T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ

The Graduate School of Natural and Applied Sciences
Industrial Engineering

DATA MINING TECHINIQUES AND A BANKING APPLICATION

M.Sc. Thesis

Ecehan ÇETİN

Assoc. Prof. Dr. Mehmet Mutlu YENİSEY

Istanbul, June'2011

ACKNOWLEDGEMENT

I would like to thank to all who gave me the possibility to complete this thesis. I am deeply indebted to my supervisor Associate Prof. Dr. Mehmet Mutlu Yenisey from İstanbul Technique University whose guidance, stimulating suggestions and encouragement helped me in all the time of research for and writing of this thesis. His moral support and continuous guidance enabled me to complete my work successfully. This thesis would not have been possible without the support of Associate Prof. Dr. Faik Tunç Bozbura.

I also would like to thank am as ever, especially indebted to my lovely friends, for their love and support throughout my life.

Finally, I would like to thank to my parents for their help and support during all my life.

İSTANBUL,2011 Ecehan ÇETİN

ABSTRACT

DATA MINING TECHINIQUES AND A BANKING APPLICATION

ÇETİN, Ecehan

INDUSTRIAL ENGINEERING

Supervisor: Assoc. Prof. Dr. Mehmet Mutlu YENİSEY

June 2011, 108 Pages

Customer data is one of the most valuable assets of any company and leading information to market success. It is very important to evaluate and reveal the valuable knowledge hidden in raw data. This is why data mining has become an inevitable and competitive tool especially in the banking and retail industries. The business way is executed in banking industry around the world has undergone a grand changes. Data mining tools that are contained several data mining techniques are being used by leading banks for customer segmentation, behavior, profitability, credit scoring, tracing payment balances, marketing new products and detecting fraudulent transactions..etc Therefore, data mining techniques conduce to optimize business decisions and increase the value of each customer information and improve customer satisfaction.

The purpose of this study is to describe fundamental concepts of data mining that are existed in a standard data mining tool and explain why these major processes are extremely important for corporations in order to deal with large data sets effectively. Also it is planned to introduce well known techniques commonly used in data mining processes that have proven effective analyzing huge data sets that have ambiguous descriptions and conditions by producing profitable business decisions. Especially, the data mining methods commonly used in banking and retail industries are selected and explained in detail by performing literature review. In the last section, one of the methods among others is chose and is applied to customer data. Therefore, all major steps of the data mining are practiced in that data and tried to find proper business results.

This research exposes clear understanding of data mining techniques and the essential steps of them. It is also present how these methods can be used by companies for which purposes and which goals by performing which processes.

Keywords; Data mining, Data mining techniques, Banking application

ÖZET

VERİ MADENCİLİĞİ TEKNİKLERİ VE BİR BANKACILIK UYGULAMASI

ÇETİN, Ecehan

ENDÜSTRİ MÜHENDİSLİĞİ

Tez Danışmanı: Doç. Dr. Mehmet Mutlu YENİSEY

Haziran 2011, 108 Sayfa

Bir firma için müşteriye ait bilgiler, bulunduğu sektördeki başarısı açısından en önemli ve yönlendirici değerleri arasındadır. İşlenmemiş, ham veriler içerisindeki bilgilerin ortaya çıkarılması ve değerlendirilmesi çok önemlidir. Bu açıdan veri madenciliği, özellikle bankacılık ve perakende sektörü için vazgeçilemez bir araç haline gelmiştir. Dünya çapında bankacılık iş yapısı hızlı bir değişime uğramaktadır. Veri madenciliği tekniklerinin kullanıldığı yazılım çözümleri vasıtası ile bankalarda müşteri sınıflandırılması, müşteri eğilimlerinin belirlenmesi, karlılık hesaplamaları, kredi puanlama, ödemelerin izlenmesi, yeni ürün pazarlama faaliyetleri ve sahte işlem incelemeleri gibi çalışmalar yapılabilmektedir. Bu şekilde, veri madenciliği teknikleri ile iş kararlarının optimize edilmesi, müşteri değeri ve müşteri memnuniyetinin artırılması çalışmalarına destek vermektedir.

Bu çalışmanın amacı, standart veri madenciliği yazılımlarında kullanılan veri madenciliği teknikleri ile ilgili temel kavramları tanımlamak ve geniş veri gruplarının etkili şekilde değerlendirilmesini sağlayan bu temel proseslerin önemini açıklamaktır. Ayrıca en çok bilinen, kullanılan ve geniş veri gruplarını başarılı şekilde analiz edip karlı iş kararları verilmesini sağlayan veri madenciliği teknikleri hakkında detaylı bilgi verilmesi planlanmıştır. Özellikle, bankacılık ve perakende sektöründe yaygın olarak kullanılan veri madenciliği yöntemlerine ait literatür taraması yapılarak detaylı bilgi verilmektedir. Son bölümde ise müşteri datası üzerinde veri madenciliği uygulaması yapılmıştır. Böylece, veri madenciliğine ait tüm temel adımlar gerçek verilere uygulanarak, sonuçları değerlendirilmiştir.

Bu araştırma ile veri madenciliği teknikleri ve adımlarına ait net bir anlayış ortaya konulmuştur. Aynı zamanda bu yöntemlerin firmalarda ne şekilde, hangi alanlarda, hangi amaç ve hedefler için uygulandığı anlatılmıştır.

Anahtar Kelimeler: Veri madenciliği, Veri madenciliği teknikleri, Bankacılık uygulaması

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

Artificial intelligent	:	AI
Customer relationship management	:	CRM
Knowledge discovery in databases	:	KDD
Online analytical processing	:	OLAP
Online transaction processing	:	OLTP
Self-organizing map	:	SOM

1. INTRODUCTION

In recent years, the capacity of collecting, generating, transferring and storing data have been increased enormously for especially in banking and retail industries There is always some information hidden in these huge amount of data called database will be vitally important for the point of business view of the company. Because of that, it is very critical issue to recognize how to evaluate this information in order to meet the profitability goals of company and effectively compete in competitive global markets.

The amount of data collected and stored in databases by enterprises has grown rapidly nowadays. Existing statistical data analysis techniques are found inadequate to cope with the large volume of these databases. That significant growth has caused the need for new data analysis techniques and tools that are containing more systematic processes and are supported by software solutions in order to detect hidden information in the databases. Consequently, the research field of data mining has arisen.

Data mining has been defined as a statistical process of analyzing data stored in a data warehouse. A data warehouse is an extension data repository consisting of information from all facets of an organization that is maintained to support decision making. Through data mining technology, large databases can be explored to find relationships and trends previously unknown, to provide support for complex decisions (Anderson & Jolly & Fairhust, 2007). As a basic definition, data mining is the list of processes that help for extracting previously unknown information from a large databases and produce significant results by the support of software program

In the meantime, data mining processes, methods and the area of their applications have been tremendously improved in both academic and business world. Especially existing ones have been very well researched and a serious of algorithm has been developed. Also, software applications have been developed according to the newly invented data mining methods by implementing new algorithms. A variety of analytic software models have been used in data mining as well as other means of analysis.

Today, data mining is being applied into several industries including banking, finance, retail, insurance, telecommunications..etc Data mining processes and the methods are also used for database marketing, sales forecasting, call behavior analysis, detecting fraud cases, churning management in telecommunications, quality control in manufacturing sector, improve food and drug safety, cancer diagnosis and biomarker discovery, forecasting of demand for utilities such as usage of energy and water.

Banking is the most important sector where data mining is effectively used and is being produced important results in business world. Especially, customer relationship management (CRM) that contains a list of processes and software support to produce business strategy to build long term profitable customer relationship has became vital business approach for the banks. CRM can also be defined as managing customer's relationships with the company by using strategic information, processes, technology and people in the area of marketing, sales, services and support. CRM is an extensive process of acquiring and retaining customers with the support of business intelligence in order to increase the values of customers to the enterprise. It can be observed CRM practices applied in banking sector in the area of customer demographic and financial information, call center, ATM facilities, credit & debit card transactions, internet banking, swift network, connection availability of branches and their financial facilities.

This research aims to provide detail information for concept, evaluation and main processes of data mining. Also it presents literature review for well known data mining techniques and comprehensive definition for each of the method. After that, one of the appropriate methods is selected in order to be applied into a sample of customer database taken from a bank.

The thesis is very well organized for the reader whose aim is to gain some information for the scope, evaluation, processes and techniques of data mining and also give an detail information for "how to apply one of the data mining techniques" into a sample of data in order to produce a significant result. In the second section, the literature review is provided. In the third and fourth sections, data mining tasks, processes will be explained in detail as well as expressing general knowledge of it. In the fifth section, the detail definition of data mining techniques is provided as well as the following methods are examined deeply in order to give clear view to the readers;

- Decision Trees
- Neural Networks
- Genetic Algorithm
- K-Nearest Neighbor Algorithm
- K-Means & Hierarchical Clustering
- Khonen Networks
- Naïve Bayesian Classifier
- Bayesian Network

In the sixth section, clustering methods (K-Means and Decision Tree) will be applied into a sample of customer data provided from a foreign bank located in İstanbul.

Since the last section that consists of conclusion, effective usage of data mining elements and techniques is provided in order to exist in competitive global market.

2. LITERATURE REVIEW

Data mining is a chain of processes aimed to extract hidden information such as data attributes, trends or patterns from large databases by analyzing data with defined some specific analysis techniques and summarize it into useful information. The extraction process is successfully produced a result by finding correlations or patterns among dozens of fields of large databases which are stored into as data warehouses.

Data mining has an organic link to Knowledge discovery in databases (KDD). KDD is an systematically executed and explorative analysis and modeling of large database. KDD is also defined as an iterative discovery and learning process that broadens the collection of data mining techniques into a knowledge management framework. Please find the typical knowledge management process in Figure 2.1 in the following page.

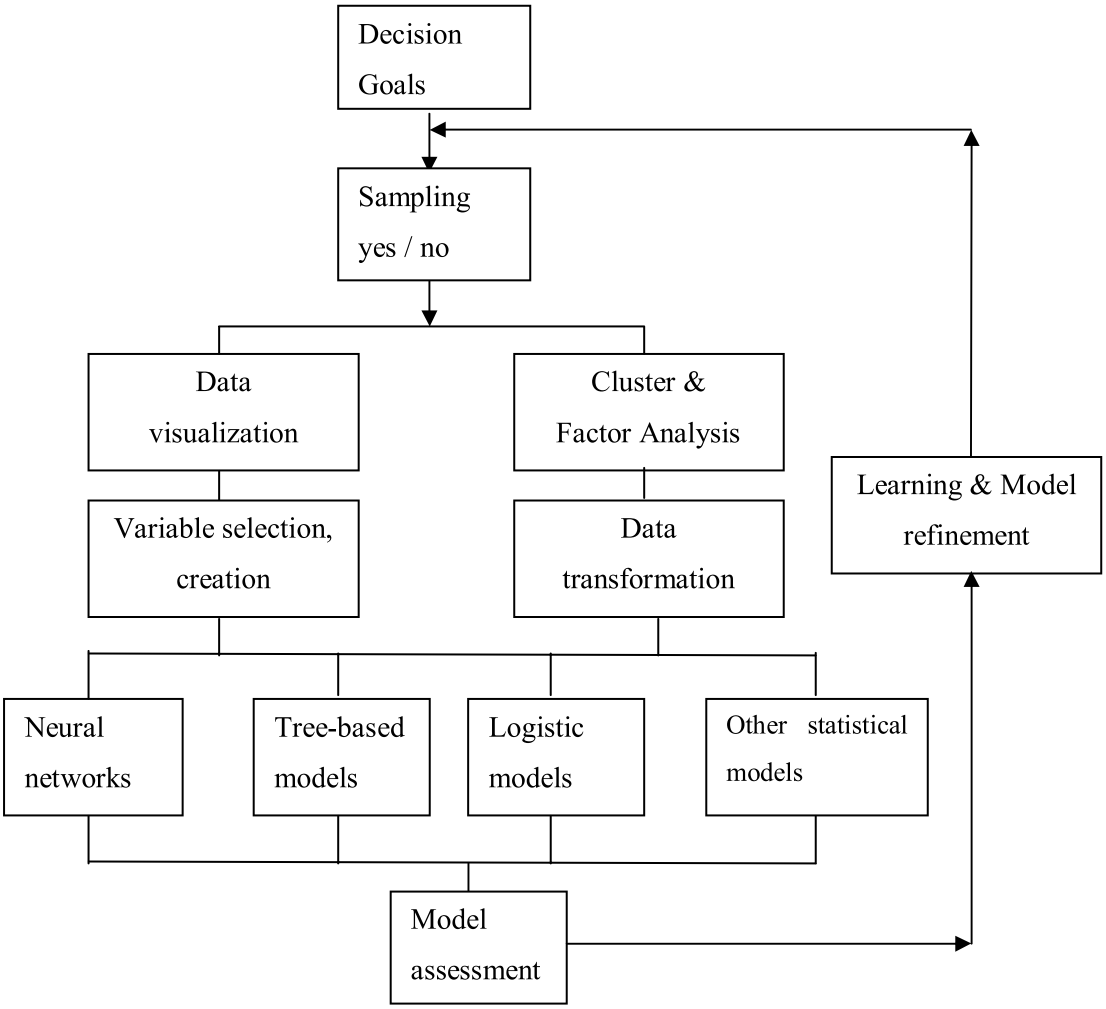


Figure 2.1 Knowledge management process
Source: Shaw & Subramanian & Tan Gek & Welge, 2001

Although data mining techniques are commonly applied to the complete database, it can be preferred to apply a statistically representative sample of data as it is illustrated in Figure 2.1 above. After making decision on using sample or complete database, the next step will be proceeding which is explored the data using the tasks such as visualization. And also appropriate data mining techniques are selected and applied over the data. The outcome of that data mining practice is evaluated to define the useful and effective resulting pattern to produce a solution for the aim of associated data mining project.

