It is important for entrepreneurs to make sales and make money so that they can continue their lives, and the market in which they will be able to survive is also important. As the experts evaluate the business plan, the market to be sold will be evaluated and scored between 1 and 10 according to the business established. Entrepreneurs will have

difficulty in selling if the supply demand balance of the product which they want to sell is low. In this case, entrepreneurs will be given low scores because their success will be low. If there is enough demand in the market where the entrepreneur is going to sell, the entrepreneur's success rate will increase and the market scoring will be higher accordingly.

- Location:

Locations where entrepreneurs set up business in are important because of factors such as transportation to raw materials, access to market, cheap labor power and energy. The location of the entrepreneur is scored between 1 and 10, depending on whether the location is established in the area concerned with the job. Entrepreneurs have often been choosing the right location, as shown in Figure 4.7, as their location for their business. The correct selection is gave in the score because it will bring success.

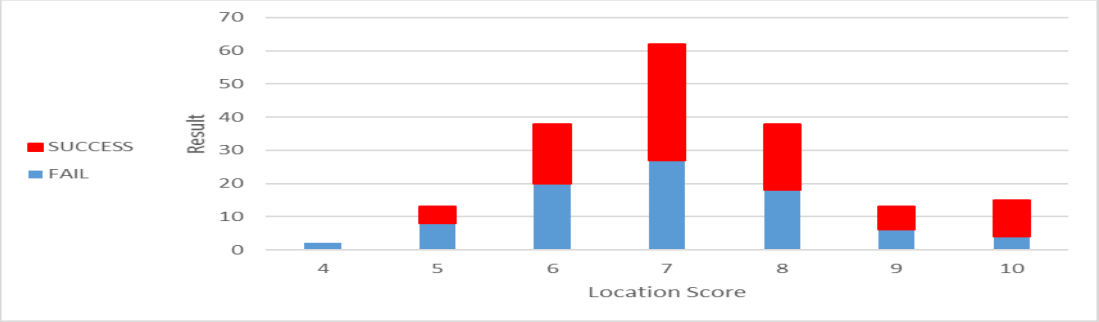

Figure 4.7 Location Graph of Entrepreneurs

- Sector:

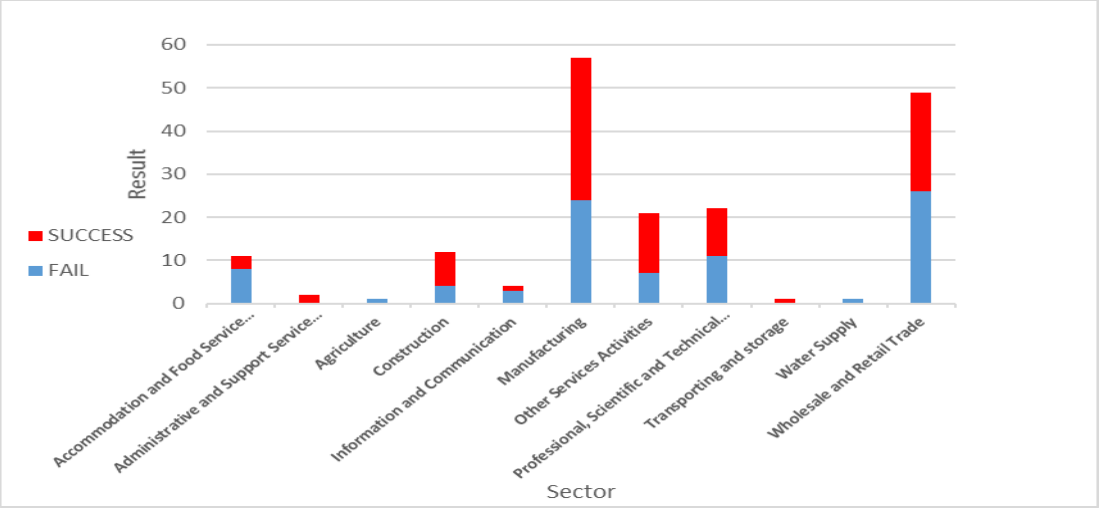

Figure 4.8 Sector Graph of Entrepreneurs

When establishing a business, they have to issue levy according to the activity areas they have chosen. As KOSGEB gives more support to manufacturers and new ideas, the surplus of entrepreneurs who choose this sector appears in Figure 4.8. 31.6% of the entrepreneurs operate in manufacturing, 26.7% in trade and 12% in sectors that require professional experience. It is due to the positive discrimination towards the manufacturers that the majority of the entrepreneurs are in the manufacturing sector.