The quality and quantity of available data are directly effect on the result of data mining and knowledge discovery projects. Because of that, data warehouse and its associated activities, such as data type, data storage techniques and the quality of the data, will be very important factors on the success of whole project and its steps such as data preprocessing, data mining representation of generated knowledge and assessment of generated model.

Data warehouse is defined as a repository of data collected in different locations (relational databases) and stored using a unified schema, is used. Data warehouses are usually created by applying a set of processing steps to data coming from multiple databases. The steps usually include data cleaning, data transformation, data integration, data loading and periodical data update. Please find the typical architecture of a data warehouse in Figure 2.2 below (Miamon & Rokach, 2005).

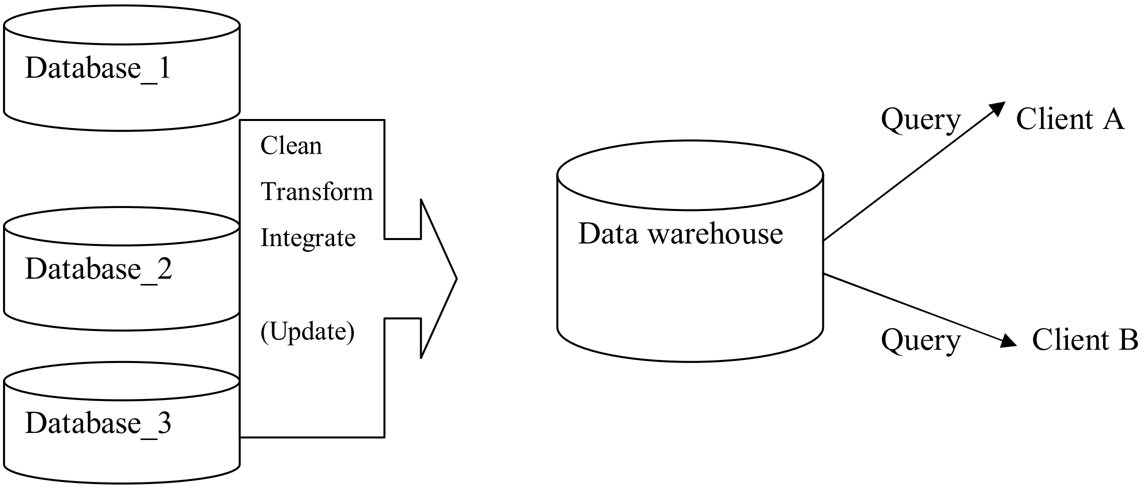


Figure 2.2 Typical architecture of a data warehouse system
Source: Miamon & Rokach, 2005

Data mining uses modern statistics, intelligent information systems, machine learning, patterns recognition, decision theory, data engineering and database management. Data mining process is also supported by a powerful software tool that is disclosed hidden and complicated relationships in large data sets. Therefore data mining methods becomes a part of Information Technology (IT) software packages. It is illustrated in

Figure 2.3 below. Concerning all these features above, it can be expressed that data mining has a great impact on business and financial decision making such as risk management of all financial transactions including fraud detection and market risk, asset allocation and trading, behavioral finance and customer relationship management (CRM).

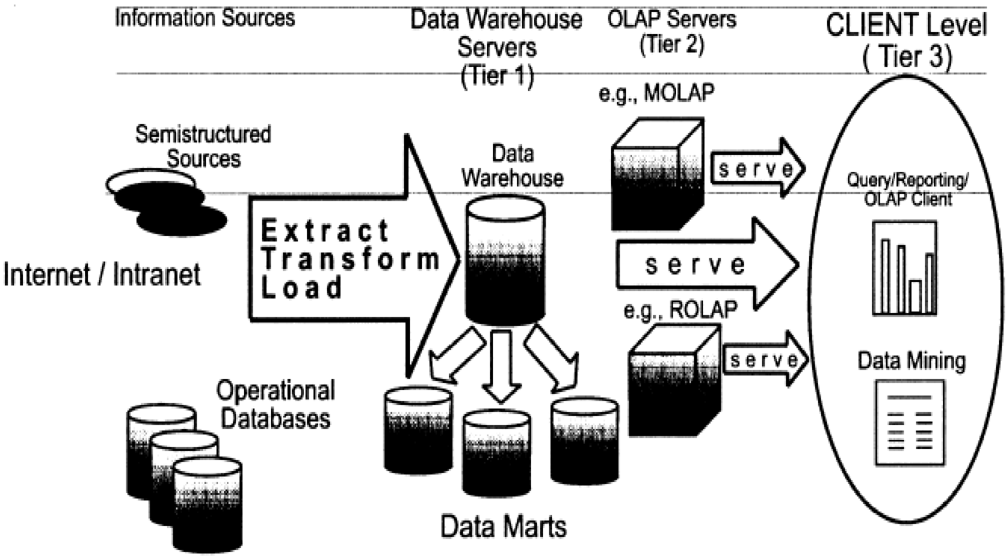


Figure 2.3 The IT decision support tiers
Source: Miamon & Rokach, 2005

Figure 2.3 illustrates the three tiers of the decision support aspect of IT. Starting from the data sources (such as operational databases, semi- and non-structured data and reports, Internet sites etc.), the first tier is the data warehouse, followed by OLAP (On Line Analytical Processing) servers and concluding with analysis tools, where Data Mining tools are the most advanced (Miamon & Rokach, 2005).

Data mining techniques and tools have a great impact on implementing customer relationship management in retail or banking industries. CRM is defined as a comprehensive strategy and process of acquiring, retaining and partnering with selective customers to create superior value for the company and the customer (Ngai & Xiu & Chau, 2009).at al, 2009). It contains the integration of marketing, sales, customer service and supply chain functions of the enterprise in order to succeed delivering customer value effectively and efficiently. That shows that CRM is very important instrument for acquiring and retaining customers with the help of business intelligence via data mining. Data mining helps to be analyzed and understood the customer behaviors within competitive CRM strategy of the enterprises and generates a model from the related data set. A graphical classification framework on data mining techniques in CRM is proposed and shown in Figure 2.4 below (Ngai & Xiu & Chau, 2009).

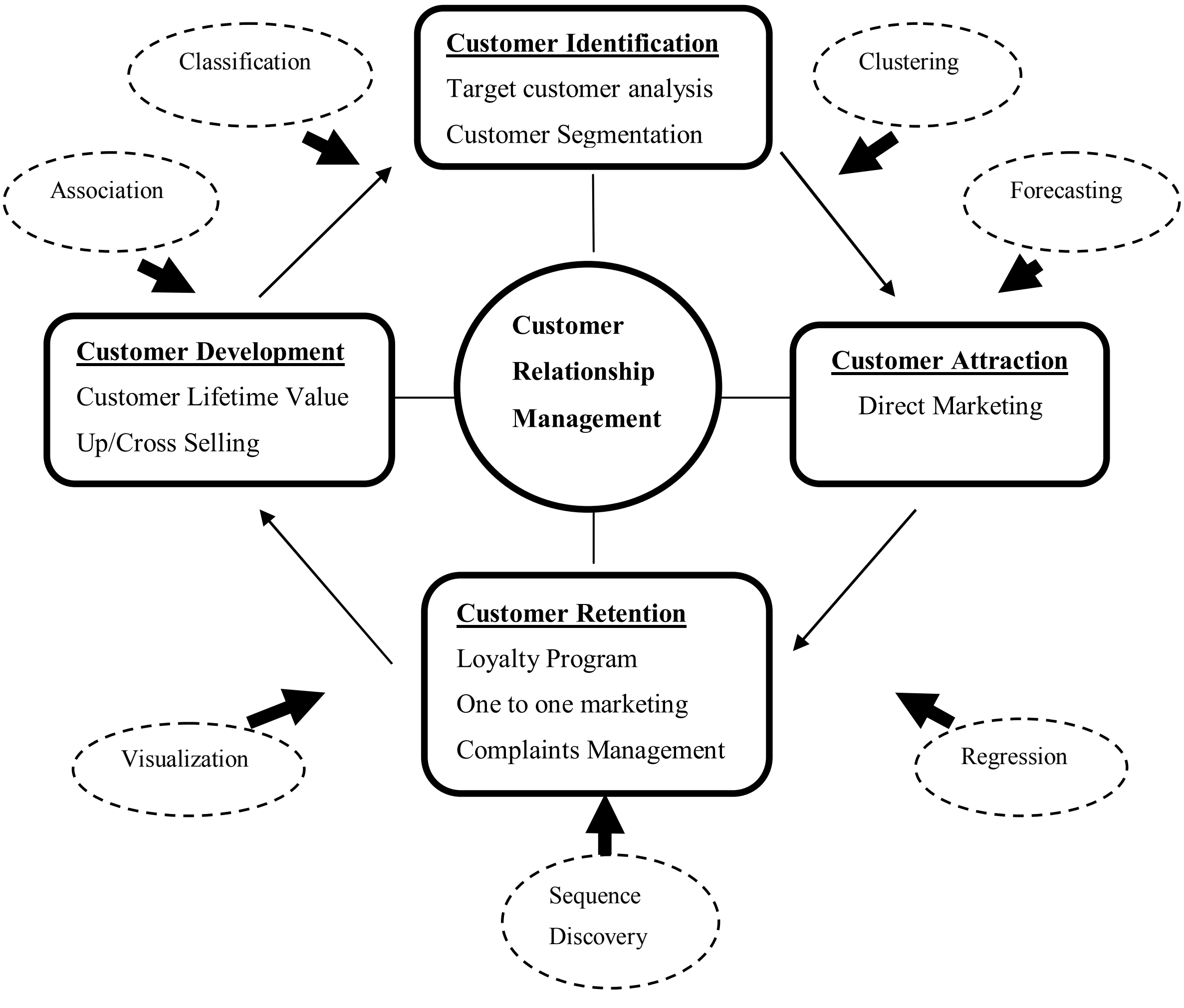


Figure 2.4 Classification framework for data mining technique in CRM
Source: Ngai & Xiu & Chau, 2009

3. DATA MINING

According to the Gartner group data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amount of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques (www.gartner.com, 2011). There are also other definitions;

Data mining is a new discipline lying at the interface of statistics, database technology, pattern recognition, machine learning and other areas. It is concerned with the secondary analysis of large databases in order to find unsuspected relationships, which are the interest or value to database owners (Hand & Mannila & Smyth, 2001).

Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and summarize the data in novel ways that are both understandable and useful to the owner (Hand & Mannila & Smyth, 2001).

Data mining is predicted to be "one of the most revolutionary developments of the next decade" according to the online technology magazine ZDNET News (Kondrad, 2001). In fact the MIT Technology reviews choose data mining as one of the ten emerging technologies that will change the world (MIT Technology Review, 2001).

According to the 1999 Information Week National Salary Survey reports; "Data mining skills are in high demand this year as organizations increasingly put data repositories online. Effectively analyzing information from customers, patterns and suppliers has become important to more companies. 'Many companies have implemented a data warehouse strategy and now are starting to look at what they can do with all that data' says Dudley Brown, managing partner of BridgeGate LLC" (Metayashuck,1999).

From looking at all the definitions and comments above, data mining has been very significant, widespread and useful discipline relies heavily on information technology for not only academic but also business world.

3.1 CHARACTERISTICS OF DATA MINING

All the definitions of data mining made by either academic authorities or partner of business world contain the same major characteristics;

Observational data; As it was mentioned in the definition of Mr. Hand (Hand & Mannila & Smyth, 2001)., main item of data mining cannot be a group of ordinary data which is collected without any reason that is called experimental data but it must a group of data that have been already collected for some purposes, that is named as observational data. That means, the types of data collection has not any important role within the objectives of data mining. That is a distinguishing feature from the statistical methods in which data are mainly collected by defined strategies to answer specific questions.

Large Data Sets; Data volume is also typical features of data mining and it should be large group of data in order to apply data mining principals. If a small data sets are exist, it should be discussed which classical exploratory data analysis techniques must be used. Just because, when the large group date are involved, new relational problem is arise and it makes impossible to solve one of the ordinary statistical methods.

Information Technology; Most of the phases of data mining has been heavily supported by software programs is called data mining tools. Data mining tools need to contain some important features in order to produce objective result. Such as, versatile, scalable, capable of accurately, predicting responses between actions and results, capable of automatic implementation. It can be applied wide variety of models if the tool is versatile. Scalable is important feature if different volumes of data sets are applied, such as large or medium. Automation is useful feature especially if some of the analytical functions are automated but if only human set up is regulated correctly. Also data transformation is often necessary. Apart from all these automation, there will be some human activities with the data mining processes, such as selecting proper group of data that is going to the subject of all research, analyst judgment for especially critical issues in implementation of data mining. Off course, the most important part is the understanding of statistical concept fundamentally.

Meaningful/Understandable Results*;* It the vital part of data mining to produce relational, structural, measurable and statistical objective results that have not been generated before. But there are also some data mining techniques considering the prior knowledge of users.

There is also an important detail related with the meaning of data mining, misnomer in the term of data mining. It is also emphasized that the term of mining is used the valuable material is being extracted but in fact the meaning of mining in the term "data mining" is vice a versa. Data is not mined; there is a large amount of data already stored and waiting to be interpreted. In other words, gold mining is extracting gold from sand, coal mining is extracting coal from ground but data mining is extracting knowledge from data.

Data mining is in the position of enlarged concept of knowledge discovery in databases KDD is also invented in the artificial intelligent (AI) search field. It is generally used as a synonym for the term "data mining". KDD is more explicit and illuminating than the term of data mining in spite of the low popularity of it. Please find the relations of data mining with other disciplines in Figure 3.1 below.

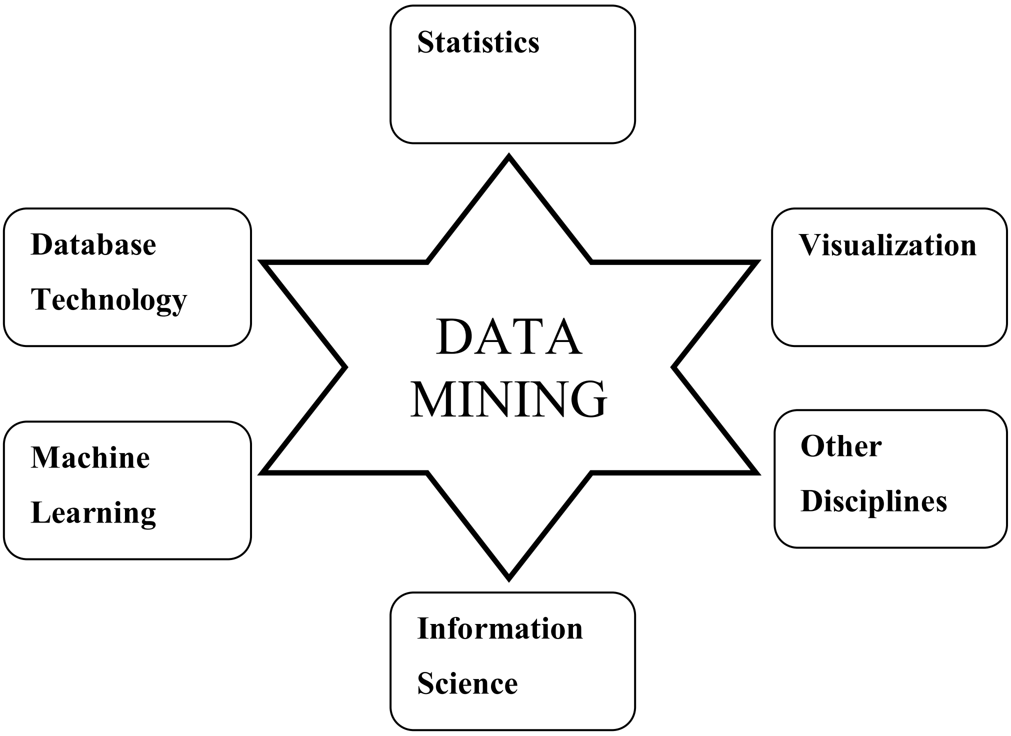


Figure 3.1 Data mining: Confluence of multiple disciplines
Source: Han & Kamber, 2001

3.2 EVALUATION OF DATA MINING

Data mining can be viewed as a result of the natural evaluation of information technology. An evolutionary has seen witnessed in the database industry in the development of the following functionalities; data collection, data management that is included data storage, retrieval, and database transaction processing, data analysis and understanding that is involved data warehousing and data mining. For instance, the early development of data collection and database creation served as a prerequisite for later development of effective mechanisms for database systems offering query and transaction processing as common practice, data analysis and understanding has naturally become the next target (Han & Kamber, 2001).

The evaluation of data mining and emergent forces are summarized in Table 3.1 below

Table 3.1 Emergent forces & evaluation of data mining

Emergent Forces	**Evolutionary Step**	**Business Question**	**Enabling Technology**
60s New Products	Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks
70s Low cost manufacturing 80s Total Quality Management	Data Access (1980s)	"What were unit sales in New England last March?"	faster and cheaper computers with more storage, relational databases
90s Customer Relationship Management and one to one marketing	Data Warehousing and Decision Support	"What were unit sales in New England last March? Drill down to Boston"	faster and cheaper computers with more storage, Online analytical processing (OLAP), multidimensional databases, data warehouses
2000s Knowledge enabled relationship management and e-business	Data Mining	"What likely to happen to Boston unit sales next month? Why?"	faster and cheaper computers with more storage, advanced computer algorithms

Source: Tiawana, A., (2001) The Essential Guide to Knowledge Management : E-businesssand CRM Application, Prentice Hail PTR, Saddle River

Since the 1960s, information technology and database management techniques have been developed systematically from primitive data processing systems to more automated and powerful database systems. The research and developments starting from 1970s has been improved from early hierarchical and network database systems to the more developed relational database systems including data modeling tools and new data organization techniques. Additionally, more flexible data access through query languages, user interfaces, optimized query processing and transaction management systems have been developed. Therefore the system became more user – friendly. Also the improvements on online transaction processing (OLTP) has been continued especially for generating more efficient methods where a query is viewed as a read only transaction have been changed considerably to the valuation and became wide acceptance of relational technology as a main tool for efficient storage, retrieval and management of large data sets. Database technology since the mid-1980s has been improved on the way of adopting relational technology and by the expanding of extensively

3.3 APPLICATIONS OF DATA MINING

Data mining is a broad technology that can potentially benefit any functional areas within a business where there is a major need or opportunity for improved performance and where data is available for analysis that can impact the performance improvement.

Data Mining Applications in Marketing/Retailing; Tight profitability targets and highly competitive market conditions pushed retailers into embracing the data warehouse technologies earlier than the other industries. Retailers have noticed that improved decision support systems lead directly to successful management on inventory management and financial forecasting. Retailers who were applied the data warehousing techniques earlier than the others in market had a better opportunity to take advantages of data mining. There are various and huge amount of point of sales data have been stored in large retail chains and grocery stores. These conditions were the important factor for retailers to adopt data mining strategies into their enterprises.

Data Mining Applications in Banking/Finance; Data mining has widely been used in the banking and financial markets. Especially in the banking industry, data mining is extensively used to generate models and prediction for fraud issues and also evaluates risks, performs trend analysis, analyzes profitability as well as launch marketing campaigns. In the financial markets, neural networks have been generally used in stock-price forecasting, option and bond trading, portfolio management, commodity price prediction, mergers and acquisition as well as forecasting financial disasters

Data Mining Applications in Internet; The World Wide Web (known as WWW or Web) is growing enormous steps. The main reason for the success of web is hidden on its simplicity. Users can easily retrieve any issue any information that they require thru the hypertext interface. There is another important feature is its compatibility with the other existing protocols ftp,telnet…etc. Because of these uncompetitive features, web became most popular information and communication tool all around the world and it was caused to store enormous volume of data and the need for data warehouse. Therefore data mining is commonly started to use in web world and called as web mining.

Data Mining Applications in Telecommunications; In recent years, the telecommunication industry has experienced tremendous changes among the other industries. The volume of the customers and their transaction performed has increased enormously. Therefore data warehouse facilities and data mining techniques found an important place in the telecommunication industry.

Data Mining Applications in Healthcare; Data mining has extensively been used in medical industry. For example, neural network, decision trees and logistic regression are used to develop prediction models using large set of data for predicting cancer, smear diagnosis and survivability. And also other data mining techniques are used protein analysis for drug development.

Data Mining Applications in Manufacturing; During the recent years, profitability terms have been changed in manufacturing industry. Not only low prices but also high quality and on-time delivery are taking a place in the profitability strategies of companies. Although these features were advantages a decade ago, they are just required features in profitability strategies for companies to stay in business. Manufacturers face with more competition and demanding customers within this global environment. Because of that, data mining became one of the main technologies in that market conditions and strategies.

Table 3.2 shows examples of business applications in various sectors and industries that can most benefit from data mining (Musaoğlu, 2003).

Table 3.2 Examples of data mining business applications in various sectors

Sector / Industry	***Application***
Marketing / Retailing	√ Market basket analysis √ Finding market segments √ Identifying loyal customers √ Predicting what type customers will respond to mailing √ Finding customer purchase behavior patterns √ Finding associations among customer characteristics √ Determine items for cross selling / up-selling √ Detecting seasonal differences in sales patterns √ Product placement √ Forecasting sales / demand / revenue
Banking / Finance	√ Predicting customers that are likely to change their credit cards √ Identifying loyal customers √ Identifying fraudulent behavior √ Detecting patterns of fraudulent credit card usage √ Credit Scoring √ Risk assessment of credit √ Determine credit card spending by customer groups √ Segmentation of customers √ Analysis of customer profitability √ Managing portfolios √ Forecasting price changes in foreign currency markets √ Distribution channel analysis

3.2 Examples of Data Mining Business Applications in Various Sectors (continued)

Sector / Industry	***Application***
Telecommunications	√ Churn analysis
Internet	√ Text Mining √ Web marketing
Manufacturing	√ Inventory Control √ Equipment failure analysis √ Resource Management √ Process / quality control √ Capacity management
Insurance / Healthcare	√ Identifying fraudulent behavior √ Predicting which customers will buy new products √ Medical treatment analysis
Transportation	√ Loading pattern analysis √ Distribution channel analysis

Source: Musaoglu, C., (2003) Customer acquisition and retention modeling in consumer finance sector using data mining,Thesis Study, Bogazici Press, İstanbul

4. DATA MINING PROCES AND PHASES

Data mining has the multidisciplinary structure as well as the multiple tasks and procedures can be applied in different areas and industries. These structure causes setting various standards and methodologies of data mining for the different industry areas. The standard application methodology can be easily applied in the specific areas on time with the less cost and will be more reliable, manageable and faster. This will also caused developers to create new integrated data mining solutions to generate overall methodology. A methodology will also bring the features of more adaptable, understandable to data mining disciplines.

There is a model called Cross Industry Standard Process for Data Mining (CRISP – DM) is commonly used different industries. CRISP-DM model was developed in 1996 by analysts representing DaimlerChrysler AG, SPSS, NCR and OHRA. The project was partly sponsored by the European Commission under the ESPRIT program as a mining project that is independent of industry sector. CRISP provides a 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.

According to CRISP-DM, generally data mining projects has a life cycle containing six phases and demonstrated in the Figure 4.1 and the phases are adaptive and also the sequence of them is adaptive. In a clear definition, the next phase in the sequence often depends on the results associated with the antecedent phase. The most important dependencies between phases are indicated by arrows. The iterative nature of CRISP-DM is symbolized by the outer circle in Figure 4.1 and the solution to a specific business or research area leads to further questions that can be caused using the same general process as before. Experiences gained from past projects should always be brought to bear as input into new projects.

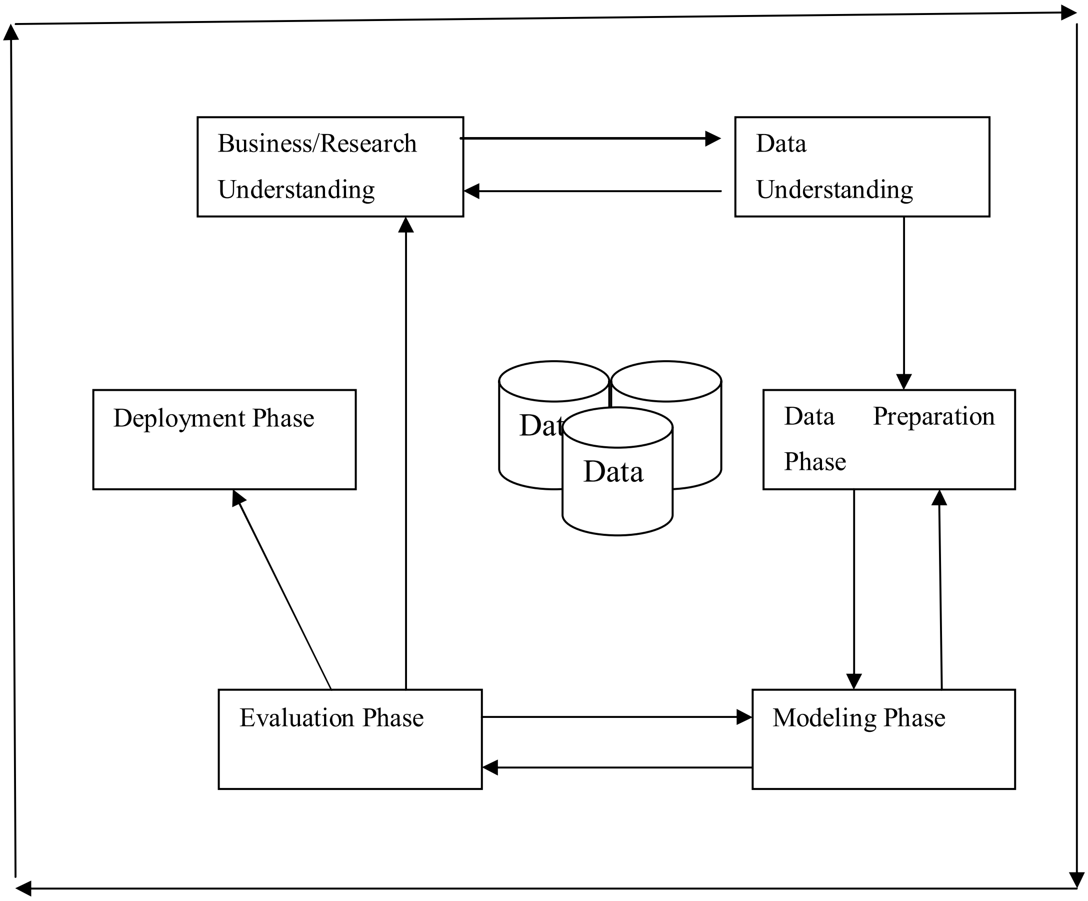


Figure 4.1 CRISP – DM process
Source: Larose, 2005

The CRISP – DM is a reference model for data mining processes and represents the general life cycle of data mining projects, typical phases, tasks and the their outputs.