- Staff:

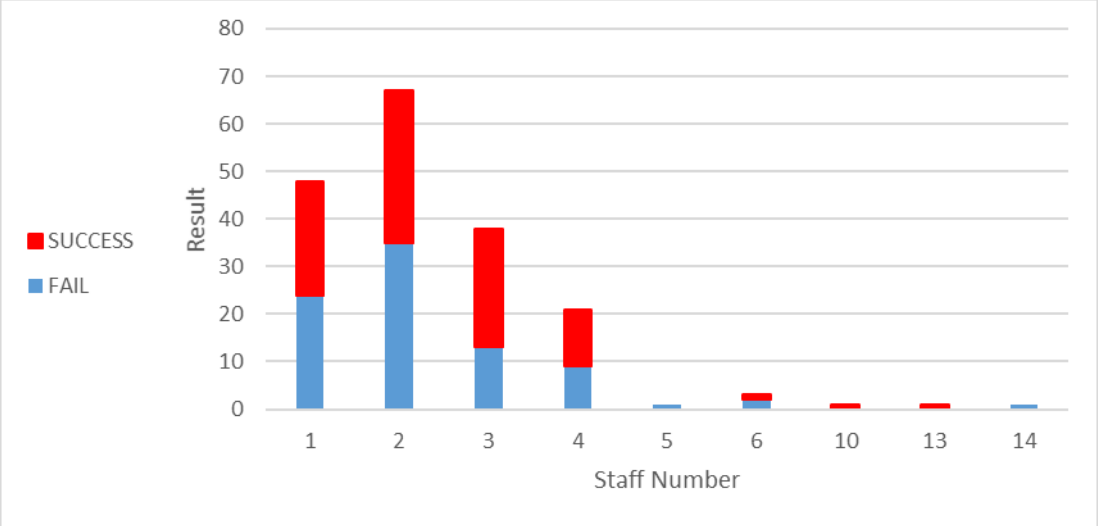

Figure 4.9 Staff Graph of Entrepreneurs

While entrepreneurs are starting business, having enough skill workers is effective for success. As can be seen in Figure 4.9, the vast majority of entrepreneurs have started with employees between 1 and 3. When entrepreneurs start to business, the right planning of the number of workers is important in terms of efficient use of limited resources and success.

- Capital Score:

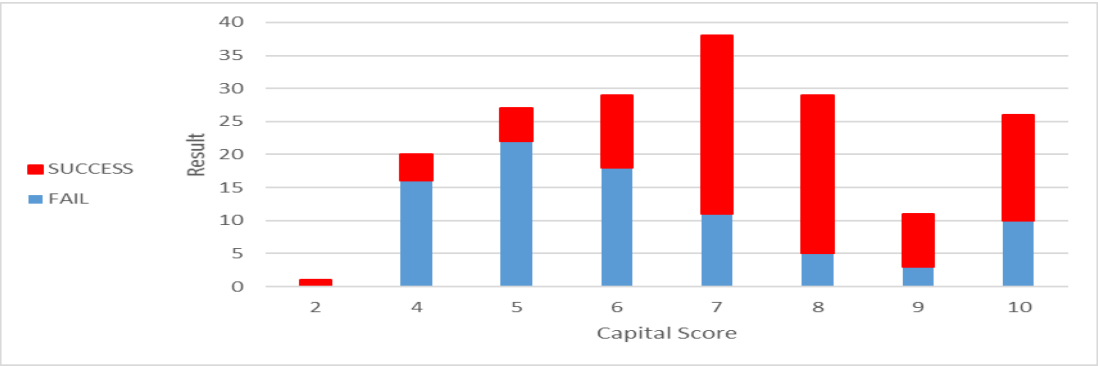

Figure 4.10 Capital Score Graph of Entrepreneurs

One of the most important features of the success of entrepreneurs is capital. Entrepreneurs must prepare a detailed financial statement when they fill out the business plan. The Expert gives score according to whether they have substantial capital for their business by looking at the analytics in the business plan. Entrepreneurs are assessed in detail according to the equity they own, the credits they have received and the debts they have owed. Entrepreneurs are scored according to whether they have adequate capital. Entrepreneurs with a nominal capital are given low points if they do not have high enough capital.