As KDD is an important concept for data mining and decision support systems. Knowledge discovery concerns the whole knowledge extraction process containing the process of data storage and accessing. It is also included efficient and scalable algorithms to analyze huge data sets with interpreting and visualizing the results. Additionally, modeling is important to support the interaction between human and machines. Both DM and KDD are vital processes for decision support systems.

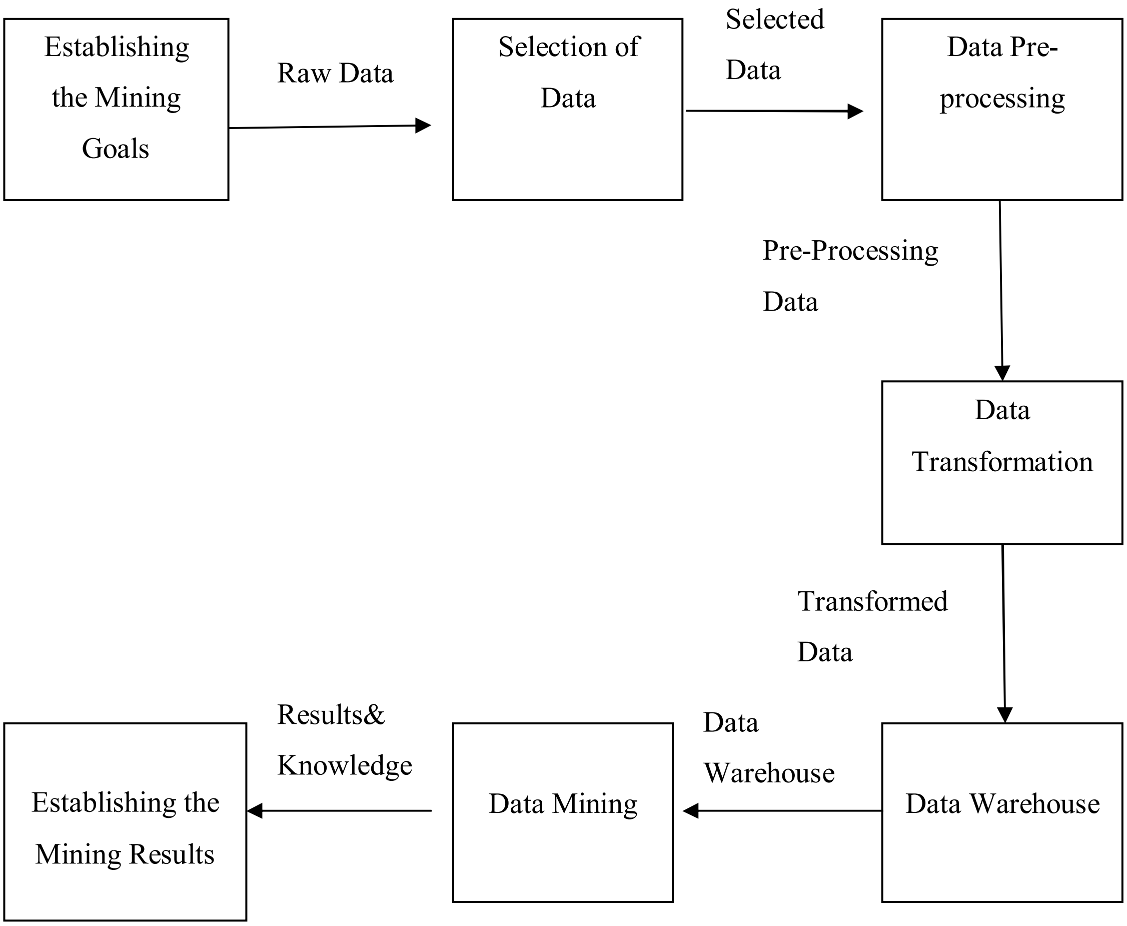


Figure 4.2 Data mining process for decision support
Source: Hui & Jha, 2000

4.1 BUSINESS UNDERSTANDING OF DATA MINING GOALS

It is a initial phase of data mining process and as well as the key element for the process to frame the scope of the study. It is focused on defining the project objectives and requirements in terms of business approach and transferring that information into the problem definition of data mining study and preliminary project plan to achieve the objectives.

The first and most important step in any data mining model project is to establish a clear goal and develop a process to achieve that goal. In the process of definition of the goal, you must first decide what you are trying to measure or predict. Targeting models generally fall into two categories, predictive and descriptive. Predictive models calculate some value that presents future activity. It can be continues value, like a purchase amount or balance, profitability of likelihood for an action. A descriptive model is just as it sounds. It creates rule that are used to group subjects into descriptive categories (Feeders & Daniels & Holsheimer, 2000).

Profile Analysis;

Acquire a new customer and protect the existing customer portfolio is an essential to stay competitive in marketplace recent days. Profile analysis is an excellent way to discover the value of the customer by measuring common characteristics within a population interest. Such as evaluating of customer demographic information, average age, gender (male, female), marital status…etc.

Segmentation;

Targeting models are focused on improving the efficiency of selected data mining topic such as marketing, risks..etc. But before developing the models, it is vital issue to understand value of the customer base. Profile analysis is competent technique to get to know the associated customers. General use of segmentation analysis is to segment customers according to profitability or market potential. For instance, customer are segmented into their purchasing behavior over their total purchasing behavior at all stores in retail industry. Through this analysis, the most potential customer range can be estimated. That method is also named as “share of wallet” analysis.

Response;

It should be the one of the first type of targeting model among the others that company focuses to develop. The goal of response model is to anticipate what kind of customer will be responsive to the product or service of the company. The model should be supported by the past behavior of similar population or some reasonable substitute. Response of the customers can be collected by the several of channels, such as e-mails, phone or internet. It is Important to manage response channels and discard duplicate transactions while collecting the results. Because of that, some rules must be defined in order to deal with multiple responses in model development phase.

Risk;

Risk models are commonly used in banking and insurance industry in order to predict or and define potential risks or loss when offering a product or services. The risks can be internal or external of the company or it can be defined or undefined in some circumstances. For instance, banks are mainly interested in financial risk of loans and decreasing the risks by collateral management. Also, fraud risk is another main area of risk models and main concern for the companies.

Activation;

Activation models are targeted to anticipate if the potentials will become full-fledged customer and commonly used in the financial service industry. For example, credit card can be delivered to customer successfully but the expected target is, the credit card should be also actively used by the customer. There are two different approaches for building an activation model. First approach is build a model and predicts responses as outcomes of the business. The second approach is using one-step modeling which is predicting the probability of activation without separating the different phases.

Cross-Sell and Up-Sell;

Cross-sell models are aimed to anticipate the probability or value of current customers who purchase a different products or services from the same company, the action called cross-sell. Up-sell models are focused on the probability or value of customers who purchase more of the same products or services. Both of the methods support to the decision of what kind of product / services will be produce in next marketing period in order to increase the amount and profitability of new customers as well as performing some marketing models to protect current customer existence and their profitability.

Attribution;

It is an important issue to avoid customer attrition or churn in many companies. They are both occurred the act of customer such as changing companies in order to take an advantage to of better based on finding new product or services with better financial or benefit. For instance, banks launch very attractive campaign with lower interest rate for credit card customer in order to lure credit card customer of other banks. Also telecommunication companies create new marketing strategies and tactics to lure existing customers away from their competitors.

These activities cause to be generated new modeling types into data mining discipline. One of the models is anticipates the possibility of reducing a product or service after a newly invented product or service is being launched in the market. Attribution is characterized a decrease in the usage of a product or service. Churn is also characterized as the closing of one product/service by opening of another product/service with the same features in order to reduce cost to consumer.

Net Present Value;

A net present value as a short name NPV model is aimed to anticipate the overall profitability of a product that has a predetermined timing. NPV is estimated based on predefined number of years of product life and discounted to the value of today. Even though there are some certain methods for the estimation of net present value, these methods are modified according to the product type or industries.

Life Time Value;

A life time value as a short name LTV model is aimed to anticipate the overall profitability of a customer for predetermined timing. It has similar terms for the estimation comparing to NPV model. It is also estimated over a predefined number of years of customer and discounted to the value of today. Even though there are some certain methods for the estimation of life time value, these methods are modified according to the product type or industries.

4.2 DATA UNDERSTANDING

There is a strong and complementary relation between business understanding and data understanding phases. Data understanding phase begin with collecting a group of data and continues with some activities with the following tasks in order to understand the data in details such as defining data quality problems.

After setting business objectives and generating project plan, data understanding is arises data requirements including initial data collection, data description, data exploration and verification of data quality. Data exploration is a kind of summary statistic.

Since data mining has task oriented approach, different business areas or aims need different group of data. The initial and important stage of data mining process is to select the related data among many databases in order to describe the business task which is the aim of project correctly. There are some main issues to be taken into consideration within the phase of data selection as follows,

i. **Clear definition of problem**; for instance identifying behavior pattern for credit card customer in black list with bad depth.

ii. **Defining the relevant group of data for problem definition**, this facility is also named as selecting data for modeling; for instance credit card transactions and demographic information of the customers are relevant for the project of credit card with bad depth. Data for modeling can be provided from different number of sources that are categorized as internal and external. Internal source is refer to the data is being provided from the activity of company. Such as, customer database, transaction database, history database, MIS database for special reporting. External source is refer to the data is being provided from out of company such as issued selling list among the other competitive companies…etc.

   Demographic data such as age, income, education, social-graphic data such as hobbies, club membership and transactional data such as credit card transactions will be included as types of data sources for business applications. These data can be classified as quantitative and qualitative data. Quantitative data is measurable containing numeric values. Qualitative data is also called categorical data contains both nominal and ordinal data.

iii. **Selecting associated variables for the relevant data** that must be independent of each other. Independent variables selection is decisive step for data mining algorithm to quickly detect the correct knowledge pattern.

After selecting relevant data according to the data mining business objectives and data mining method, data preparation phase should be followed.

4.3 DATA PREPARATION

The purpose of data preparation phase is to perform all activities to generate the final data set which will be processed into data mining tool, from the raw data. The tasks of data preparation are generally performed in multiple times and include tables where the transactions are saved, transactions, attribute of transactions, data cleaning a, construction of new attributes and transforming of data for data mining tool.

Raw Data / Sampling;

The initial task of that phase is sampling which is supported by the information technology. Because of that, the time spends for sampling has been improved and reduced. Since the information technology speed up the sampling process, this reduces the whole time for data preparation process.

Maintain Data Quality;

Data is infrequently hundred percents clean. So that the purpose of that task is clean the selected data to generate data set in better quality. Even some selected data may have collected from completely different sources with different formats.

i. ***Redundant Data;*** It is occurred when the duplicated transactions or impossible accuracies are existed in the data set. For instance, one customer purchased the same shirt as 10000 times in the same day. These 10000 transactions are suspicious to be redundant.

ii. ***Incorrect or Inconsistent;*** It is occurred when the invalid data or consistent accuracies are existed in the selected data. For instance, observing absurd or null values are listed in the customer data as customer name or surname.

iii. ***Typos;*** There can be some typing mistakes especially in the text type of data. Even though some of the databases are case sensitive, capital letters cause problems. For instance, Istanbul, Itanbul, Isstanbul..etc

iv. ***Stale Data;*** The data that have not been maintained since it is entered. Birth date, addresses are typical examples of stale data. But data staleness is an important factor for the other accurate data such as customer behavior..etc

v. ***Variance In Defining Terms;*** If the data is collected from different sources, there may be variances in the definition of data fields. For instance, suppose data collected from two different plants producing the same products.

Outliers;

Outlier analysis is generally applied into continues. It is determined if a value is an outlier or data error is an art as well as a science. With the support of that analysis, all aspects of the selected data will defined and weakness or strength side of the data will be determined. Outliers can be also systematically detected by various statistical analyses.

Missing Values;

After collecting and combining the data, there will be always some missing values are occurred. These missing values can be ignored or evaluated as nuisance by some data mining tools. It is important to ensure missing values by prediction and they can be substituted by some methods explained below;

i. ***Single Value Substitution;*** Single value substitution method is the most easiest and simplest method among the others and there are three methods to be applied called ‘mean’,’median’,’mode’

ii. ***Class Mean Substitution;*** The mean values among the subgroup of other variables combination are used in class mean substitution method.

iii. ***Regression Substitution;*** It is similar to class mean substitution method among the subgroup of other variables combination are used in this method.

Selecting and Transforming Variables;

There can be some additional processes are needed to be performed before mining the data. For instance, in order to predict customer behavior, it may be required to create new variables that have to be derived from the data of transaction table. Also, for transaction data of existing customer, RFM (Recency, Frequency, Monetary) Recency is refer to be some measure of time since the last transaction is performed. Frequency refers to be the numbers of transaction are performed in specified period. Monetary refers to be the total transaction in the specified period and also an average per transaction. As a result, there need to be generated some additional variables in order to create more meaningful results and useful relationship among the variables by using data mining tool. Please find the conceptual framework of data quality in Figure 4.3 below.

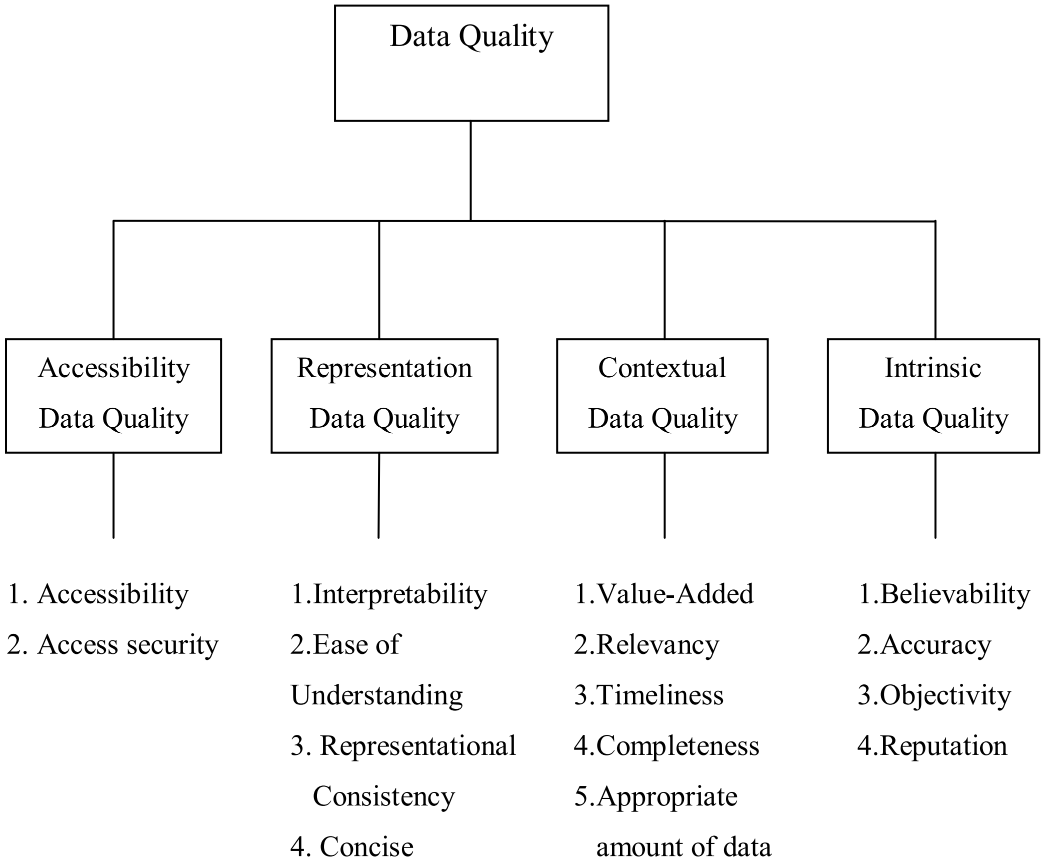


Figure 4.3 A conceptual framework of data quality
Source: Wang & Strong & Guarascio, 1994

i. ***Accuracy;*** The recorded value is consistent with the actual value. The accuracy dimension is the most straightforward and is the difference between the correct value and the one actually used.

ii. ***Timeliness;*** The recorded value is not out-of-date. Any data item will become out-of-date as time passes.

iii. ***Completeness;*** All the values of certain variable are recorded. Completeness can be handled in a satisfactory manner. For example, a default value or an estimated value can be assigned to fill the missing value. However, such assigned values could affect the level of accuracy quality.

iv. ***Consistency;*** The representation of the data value is consistent in all cases.

The total score of the data quality is defined as follow;

$$S_{quality} = (S_{accuracy} + S_{timeliness} + S_{completeness} + S_{consistency} + S_{years}) / 5 \qquad (4.1)$$

4.4 MODELING

The purpose of modeling phase is to perform all modeling activities in order to generate most appropriate data mining model. During these activities, some alternative models are explored in order to find the most effective and useful solution with optimum results among the other models. Therefore it is built final data mining model. While performing all these activities, there need to return to the data preparation phase in order to apply some additional data treatment activities. Main steps of this phase are as follows;

i. **Select Modeling Technique;** After defining a clear business goal and selecting data mining modeling algorithm should be selected. In the following chapter, common data mining techniques and characteristics of them will be summarized. Statistical methods can be used as logistic regression as well as no statistical or blended methods like neural networks, genetic algorithms, and decision tress.

ii. **Building Model;** A various data mining techniques should be choose and applied into the selected dataset in order to select appropriate methods in this phase.

iii. **Assessing Model;** Alternative models are being assessed according to the some success criteria of associated data mining technique and experience of researcher. This activity is different than the evaluation phase of the data mining, that step is interested in the correction of associated data mining techniques and the model selection even though evaluation phase is focused on all other results of whole project.

After constructing an appropriate model whose settings are calibrated in order to produce optimize results, evaluation phase should be followed.

4.5 EVALUATION

The Model and the outputs should be determined if they are achieves the business objectives that are set in the first phase of data mining project. The interpretation of data is very critical issue because evaluating the model from the business perspective is the approval of whole project, if it is successful or not. The evaluation of the model should be performed according to the cost benefit analysis and return on investment.

There are two issues are essential in evaluation phase. First one is to define the way of recognizing the business value from knowledge patterns in data mining stage in modeling phase. Other one is choosing visualization tool in order to show the data mining results. The mined raw data is very complicated to analyze and needs to be processed and combined according to the business purposes. That operation heavily depends on combining the outcomes properly that are produced at the end of business understanding, data understanding, data preparation and modeling phases where data mining techniques are effectively used. In order to interpret knowledge patterns properly, it is important to select appropriate visualization tool. Efficient business decisions are made by using proper interpretation of knowledge pattern.

It is very important to evaluate whole model in the perspective of business goals of the project. One of the basic ways to evaluate a model is test the results in the real world. It should be selected a sample of data in the whole data set and test a prediction of the model and observe the results how they are close to the predicted results.

At the end of this phase the following issues should be completed.

i. Evaluating one or more model are decided and generated in the modeling phase for checking effectiveness and quality before deploying them.
ii. Determining if the selected model achieves the objectives are set in the phase of business understanding.
iii. Establish if some business and research problems are occurred or not
iv. Decision of using the data mining results.

After completing all these tasks, deployment that is last phase should be followed

4.6 DEPLOYMENT

After building the model, the findings of the project must be reported back to the management to decide. So that management can apply the new approach of their business environment in the company. In other words, the knowledge discovered from data mining project is needed to be formed and presented to the project sponsor, top management and related project holders.

Depending on the needs of requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. As a common practice, constructed data mining models are adopted in software therefore the requested data will be reported in requested format effectively. Any kind of development over the data mining tool in order to perform necessary reporting such as customer relationship, churn prediction, pricing, product customization .etc

At the end of this phase the following issues should be completed

i. Performing activities in order to provide the models in usage: Creation of model is not at the end of project.
ii. Provide an evidence of the deployment process : Generating reports
iii. Provide some complex deployments : Implement a parallel data mining process for another department

4.7 TASKS IN DATA MINING

Data mining tasks can be categorized in six groups according to the purpose of data mining model.

i. Description
ii. Estimation
iii. Prediction
iv. Classification
v. Clustering
vi. Association

Classification, estimation and prediction are directed data mining applications; however the rest of the tasks are undirected ones. Association, clustering and description tasks have a goal of discovering hidden structure of data with no respect to a specific target variable (Berry & Linoff, 1997).

4.7.1 Description

The one of the main aims of the researcher in data mining process is to describe the hidden pattern and trend lying in data set. Because of that reason data mining models should be as transparent as possible to perform transparent interpretation therefore hidden knowledge pattern should be discovered by data miners.

For instance, a pattern present people who had been laid off are more likely to support the opposition candidate will be extracted from the results of a new public survey. The explanation for this reference will be the financial problems that the laid off works are living and preferring to vote another candidate that can do better than the current president (Larose, 2005).

The results of data mining study should be explanatory. The pattern should cover clear patterns that can be used in the decision making process. Some data mining methods like decision trees are easier to understand and have more human friendly explanations. Some other methods like neural networks or genetic algorithm are tended to be more complex and difficult to interpret (Witten & Frandk, 2005). Explanatory data analysis is a way to high quality description generation. The method investigates the data by graphs to find meaningful pattern and trends (Larose, 2005).

4.7.2 Classification

There exists a target categorical variable in classification which can be divided into number of classes. As an example, the target variable can be age of a population and the population can be divided to four segments like child, young, middle-aged and old. It is also possible for researchers to classify the population according to their ages by training data. Suppose that for a smaller group of people age, income and gender data are known. This data can be used as training data to catch the relationship between gender, income and age. For more complex data, the model can classify the people according to their income by using the gender and age data. It can achieve it from what the model had learnt from the training set. For example the model will guess a male engineer with a high income as a member of middle aged class (Larose, 2005).

Other examples of classification tasks in business are as follows;

- Assessing whether a customer is profitable or not
- Diagnosis whether a patient has a specific disease
- Determining whether a student passes a class or fail
- Determining the type of drug that a doctor should prescribe to a particular patient

If they are two or three dimensional relationship in the data, it is possible to analyze with graph and plots. But when it comes to multidimensional classification data mining is forced to be used. The most common methods of data mining used for classification are k-nearest neighbor, decision tree and neural network (Berry & Linoff, 1997).

4.7.3 Estimation

Estimation is similar to classification method but the difference between classification and estimation is that the target variable is numerical in estimation rather than categorical. Researchers work with the current data and use the relationship between target variable and predictors in new observations. As an example normal blood pressure of adults can be modeled by using gender, age, height and weight. The model enables to calculate normal blood pressure of a new patient by gender, age, height and weight (Larose, 2005).

Some examples of estimation are as follows;

- Estimating the amount of money that a family of five will spend for kitchen expense in a month.
- Estimating the number of goals per match Fenerbahçe will score in Turkish League.
- Estimating the normal weight of an adult using gender and height.
- Estimating the GDP per person using economic values.

There are several and widely used methods for estimation. Point estimation, confidence interval estimation, simple linear regression and correlation, multiple regression and neural networks can all be used for estimating future values (Witten & Frandk, 2005).

4.7.4 Prediction

Any techniques and methods used for classification and estimation may also be used for prediction according to the nature of business needs. In terms of techniques in usage, they have a lot in common except that prediction interested in the future values that are expected. These include the traditional statistic methods for estimation such as regression and correlation techniques as well as knowledge discovery methods used in data mining, such as decision tree, neural network and k-nearest neighbor.

The examples of prediction tasks in business and research are as follows;

- Predicting sales amount of a product for six months.
- Predicting the breast cancer survivability in next 5 years.
- Predicting the campaign of football team in first league.
- Predicting the general election result of Turkey for June'2001
- Predicting Gross National Income per capita of Turkey for next five years.

Prediction analysis is mainly associated to regression techniques. The purpose of the prediction is to discover the relationship between dependent and independent variables by using historical data in order to complete prediction task realistically and successfully.

4.7.5 Clustering

Clustering refers to the grouping of records; observations or situations are divided into smaller groups containing similar objects. A cluster can be defined as "the combination of elements that are alike each other and different from the elements of other clusters". Apart from classification, clustering does not include a target variable. Clustering is only interested in dividing the data set into homogenous subgroups. The segmentation studies are important examples of clustering methods (Larose, 2005).

Examples of clustering tasks in business and research are as follows;

- Target marketing of a specific product for a small capitalization business that does not have a large marketing budget.
- Reducing the dimension for the data sets that have hundred of attributes.
- Producing effective marketing campaigns for the group of customers that have similar attributes.

Clustering can frequently be applied as initial step in data mining process with the results of clustering are being used as an further input into different data mining techniques such as neural network, k-means clustering….etc

4.7.6 Association

Association task in data mining is focused on the job of finding relationships between variables. It is commonly used in data mining area and also known as market basket analysis or affinity analysis (Delen & Walker & Kadam, 2005). The most common example for association is as follows; It is found that 200 people from 1000 people shopping on Sunday bought beer. The 50 people from 200 people also bought diapers with the beer. It is surprising for these two products to be sold together but it happens. The association rule generated from this shopping experience is as follows;
"If buy diapers, then buy beer", the support of this statement is 200/1000= 0, 20 and confidence is 50/200= 0, 25

Some examples of association tasks in business are as follows;

- Investigating the percent of potential customer who are offered to involve the special pricing schedule to be a customer of Turkcell
- Finding out which items in super market are definitely purchased together for families with children
- Determining the percents of effects of smoking for lung cancer.

There are two algorithm for generating association rules called Generalized Rule Induction (GRI) algorithm and priori algorithm.