- Result:

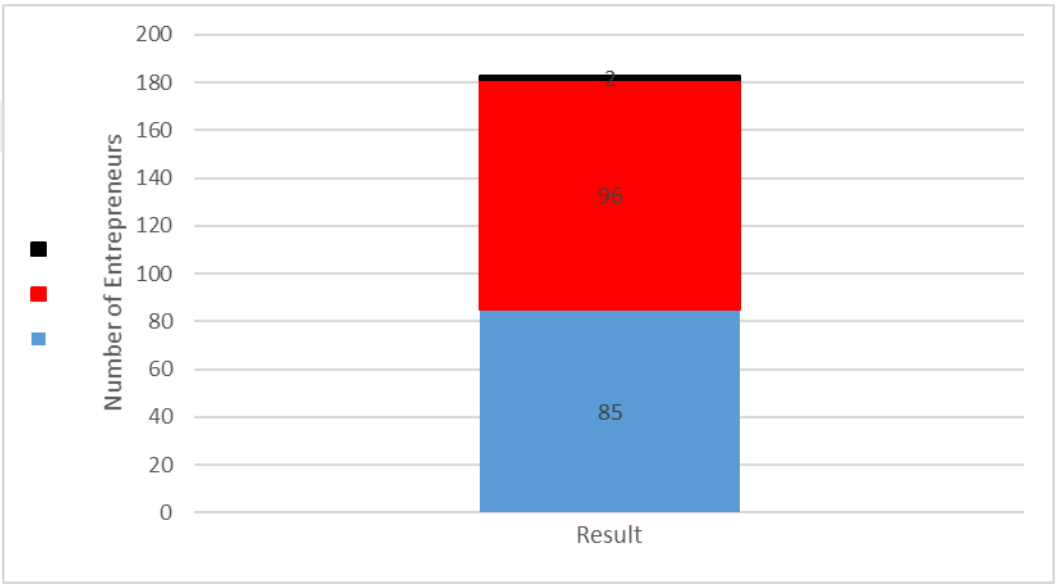

Figure 4.11 Result Graph of Entrepreneurs

Figure 4.11 shows successful or unsuccessful enterprises in accordance with whether they have been open at the end of four years. 181 entrepreneurs in the database were entered fully and 2 were as missing data since they could not be reached. Of the 181 entrepreneurs, 96 were successful and 85 were unsuccessful.

4.2 Preparing the Data for Application

Once transferred to the excel file, the entrepreneur data is made available to the WEKA program and run on the system. Attributes are defined in the WEKA program.

@relation entrepreneurship

@attribute AGE numeric

@attribute GENDER {MR, MSS}

@attribute EXPERIENCE numeric

@attribute EDUCATION {1,2,3,4}

@attribute PARTNER {YES, NO}

@attribute MARKET numeric

@attribute LOCATION numeric

@attribute SECTOR {A, C, E, F, G, H, I, J, M, N, S}

@attribute STAFF numeric

@attribute CAPITAL numeric

@attribute RESULT {FAIL, SUCCESS}

Once the attributes are defined, data is entered sequentially after the @data command.

Table 4.2 Dataset Attributes

Attributes names	**Attribute Type**	**Units**
Age	numeric	number
Gender	binary	Mr, Mss
Experience	numeric	{1-10}
Education	numeric	{1-4}
Partner	binary	Yes, No
Market	numeric	{1-10}
Location	numeric	{1-10}
Sector	nominal	{A, C, E, F, G, H, I, J, M, N, S}
Staff	numeric	{1-99}
Capital	numeric	{1-10}
Result	binary	Fail, Success

4.3 Implementation of Methods

In this part of the study all algorithms are applied to the prepared data set. The data is run with the WEKA program. In this section, accuracy, true positive rates, false positive rates and F-measures of the algorithms were found for each algorithm first.

For these values of the algorithms, cross-validation 10 folds was applied. Then, each algorithm was compared to each other for 100 iterations, and the results were compared. Then ROC areas were calculated and interpreted.

4.3.1 Implementation of C4.5 Algorithm

The C4.5 algorithm is one of the most preferred decision tree algorithms. It is preferred because of its high accuracy and visual results. The results of C4.5 algorithm applied to the dataset are shown in Table 4.3.

Table 4.3 C4.5 Algorithm Result

Accuracy	Class	TP Rate	FP Rate	F-Measure
71.82	Fail	0.659	0.229	0.687
	Success	0.771	0.341	0.744

The tree obtained as a result of the algorithm is shown in figure 4.12 and capital is determined as the most important feature.