5. DATA MINING METHODS

5.1 DECISION TREES

The method of decision trees is a classification method among the data mining techniques. Basically it includes the building a decision tree with decision nodes, connecting by branches. It is a kind of generating a tree structure starting from root node until ending by leaf nodes. All the rules and conditions must be set at the beginning of constructing a decision tree diagram and all attributes should be tested at the decision nodes therefore all possible results will be finalized and come up in a branch. Each branch conducts to another decision or finishing to a leaf node. Please find the example of simple decision tree in Figure 5.1

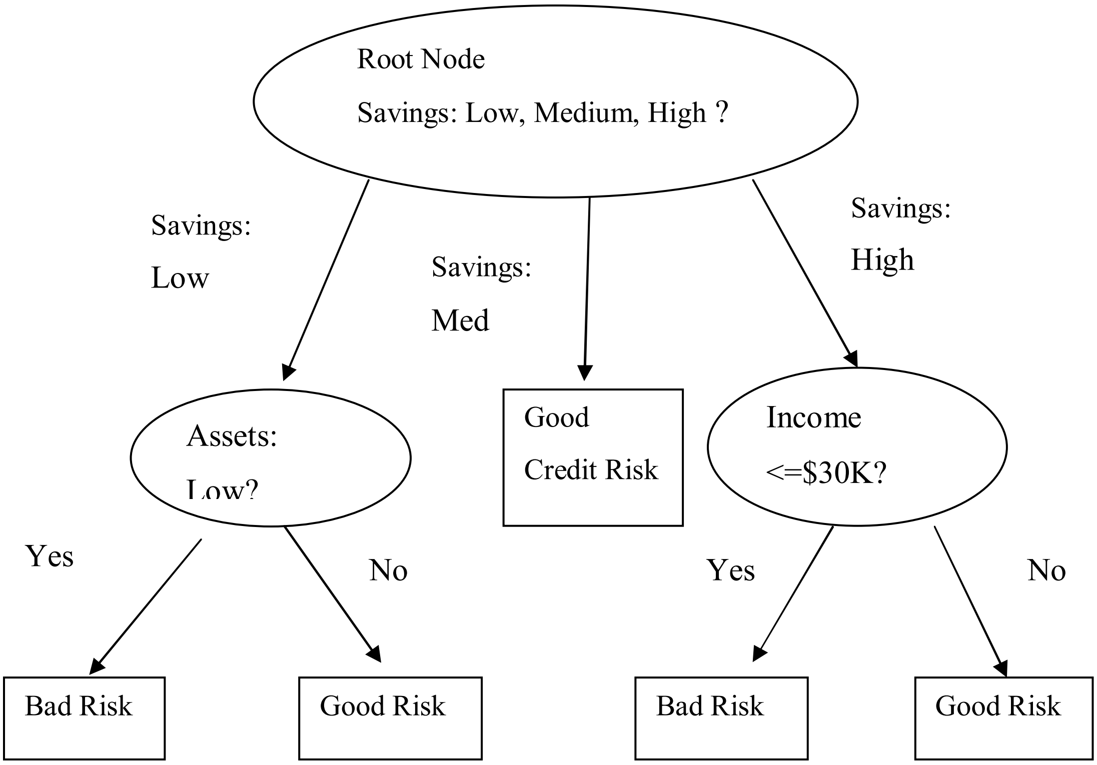


Figure 5.1 Simple decision tree
Source: Larose, 2005

Decision tree algorithms contain some supervised learning techniques such as using pre classified target variables. Because of that, the target variables must be selected and defined in the sample of data set. Also the sample of data set should contain all types of conditions and variables in order to be applied the algorithm properly. The decision tree study and improve by processing sample of data. Therefore, performing all the steps of decision algorithm and using all types of variable within the defined conditions. The results will be produced systematically by performing the classification and prediction steps one by one. Because of that target feature classes should be carefully and clearly defined from each other, there must be confusion on these terms.

One of the most useful aspects of decision trees is their explainable and understandable easily even all the rules while setting decision rules. Decision rules are applied any path starting from root node to leaf. But the whole decision rules are equal to the decision tree in terms of classification purposes.

There are two important approaches are applied in decision tree;

i. **The classification and regression trees (CART) algorithm**; It is proposed by Breiman in 1984 and has a statistical approach as binary containing two branches for each decision node (Breiman & Friedman & Olshen & Stone, 1984). CART is stepping on the divisions repeatedly applied in order to reach target attributes by making decision on sample of data set. The CART algorithm expands the tree by executing each decision node.

ii. **C4.5 algorithm**; The C4.5 algorithm is Quinlan's extension of his own ID3 algorithm for generating decision trees (Quinlan, 1992). Comparing to CART, the C4.5 algorithm regularly check each decision node in order to select appropriate split until there is no splits are exist. Also, binary splits are not limited in C4.5 compare to CART that is always produces a binary tree. But C4.5 produces a tree of more variables. C4.5 produces separate branch for each value of attribute class

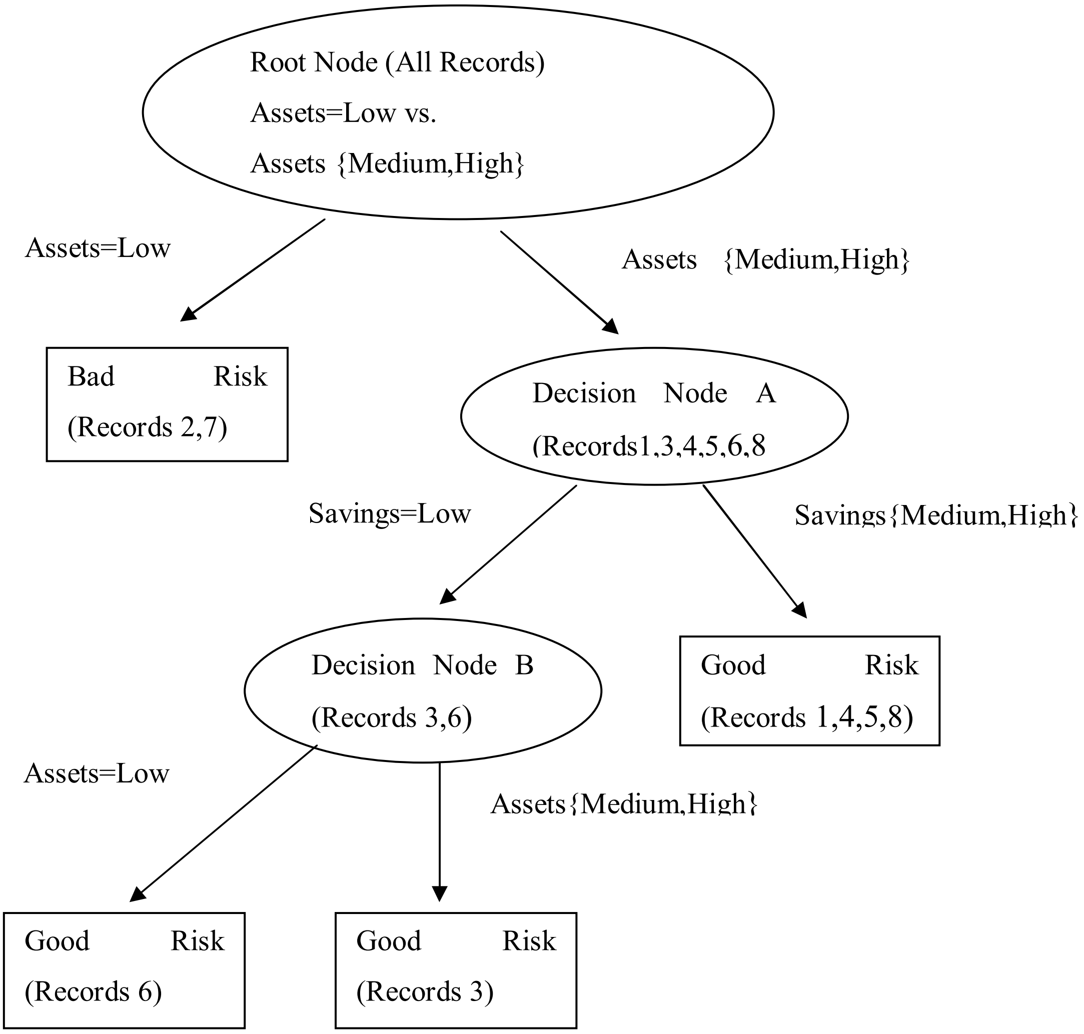


Figure 5.2 CART decision tree
Source: Larose, 2005

The decision algorithm has the following major steps;

i. Target variables are selected from data set carefully and be sure that all the variables must be presented in the data set but independent variables are selected by user.

ii. Each variable that is any impact on the outcome should be questioned and evaluated. Also an iterative function is performed on the grouping values that contain outcome impact

iii. After each variable have been calculated for grouping, it is decided the most predictive for dependent variables and is used for creating the leaf nodes of the tree.

Most people understand decision trees intuitively. This is the greatest strength of technology. The negative side of decision trees is the fact that they get harder to manage as the complexity of the data increases. This is because of the increasing number of branches in the tree. There is also an issue with the handling of missing data, because without a data element being present, how do you traverse a tree node dependent on the data (Groth, 1999).

5.2 NEURAL NETWORK

As a supervised learning technique, neural networks are generally used to produce effective solutions for complex problems by building prediction models. It is widely used in finance, manufacturing and health sector, such as constructing models for fraud in credit card transactions or cancer survivability researches.

The inspiration for neural networks is the complex learning systems in animal brain consisted of closely interconnected sets of neurons. Although a particular neuron may be relatively simple in structure, consistent networks of interconnected neurons could perform complex learning tasks such as classification and pattern recognition. The main important issue is complex structure is able to manage the learning process (Larose, 2005).

Figure 5.3 demonstrates a real neuron and artificial neuron models and their inputs and the outputs. Dendrites of a neuron cell collect the information and through axon located in cell body it transfers it to the neighbor neuron by dendrites. In this process the neuron generates a nonlinear response to the other neurons when a threshold is reached. The artificial neuron is also able to collect data set from the neighbor neurons and gather it by using a predefined usually non linear activation function (Witten & Frandk, 2005). The result of the activation function will be notified to the other neurons to be used as an input for their activation function (Larose, 2005).

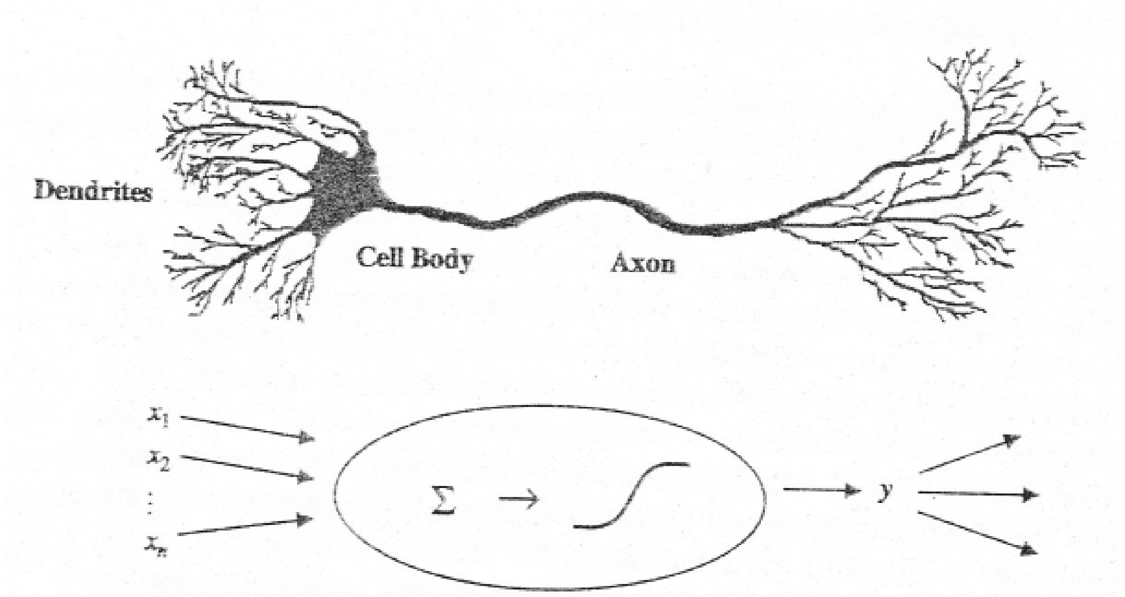


Figure 5.3 Real neuron and artificial neuron models
Source: Larose, 2005

All of the attribute values in neuron network model should be coded with a standard type. The standardization usually includes taking values between zero and one. The standardization issue is important for back propagation.

The variables should be normalized. The chosen normalization method for continuous variables is the min-max normalization; however for category variables. It is more complex to normalize the variables. Because of the output generated by neural networks being continuous, the method is widely preferred for estimation and prediction (Witten & Frandk, 2005).

Artificial neural network are constructed by simulating the neurons in human brain. Each link defined as a processing element (PE). Neural network gains information from the experiences and it is very successful to discover unknown relationship between input data set and the associated output. Neural network discovered knowledge pattern in data set and standardize relationships found in the data then predict the outcome. Neural network is very good at predicting complex processing.

Processing elements (PE) summarize and transform the data by using a serious of mathematical function. A PE that is linked to inputs and outcomes has limited ability to produce a solution but when the neuron or processing elements (PEs) connected to each other and generates a system, it is created an intelligent model by interconnecting PEs in any number of ways and they can be retrained hundreds of iterations to produce effective outcomes. The connections between inputs and output are modified in the process of network training as evaluating strength and/or weight of it. While producing the proper outcome, it is very important to evaluate the changes (increase or decrease) on the strength of connection. Strength of a connection depends on a weight and receives on a trial and error process. A mathematical process is used for that process which is called a learning rule, in order to adjust the weights

Training operation continues until a neural network produces outcome values that match the known outcome values within a specific accuracy level or until it satisfies some other criteria’s are provided.

Each of the processing units takes many inputs and generates an output that is a nonlinear function of weighted sum of the inputs. The weights assigned to each of the inputs are obtained during a training process (often a propagation) in which outputs generated by the net are compared with target output. The answers you want the network produce are compared with generated outputs and the deviation between them is used as feedback them is used as feedback to adjust the weights (Groth, 1999).

The greatest strength of neural networks is their ability to accurately predict outcome of complex problems. Neural networks are preferred technique in performing estimation or continue numeric outputs, which are popular in financial markets and manufacturing (Groth, 1999).

5.3 GENETIC ALGORITHMS

The inspiration for genetic algorithm is based on the biological evolution and natural selection such as neural networks is. In other words, genetic algorithms are methods of integrative optimization those are based on processes in biological evolution.

Evolution makes it possible for animals to be more adapted to the nature and it increases the compatibility to habitat that they are living in. This improvement is a result of most suitable genetic material to be selected by ancestors and transmitted to the new generation. This idea also stands in the base of genetic algorithm method. Genetic algorithms are applied to optimization studies because of its nature (Zikmund & McLeod & Gilbert, 2003).

The usage of genetic algorithm is not as common as the other methods and it does not common exist in most commonly used data mining software applications. Genetic algorithm specially works on optimization, which is not a favorite issue in data mining like clustering or classification. It is usually used with other methods to increase the overall performance and specially generated software is used for this purpose (Berry & Linoff, 1997).

Genetic algorithms are a group of mathematical procedures that use the process of genetic inheritance. The algorithms are very good at being applied to a wide variety of analytic problems successfully and produce a appropriate evaluation. The human understanding of data and automatic analysis of data are combined in data mining in order to discover patterns or key relationships.

Basic operators in Genetic algorithm is shown in Figure 5.4

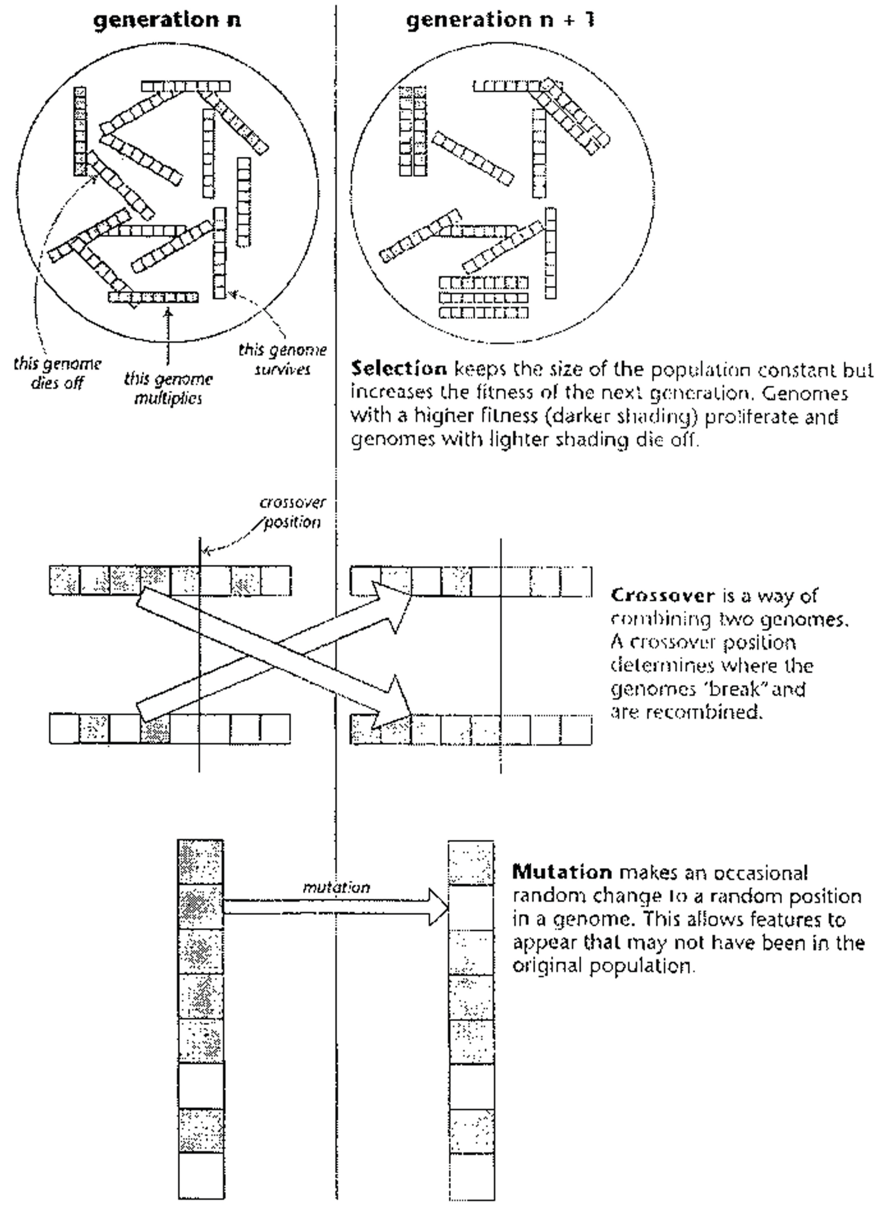

generation n
generation n + 1
this genome dies off
this genome multiplies
this genome survives
Selection keeps the size of the population constant but increases the fitness of the next generation. Genomes with a higher fitness (darker shading) proliferate and genomes with lighter shading die off.
crossover position
Crossover is a way of combining two genomes. A crossover position determines where the genomes 'break" and are recombined.
mutation
Mutation makes an occasional random change to a random position in a genome. This allows features to appear that may not have been in the original population.

Figure 5.4 The Basic operators in genetic algorithm
Source: Witten & Frandk, 2005

Genetic algorithms require definite data structure and operate on a group of data with characteristics classified in a certain form.

The typical genetic algorithm processes as follows (Olson & Delen, 2008);

i. Randomly select parent

ii. Reproduce through crossover. Reproduction is the operator choosing which individual entities will survive. In other words, some objective function or selection characteristic is needed to determine survival. Crossover relates to changes in future generations of entities.

iii. Select survivors for the next generation through a fitness function.

iv. Mutation is the operation by which randomly selected attributes of randomly selected entities in subsequent operations are changed.

v. Iterate until either a given fitness level is attained or the preset number of iterations is reached.

Genetic algorithm parameters include population size, crossover rate (the probability that a certain entity mutates) and the mutation rate (the probability that a certain entity mutates).

The advantage of genetic algorithm is in the ability of dealing with complicated data sets. They are also very easy to develop and validate efficiently, quickly when they are planned to be applied into a large data set. That makes the method very popular to choose especially for large data set in order to produce rapid progress for efficient solution. It makes the method capable of defining global optima in spite of nonlinear problems existence by using mutation features. The method also does not need to know any information for distribution of the data.

The main disadvantage of genetic algorithms is that the associated data sets are required to map into a form where attributes are separated to the values for the genetic algorithm to work with. But there are some cases occurred that great deal of detail information can be lost while dealing with continues variables. Forming data into a categorical structure can unintentionally cause biases in the data.

There are also some limit ranges related to the size of data set that are supposed to be analyzed with genetic algorithm. The sampling process will be necessary especially for very large data sets and the process leads to different results runs over the same data sets.

Genetic algorithms can be very useful within a data mining analysis dealing with more attributes and many more observations. It saves the brute force checking of all combinations of variable values, which can make some data mining algorithms more effective. However, applications of genetic algorithms require expression of the data into discrete outcomes, with a calculated functional value upon which to base selection. This does not fit all data mining applications but they are useful when they are fit. Genetic algorithms are usually applied in conjunction with other data mining techniques. They can be used to enhance the efficiency of other methods or can be more directly applied (Olson & Delen, 2008).

5.4 K-NEAREST NEIGHBOR ALGORITHM

K-nearest neighbor (KNN) algorithms are based on learning analogy. When given an unknown sample, a KNN searches the pattern space for the KNN that are closest to the unknown sample. Closeness is defined in terms of distance. The unknown sample is assigned the most common class among its KNN. The major advantage of this approach is that it is not required to established predictive model before classification. The disadvantages are that KNN does not produce a simple classification probability formula and its predictive accuracy is highly affected by measure of distance and cardinality k of the neighbor (Yeh & Lien, 2007).

KNN algorithms are mainly used for classification as well as estimation and prediction purposes. K-nearest neighbor is an example of instance-based learning and the following steps are applied. First of all, the training data set should be stored, so that a classification for a new unclassified record can be discover basically by comparing it to the similar records in the training set.

While building a classifier using by k-nearest neighbor algorithm, there will be some issues should be clarified (Olson & Delen, 2008);

i. How many neighbors should we consider? What is k?

ii. How do we measure distance?

iii. How do we combine the information from more than one observation?

iv. Should all points be weighted equally or should some points have more influence than others?

There is a method of determining which records are similar to the new, unclassified record; we need to establish how these similar records will combine to provide a classification decision for the new record. It is needed a combination function and the most basic combination function is simple un-weighted voting. Please find the main steps of it below (Olson & Delen, 2008);

i. Before running the algorithm, decide on the value of k, that is, how many records will have a voice in classifying the new record.
ii. Then, compare the new record to the k nearest neighbors, that is, to the k records that are of minimum distance from the new record in terms of the Euclidean distance or whichever metric the user prefers.
iii. Once the k records have been chosen, then for simple un-weighted voting, their distance from the new record no longer matters. It is simple one record, one vote.

There is also weighted voting and one may feel that neighbors that are closer or more similar to the new record should be weighted more heavily than more distant neighbor. Instead, the analyst may choose to apply weighted voting, where closer neighbor have a larger voice in the classification decision than do more distant neighbor. Weighted voting also makes it much less likely for ties to arise. In weighted voting, the influence of a particular record is inversely proportional to the distance of the record from the new record to be classified (Olson & Delen, 2008).

5.5 HIEARHICAL& K-MEANS CLUSTERING

Clustering refers to the grouping of records, observations or cases into classes of similar objects. A cluster is a collection of records that are similar to one another and dissimilar to records in other clusters. Clustering differs from classification in that there is no target variable for clustering. The clustering task does not try to classify, estimate or predict the value of a target variable. Instead, clustering algorithms seek to segment the entire data set into relatively homogeneous subgroups or clusters where the similarity of the records within the cluster is maximized and the similarity to records outside this cluster is minimized. Examples of clustering tasks in business and research include (Yeh & Lien, 2007);

i. Target marketing of a niche product for a small – capitalization business that does not have a large marketing budget
ii. For accounting auditing purposes to segment financial behavior into benign ans suspicious categories
iii. As a dimension-reduction tool when a data set has hundred of attributes
iv. For gene expression clustering, where very large quantities of genes may exhibit similar behavior

5.5.1 Hierarchical Clustering

Clustering algorithms are either hierarchical or nonhierarchical. In hierarchical clustering, a treelike cluster structure (dendrogram) is created through recursive portioning (divisive methods) or combining (agglomerative) of existing clusters. Agglomerative clustering methods initialize each observation to be a tiny cluster of its own. Then, in succeeding steps, the two closest clusters are aggregated into a new combined cluster. In this way, the number of clusters in the data set is reduced by one at each step. Eventually, all records are combined into a single huge cluster. Divisive clustering methods begin with all the records in one big cluster with the most dissimilar records being split off recursively into a separate cluster, until each record represents its own cluster. Because most computer programs that apply hierarchical clustering use agglomerative methods (Yeh & Lien, 2007).

Distance between records is rather straight forward once appropriate recoding and normalization has taken place. There are several criteria for determining distance between arbitrary clusters A and B (Yeh & Lien, 2007);

i. ***Single Linkage;*** Sometimes termed the nearest-neighbor approach is based on the minimum distance between any record in cluster A and any record in cluster B. In other words, cluster similarity is based on the similarity of the most similar members from each cluster. Single linkage tends to form long, slender clusters, which may sometimes lead to heterogeneous records being clustered together.

ii. ***Complete Linkage;*** Sometimes termed the farthest-neighbor approach is based on the maximum distance between any record in cluster A and any record in cluster B. In other words, cluster similarity is based on the similarity of the most dissimilar members from each cluster. Complete-linkage tends to form more compact, spherelike clusters, with all records in a cluster within a given diameter of all other records.

iii. ***Average Linkage*** is designed to reduce the dependence of the cluster-linkage criterion on extreme values such as the most similar or dissimilar records. In average linkage, the criterion is the average distance of all the records in cluster A from all the records in cluster B. The resulting clusters tend to have approximately equal within-cluster variability.

Please find an example of how these linkage methods work, using the following small, one-dimensional data set ;

12 15 19 25 26 28 35 43 43 55

i. Single-Linkage Clustering;

The initial step for agglomerative methods is to assign record to its own cluster. Single linkage aims the minimum distance between any records in two clusters. The Figure 5.5 illustrates how single-linkage clustering is accomplished for this data set. The minimum cluster distance is between the single-record clusters that each contain the value 43 for which the distance must be zero for any valid metric. Therefore, these two clusters are combined into a new cluster of two records, both of value 43 as shown in figure 5.5. After step 1, only nine (n-1) clusters remain. Next, in step 2, the cluster containing values 25 and 26 are combined into a new cluster, since their distance of 1 is the minimum between any two clusters remaining.