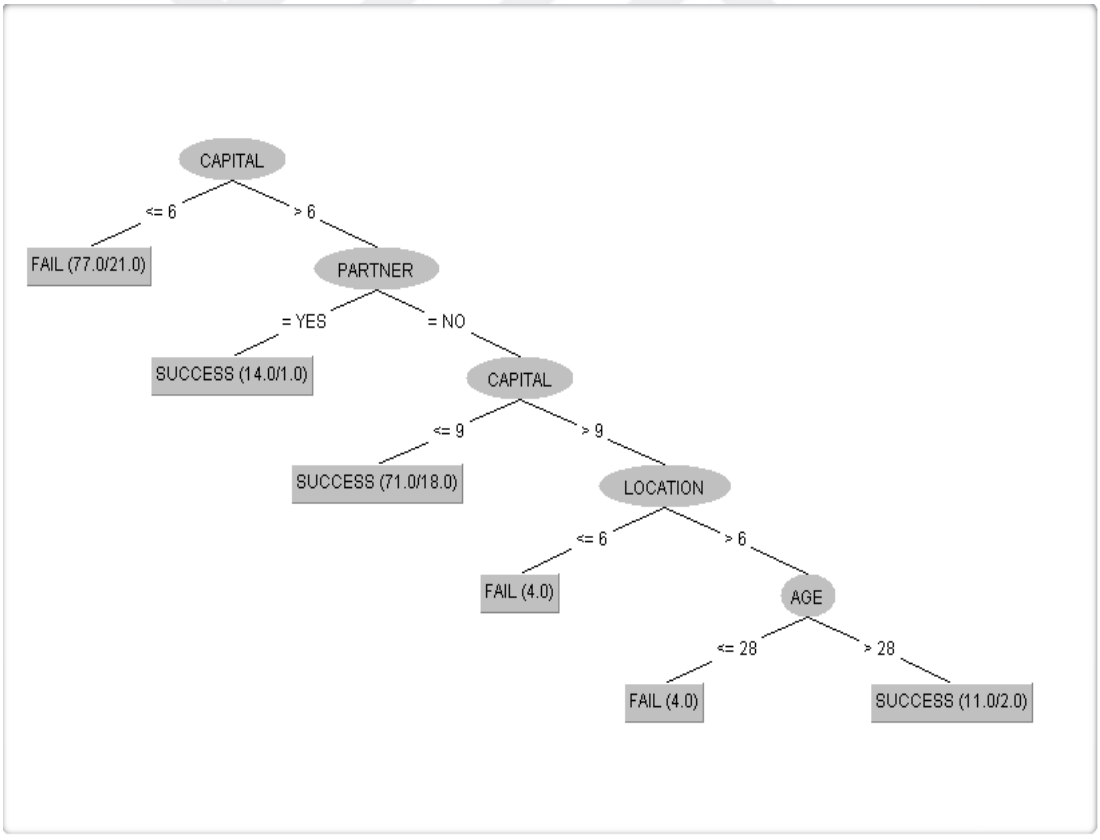

Figure 4.12 C4.5 Algorithm Tree

4.3.2 Implementation of Naive Bayes Classifier

The results are obtained by applying Naive Bayes classifier algorithm, which tries to determine the category of the new test data with the probability procedures based on the previous training data in the system, and is shown in table 4.4.

Table 4.4 Naive Bayes Classifier Result

Accuracy	Class	TP Rate	FP Rate	F-Measure
62.98	Fail	0.588	0.333	0.599
	Success	0.667	0.412	0.656

4.3.3 Implementation of Logistic Regression

Logistic Regression analysis generally aims to produce a model that predicts the value of a variable using certain variables. As a result of the analysis, the realization probability of the values is calculated and classification is based on these probabilities. The results of the regression analysis classification are shown in Table 4.5.

Table 4.5 Logistic Regression Classifier Result

Accuracy	Class	TP Rate	FP Rate	F-Measure
66.29	Fail	0.635	0.313	0.639
	Success	0.688	0.365	0.684

4.3.4 Implementation of Random Forest

Table 4.6 shows the results of the random forest algorithm, which generates multiple classifiers instead of just one classifier, and then calculates the votes from their estimates and the new classifier learning algorithm.

Table 4.6 Random Forest Classifier Result

Accuracy	Class	TP Rate	FP Rate	F-Measure
67.4	Fail	0.659	0.313	0.655
	Success	0.688	0.341	0.691

4.3.5 Implementation of Support Vector Algorithm

Table 4.7 shows the results of the support vector algorithm for achieving the optimal separation hyperplane to divide the classes.

Table 4.7 Support Vector Classifier Result

Accuracy	Class	TP Rate	FP Rate	F-Measure
67.95	Fail	0.659	0.302	0.659
	Success	0.698	0.341	0.698

4.3.6 Comparison of Algorithms

All the algorithms applied to the dataset were selected from the Weka Experimenter panel and 100 iterations were made for each algorithm. Because 10-fold cross-validation was selected as an Experimenter type, 1000 results were obtained for each algorithm. Paired T-Test (corrected) was used to analyze the results and significance 0.01 was selected.

Table 4.8 Comparison of Result

Naive Bayes	Logistic Regression	Support Vector	C4.5	Random Forest
61.42	62.96	64.06	70.75	66.37
(v / / *)	(0/ 1 / 0)	(0/ 1 / 0)	(1/ 0 / 0)	(0/ 1 / 0)

Analyzing the results in Table 4.8, the other algorithms were compared according to the naive bayes algorithm. C4.5 algorithm was more successful than the based algorithm. Other algorithms could not be compared to the naive bayes algorithm in terms of good or bad results.

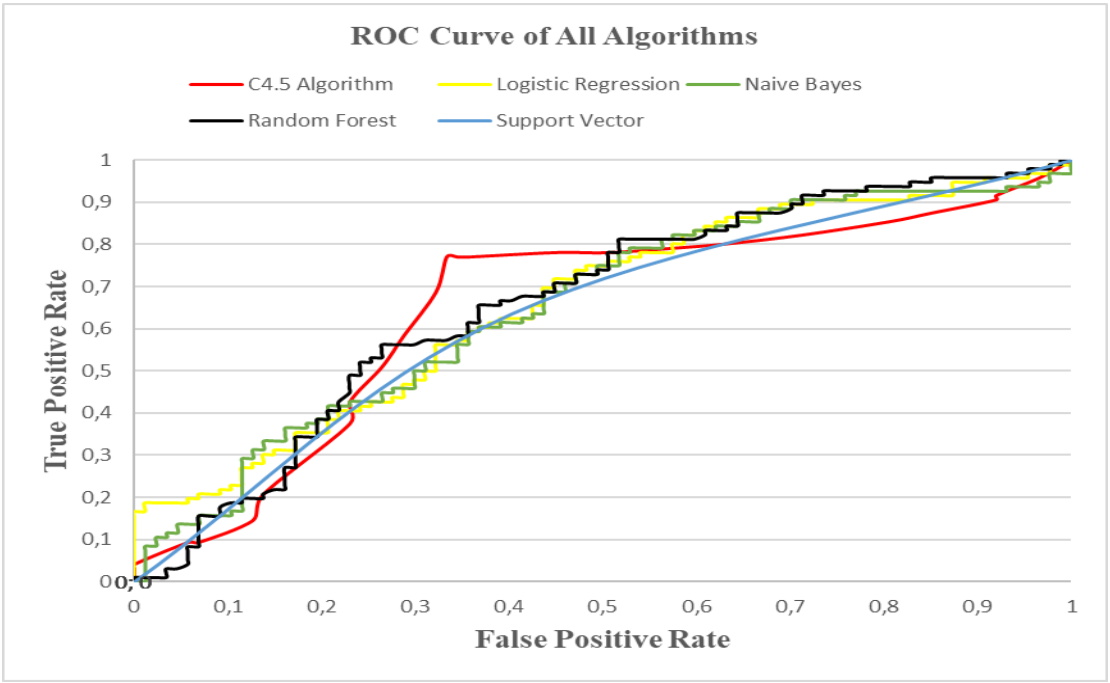

Figure 4.13 ROC Curve of All Algorithms

One of the most commonly used method of comparing the algorithms is the ROC curve. As the area under the curve increases, the success of the algorithm increases as well. Results indicated that the area under ROC was 0.65 for the Naive Bayes algorithm, 0.66 for the Logistic Regression algorithm, 0.64 for the Support Vector algorithm, 0.71 for the C4.5 algorithm (j48 in WEKA), and 0.69 for the Random Forest algorithm, respectively. The ROC analysis show that the most successful algorithm was C4.5 algorithm.

The most popular and simple method used to measure model performance is the accuracy of the model. The correct number of classified samples (TP + TN) is divided by the total number of samples (TP + TN + FP + FN). Based on the accuracy calculations, we were able conclude that the most accurate estimate obtained through the C4.5 algorithm was 71,82.

Precision and sensitivity measures alone are not enough to make a meaningful comparison. Evaluating both measures together gives more accurate results. F-measurement is a harmonic average of precision and sensitivity. When the F-measurement values were taken into account, the algorithm the C4.5 gave the best results as 0,744 for Success and 0,687 for Fail class.

The results of the C4.5 algorithm, which achieved the most successful result, were shown in detail in Figure 4.12. As a result of the algorithm, the most significant feature was capital. As capital score increases from 6 to 10, the success rate increases as seen in Figure 4.10. Similarly, in the algorithm result, it was obiviously seen that a majority of businesses with a score less than 6 were unsuccessful. When the algorithm determined the next leaves, the businesses with a capital score of 6 and above, which were likely to give a successful result, were identified. According to the result of the algorithm, another important feature that affected success was the structure of the partnership. It was clear that if the capital score was higher than 6, the success rate of the entrepreneurs who established their businesess with a partner was very satisfaying. Age and location were other features that affected the algorithm. But these features were not as decisive as capital and partnership.

CHAPTER 5

CONCLUSION

With today's increasing volume of data, its process is getting more challenging day by day. Data mining allows processing high amounts of data and acquiring meaningful patterns. Data mining, which is used in many different sectors, can also be used for the public support processes. One of the data mining techniques, the classification technique, allows the data to be categorized according to predetermined output. The outputs are known and the analysis is supervised because they are known in advance. In this study, 5 algorithms, which are most commonly preferred in classification, were applied to the dataset and it was found that C4.5 gave the best result.

In this study, the data of 183 entrepreneurs supported by KOSGEB between 2012 and 2014 were examined. In order to obtain the results, the scores given to the business plans of the entrepreneurs by the experts were benefited in the algorithm. Age, gender, experience, education, partner, market, location, sector, staff, and capital characteristics of the entrepreneurs were analyzed and evaluated according to whether they were successful or not.

When the graphs showing age was examined, it was observed that entrepreneurs in the age range of 24-29 were supported more. However, it was understood from the obtained result that age did not affect success significantly.

Graphs of gender showed that 98 of the entrepreneurs were male and 85 of the entrepreneurs were female. Looking at the success of the entrepreneurs, it was observed that gender had no direct influence on the success status. Despite the positive discrimination for female entrepreneurs, their success rates were similar to

male entrepreneurs' and this means that positive discrimination may not work as intended.

When the experience of entrepreneurs was analyzed, it can be deduced that the more experience is, the more success rate is. However, this feature did not affect the obtained tree. According to the analysis, the underlying reason here is that the experienced entrepreneurs did not have sufficient capital. In addition, generally the committee is hesitant to support less experienced entrepreneurs during the evaluation of business plans. In this study the features of the supported entrepreneurs were benefited as the dataset. It was seen that 80% of the entrepreneurs have good experience points than on the average. Therefore, experience was not a distinctive feature to affect the tree.

When the educational status of entrepreneurs was evaluated, it was seen that the supported entrepreneurs were evenly distributed in the levels of the primary school, high school, and under graduate. However, this factor did not have a direct effect on success. Only 3 of the 183 entrepreneurs had graduate level of education, and 3 of them were successful, too. Actually, entrepreneurs with graduate degrees were observed to be successful. The reason for not affecting the tree was due to the insufficient number of entrepreneurs who had graduate degrees in the dataset. In other words, 3 entrepreneurs were not enough to generalize the result.

In the partner graph it was seen that almost all of the entrepreneurs who started to work with a partner were successful. Having analyzed the tree algorithm results, the impact of the partnership situation on success was proved. When estimating the algorithm, the partnership was in the second rank after the capital.

According to the literature and experts' opinion, the market of the entrepreneurs was supposed to be important in terms of making money and succeeding. However, it was realized that there was no success effect of the market when looked at the graph and the result of the algorithm. The most important reason for this was the emphasis on the importance of the market given in entrepreneurship education. When looking at the market graph, it was observed that the majority of the entrepreneurs made the right business place choice.

Given the location where entrepreneurs established their business, few showed that they had established in the wrong place. As a result of the algorithm obtained, the location has emerged as a feature that influenced the tree. Entrepreneurs who established their business in the correct location have become more successful.

When sector graph was reviewed, it was observed that entrepreneurs who received KOSGEB support were the ones who were manufacturing and operating in a technical business line. Although KOSGEB supports many sectors, the reason for importance in this field is its contribution to the economy.

The most important feature that influenced the success of the entrepreneurs was the capital. When the amount of capital increases, the success rates of entrepreneurs increased as well. According to the algorithm result obtained, the most important feature that had an effect on success was capital. It emerged that entrepreneurs should have a certain capital in order to survive while establishing their businesses. By considering this result, even if entrepreneurs are supported by the government, it is not certain that they will be successful. This result is crucial for the entrepreneurs who dream of establishing a business without KOSGEB support and capital. In terms of correcting this wrong attitude and not wasting government support, it has been concluded that the equity of entrepreneurs needs to be sufficient, only government support is not enough for success.

Of the 183 entrepreneurs in the dataset, 96 were successful and 85 were unsuccessful. Two entrepreneurs' information could not be reached. Despite the support given to the entrepreneurs, some were unsuccessful since the support decision was not made within a scientific framework. With this study, it was aimed to find the critical success features and to prevent resource waste by using the historical data. Alos, in this study it was targeted to increase the accuracy of the estimates and to ground REFUSED and ACCEPTED decisions about the projects on scientific base. As a result, it is assumed that the objectives of the study, mentioned at the beginning can be achieved if support is given to the entrepreneurs who have sufficient capital and partnership, the features that affect the entrepreneurs according to the algorithm result.

This study was prepared using only the data of the entrepreneurs in Gaziantep province. This local population is a limitation for this thesis. For this reason, generalizing the results of this study to all entrepreneurs may not be right.

In future studies, it can be compared that data mining will be used in estimating KOSGEB support with data belonging to entrepreneurs in different years. The success of the government support in Turkey in previous studies was not analyzed. This study can be a pioneer and a model of state support can be established to be used for the prediction of success for further studies.

REFERENCES

1. Beaver, W. H. (1966). Financial Ratios As Predictors of Failure. *Journal of Accounting Research*. **4**, p. 71-111.

2. Altman, E. I. (1968). Financial Ratios, Discriminant Analysis And The Prediction Of Corporate Bankruptcy. *The Journal of Finance*. **23**, p. 589-609.

3. Izan, H. (1984). Corporate distress in Australia. *Journal of Banking & Finance*. **8**, p. 303-320.

4. Donald A.Duchesneau, W. B. (1990). A profile of new venture success and failure in an emerging industry. *Journal of Business Venturing*. **5**, p. 297-312.

5. Arnold C. Cooper, J. G. (1991). A resource-based Prediction of New Venture Survival and Growth. *Academy of Management Annual Meeting Proceedings*. p. 68-72.

6. Lussier, R. N. (1995). A nonfinancial business success versus failure prediction mo. *Journal of Small Business Management.* **33**, p. 8.

7. Lussier, R. N. (1996). A business success versus failure prediction model for service industries. *Journal of Business and Entrepreneurship*. **8**, p. 23.

8. Lussier, R. N. (1996). A startup business success versus failure prediction model for the retail industry. *The Mid-Atlantic Journal of Business.* **32**, p. 79.

9. Robert N. Lussier, J. C. (1996). A business success versus failure prediction model for entrepreneurs with 0-10 employees. *Journal of Small Business Strategy*. **7**, p. 21-36.

10. Robert N. Lussier, S. P. (2001). A crossnational prediction model for business success. *Journal of small business management*. **39**, p. 228-239.

11. Feng Yu Lin, S. M. (2001). A data mining approach to the prediction of corporate failure. *Knowledge-based systems.* **14**, p. 189-195.

12. Yadollah Mehralizadeh, S. H. (2005). A study of factors related to successful and failure of entrepreneurs of small industrial business with emphasis on their level of education and training. *European Conference on Educational Research.*

13. Nguyen, H. G. (2005). Using neutral network in predicting corporate failure. *Journal of Social Sciences*. **1**, p. 199-202.

14. Defu Zhang, S. C. (2008). A decision tree scoring model based on genetic algorithm and k-means algorithm. *Third 2008 International Conference on Convergence and Hybrid Information Technology*. p.1043-1047.

15. Gepp, A. K. (2010). Business failure prediction using decision trees. *Journal of forecasting*. **29**, p. 536-555.

16. Hui Li, J. S. (2010). Predicting business failure using classification and regression tree: An empirical comparison with popular classical statistical methods and top classification mining methods. *Expert Systems with Applications.* **37**, p. 5895-5904.

17. Bee Wah Yap, S. H. (2011). Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert Systems with Applications.* **38**, p. 13274-13283.

18. Robert N. Lussier, C. E. (2010). A Three-Country Comparison of the Business Success versus Failure Prediction Model. *Journal of Small Business Management.* **48**, p. 360-377.

19. Shaike, M. (2014). A business success versus failure prediction model for small businesses in Israel. *Business and Economic Research*. **4**, p. 63.

20. Nurcan Öcal, M. K. (2015). Predicting Financial Failure Using Decision Tree Algorithms: An Empirical Test on the Manufacturing Industry at Borsa Istanbul. *International Journal of Economics and Finance.* **7**, p. 189.

21. Shabir, H. (2016). Why businesses succeed or fail: a study on small businesses in Pakistan. *Journal of Entrepreneurship in Emerging Economies*. **8**, p. 82-100.

22. Wang, X. Z. (1999). Data Mining and Knowledge Discovery, in Data Mining and Knowledge Discovery for Process Monitoring and Control. *X.Z. Wang, Editor Springer London.* p. 13-28.

23. Usama Fayyad, G. P.-S. (1996). Piatetsky-Shapiro, and P. Smyth, From data mining to knowledge discovery in databases. *AI magazine*. **17**, p. 37.

24. Ian H. W. 1999. Data Mining: Practical machine learning tools and techniques. 3rd edition. Waltham, USA: Elsevier.

25. Thakare, M. S. (2010). Data mining system and applications: A review. International Journal of Distributed and Parallel systems (IJDPS). **1**, p. 32-44.

26. Larose D. T. 2014. Discovering knowledge in data: an introduction to data mining. 2nd edition. John Wiley & Sons Inc.

27. Han J. J. 2011. Data mining: concepts and techniques. 3rd edition. Waltham, USA: Elsevier.

28. Wagner M, Mohammed JZ. 2014. Data mining and analysis: fundamental concepts and algorithms. Cambridge University Press.

29. Aggarwal C. C. 2015. Data mining: the textbook. Springer.

30. Carlin B. P. 2008. Bayesian methods for data analysis. CRC Press.

31. Stigler, S. M. (1981). Gauss and the invention of least squares. *The Annals of Statistics*. p. 465-474.

32. Turing, A. M. (1937). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London mathematical society*. **2**, p. 230-265.

33. McCulloch, W. S. (1943). A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics. **5**, p. 115-133.

34. Azevedo, A., Santos M.F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *International Journal of Intelligence Science*. p.182-185.

35. Chapman, P. (1999). The CRISP-DM user guide. *4th CRISP-DM SIG Workshop in Brussels.*

36. Mariscal, G. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*. **25**, p. 137-166.

37. Olson DL, Delen D. 2008. Advanced data mining techniques. Heidelberg: Springer Science & Business Media.

38. Berkhin, P., *A survey of clustering data mining techniques*, in *Grouping multidimensional data*. 2006, Springer. p. 25-71.

39. Quinlan J.R. (2014). C4. 5: programs for machine learning. San Mateo, California: Morgan Kaufman Publishers.

40. Salzberg, S. L. (1994). C4. 5: Programs for machine learning by j. ross quinlan. *Morgan kaufmann publishers.* **16**, p. 235-240.

41. Brijain, R. P., Kushik, K. R. (2014). A survey on decision tree algorithm for classification. *International Journal of Engineering Development and Research.* **2**, p. 1-5.

42. Shannon, C. E. (1948). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*. **5**, p. 3-55.

43. Quinlan, J. R. (1987). Simplifying decision trees. *International Journal of Human-Computer Studies.* **51**, p. 497-510.

44. John, G. H., Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. *The Eleventh conference on Uncertainty in artificial intelligence*.

45. Liaw, A. (2002). Classification and regression by randomForest. *R news*. **2**, p. 18-22.

46. Cortes, C. (1995). Support-vector networks. *Machine learning*. **20**, p. 273-297.

47. Zheng, J.G., Jiao, L.C. (2001). Using deviation detection for data mining. *Springer*

48. Zhang, C., Zhang, Z. (2002). Association rule mining: models and algorithms. *Springer*.

49. Sing, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics.* **21**, p. 3940-3941.

50. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*. **27**, p. 861-874.

51. Tan PN. 2006. Introduction to data mining. Pearson Education India.

52. Paidi, A. N. (2012). Data mining: Future trends and applications. *International Journal of Modern Engineering Research (IJMER)*. **2**, p. 4657-4663.

53. Hall, M. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*. **11**, p. 10-18.

54. Kirkby RE. 2006. Weka explorer user guide for version 3.5.8. University of Waikato.

55. KOSGEB. 2017. 2016 Yılı KOSGEB Faaliyet Raporu.

56. KOSGEB. 2015. 2016-2020 Stratejik Plan.

57. KOSGEB. 2018. Girişimcilik Destek Programı Uygulama Esasları.