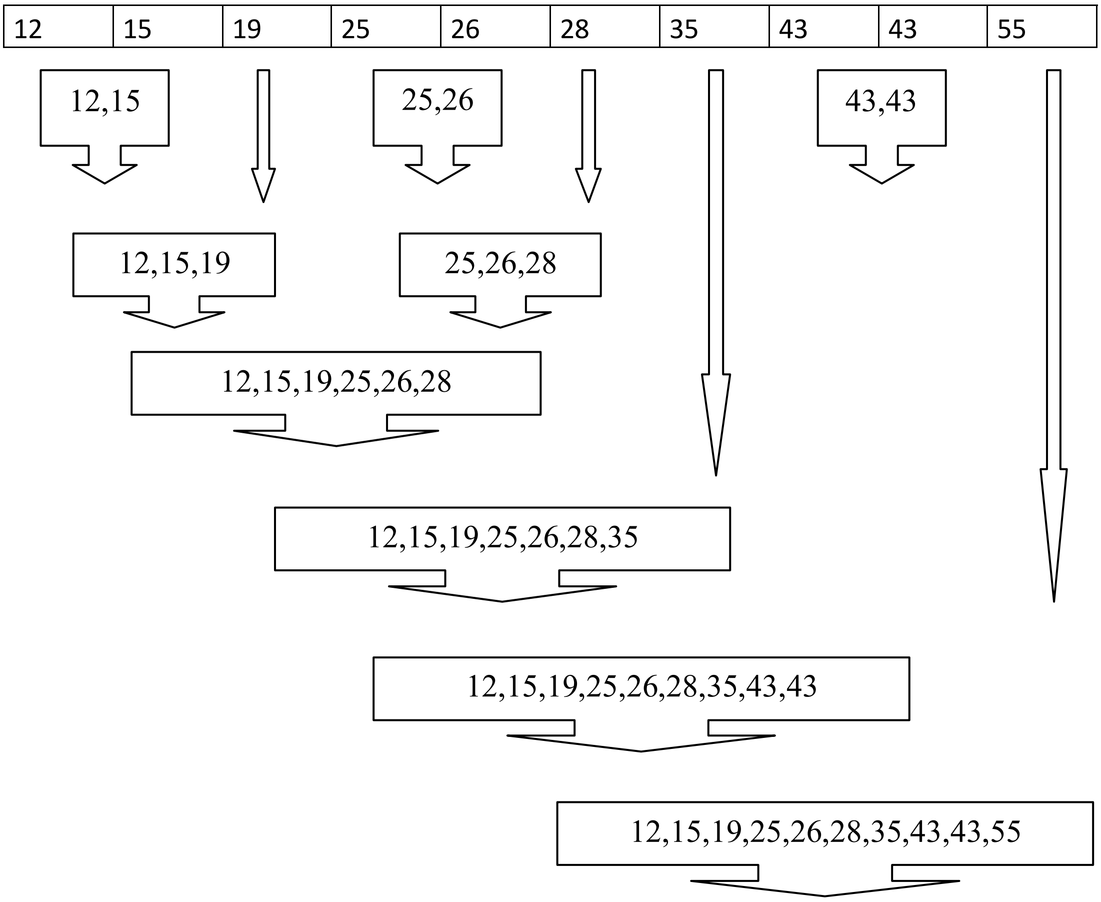


Figure 5.5 Single-linkage agglomerative clustering on the sample data set

Please find the remaining steps below;

i. *Step 3:* The cluster containing values 25,26 (cluster {25,26}) is combined with cluster {28}, since the distance between 26 and 28 (the closest records in each cluster) is two, the minimum among remaining clusters.

ii. *Step 4 :* Clusters {12} and {15} are combined.

iii. *Step 5 :* Clusters {12,15} is combined with cluster {19}, since the distance between 15 and 19 (the closest records in each cluster) is four, the minimum among remaining clusters.

iv. *Step 6:* Cluster {12,15,19} is combined with cluster {25,26,28}, since the distance between 19 and 25 is six, the minimum among remaining clusters.

v. *Step 7 :* Cluster {12,15,19,25,26,28} is combined with cluster {35},since the distance between 28 and 35 is seven, the minimum among remaining clusters

vi. *Step 8 :* Cluster {12,15,19,25,26,28,35} is combined with cluster {43,43},since the distance between 35 and 43 is eight, the minimum among remaining clusters

vii. *Step 9 :* Cluster {12,15,19,25,26,28,35,43,43} is combined with cluster {55},this last cluster now contains all the records in the data set.

ii. Complete-Linkage Clustering;

Complete-linkage aims to minimize the distance among the records in two clusters that are farthest from each other. Please find the remaining steps below;

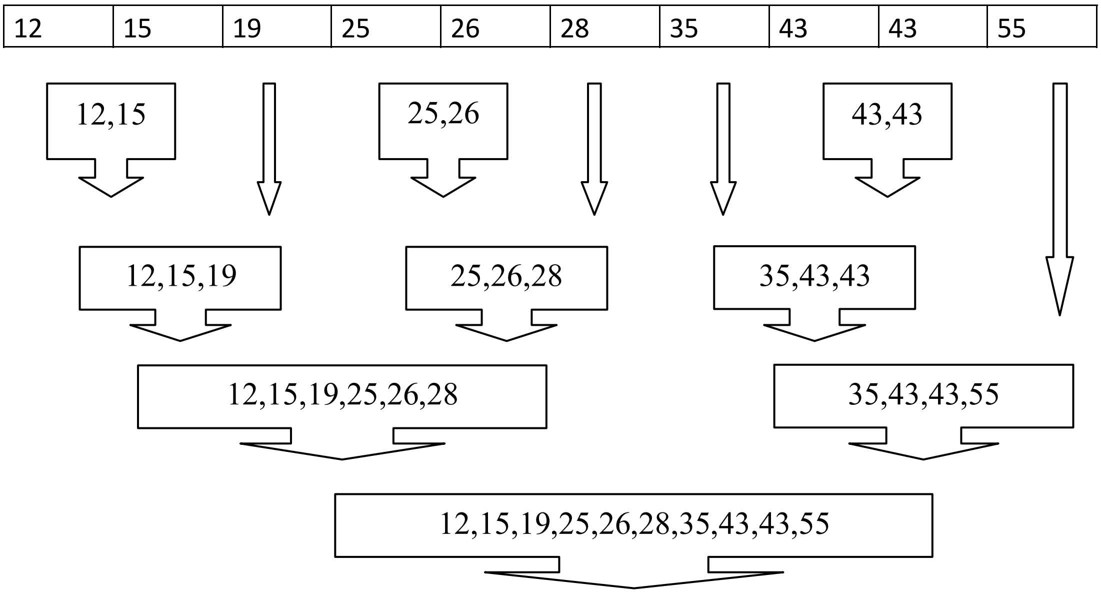


Figure 5.6 Complete-linkage agglomerative clustering on the sample data set

i. *Step 1:* There is no difference between single linkage and complete linkage at the step 1 as each cluster contains a single record.

ii. *Step 2:* As in the single linkage the clusters containing values 25 and 26 are combined into a new cluster. There is no difference between these two method at this step also.

iii. *Step 3:* In the single linkage, cluster {25,26} was at this point combined with cluster {28}. On the other hand, complete linkage searches the farthest neighbors not the nearest neighbors. The farthest neighbors for these two clusters are 25 and 28 for a distance of 3. That is the same distance separating clusters {12} and {15}. The complete-linkage criterion is silent regarding ties so that it is arbitrarily select the first such combination found, thus combining the clusters {12} and {15} into a new cluster.

iv. *Step 4:* Cluster {25,26} is combined with cluster {28}

v. *Step 5:* Cluster {12,15} is combined with cluster {19}, since the complete linkage distance is 7, the smallest among remaining clusters.

vi. *Step 6:* Cluster {35} is combined with cluster {43,43}, with a complete linkage distance of 8

vii. *Step 7:* Cluster {12,15,19} is combined with cluster {25,26,28}, with a complete linkage distance of 16.

viii. *Step 8:* Cluster {35,43,43} is combined with cluster {55}, with a complete linkage distance is 20.

ix. *Step 9:* Cluster {12,15,19,25,26,28} is combined with cluster {35,43,43,55}. All records are now contained in this last large cluster.

5.5.2 K-Means Clustering

The k-means clustering algorithm is a straightforward and effective algorithm for finding clusters in the data. The algorithm steps are below (Yeh & Lien, 2007);

i. *Step 1:* Ask the user how many clusters k the data sets should be partitioned into.

ii. *Step 2:* Randomly assign k records to be the initial cluster center locations.

iii. *Step 3:* For each record, find the nearest cluster center. Therefore, each cluster center "owns" a subset of records, thereby representing a partition of the data set, have k clusters, C1,C2,C3……Ck

iv. *Step 4:* For each of the k clusters, find the cluster centroid, and update the location of each cluster center to the new value of centroid.

v. *Step 5:* Repeat steps 3 to 5 until convergence or termination.

The nearest criterion in step 3 is usually Euclidean distance, although other criteria may be applied as well. The cluster centroid in step 4 is found as follows. Suppose that we

have n data points (a_1, b_1, c_1), (a_2, b_2, c_2),……….. (a_n, b_n, c_n), the centroit of these points is the center of gravity of these points and is located at point
($\sum a_i / n$, $\sum b_i / n$, $\sum c_i / n$)

The algorithm terminates when the centroids no longer change. In other words, the algorithm terminates when for all clusters C_1,C_2,C_3……C_k, all the records owned by each cluster center remain in that cluster.

5.6 KOHONEN NETWORKS

Kohonen networks were introduced in 1982 by Finnish researcher Tuevo Kohonen (Kohonen, 1982). In spite of being applied to image and sound analysis, Kohonen networks are a commonly used for cluster analysis efficiently. Kohonen networks express a type of self-organizing map (SOM) which itself produce a special class of neural network.

The goal of self- organizing maps is to convert a complex high-dimensional input signal into a simpler low-dimensional discrete map (Haykin, 1990). Therefore, as structures of SOMs are entirely appropriate for cluster analysis where there will be some hidden patterns among records and fields are perceived. The output nodes into clusters of nodes in SOMs structure where nodes in closer proximity are more similar to each other than the other nodes are farther apart. Ritter has shown that SOMs represent a nonlinear generalization of principal component analysis, another dimension-reduction technique.

The structure of self-organizing maps is built on competitive learning and associated output nodes compete among the others to be the succeeded node. The particular input observation is activated accordingly. As Haykin describes it; The neurons become selectively tuned to various input patterns (stimuli) or classes of input patterns in the course of a competitive learning process.

A typical SOM architecture is shown in Figure 5.7 below.

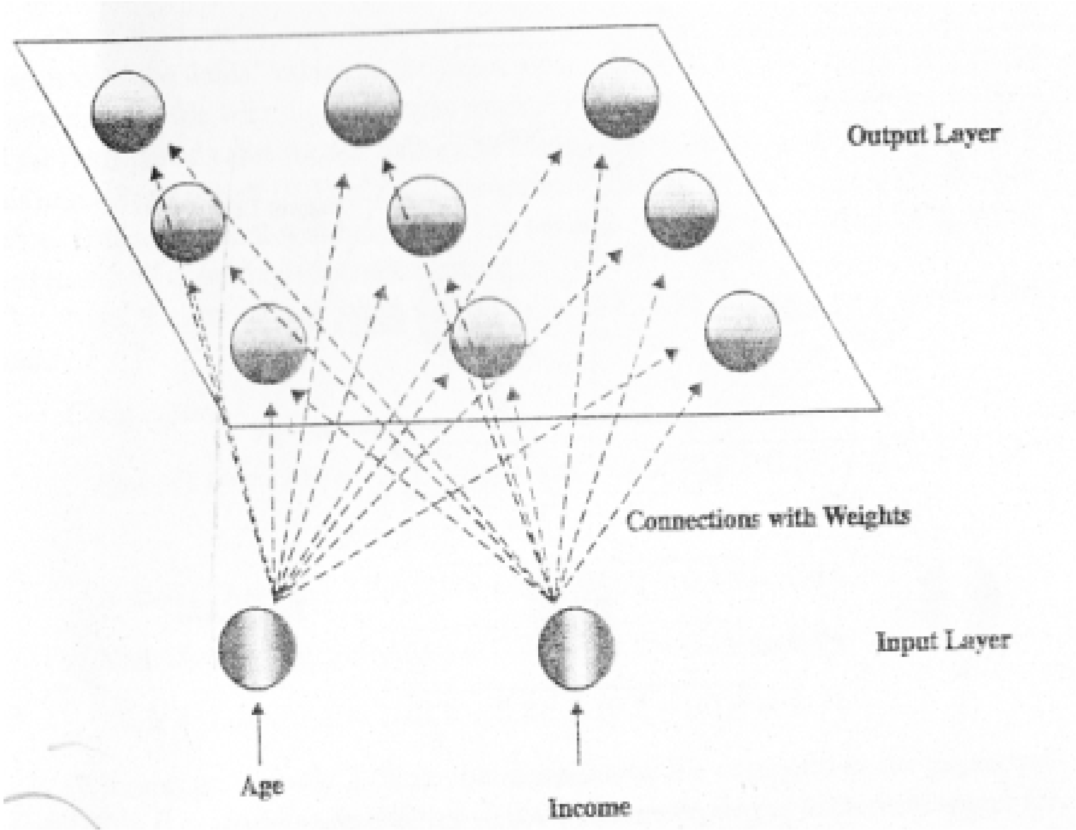


Figure 5.7 Topology of a simple self-organizing map for clustering records by age and income
Source: Larose, 2005

The input layer is shown at the bottom of the figure, with one input node for each field. Just as with neural networks, these input nodes do no processing themselves but simply pass the field input values along downstream (Larose, 2005).

Self –organizing maps expose three distinctive processes (Larose, 2005).

i. ***Competition:*** The output nodes compete with each other to produce the best value for a particular scoring function, most commonly the Euclidean distance. In this case, the output node that has the smallest Euclidean distance between field inputs and the connection weights would be declared the winner (succeeded).

ii. ***Cooperation:*** The winning node therefore becomes the center of a neighborhood of excited neurons. This emulates the behavior of human neurons, which are sensitive to the output of other neurons in their immediate neighborhood. In self-organizing maps, all the nodes in this neighborhood share in the excitement or reward earned by the winning nodes, that of adaptation. Thus, even though the nodes in the output layers are not connected directly, they tend to share common features, due to this neighborliness parameter.

iii. ***Adaptation:*** The nodes in the neighborhood of the winning node participate in adaptation that is learning. The weights of these nodes are adjusted so as to further improve the score function. In other words, these nodes will have an increased chance of winning the competition once again, for a similar set of field values.

Kohonen networks are self – organizing maps that expose Kohonen learning. Suppose that it is considered the set of m field values for the nth record to be an input vector

$X_{n=} X_{n1,} X_{n2,} X_{n3,} X_{n4\ldots..,} X_{nm}$

and the current set of m weights for a particular output node j to be a weight vector

$w_{j=} w_{1j,} w_{2j,} w_{3j,} w_{4j,\ldots..} w_{mj.}$

In Kohonen learning, the nodes in the neighborhood of the winning node adjust their weights using a linear combination of the input vector and the current weight vector;

$$W_{ij,new} = W_{ij,current} + r\ (X_{ni} - W_{ij,current}) \qquad (5.1)$$

Where r, $0 < r < 1$, represents the learning rate, analogous to the neural networks case. Kohonen indicates the learning rate should be a decreasing function of training epochs (run through the data set) and that a linearly or geometrically decreasing r is satisfactory for most purposes.

5.7 BAYESIAN NETWORKS AND CLASSIFICATION

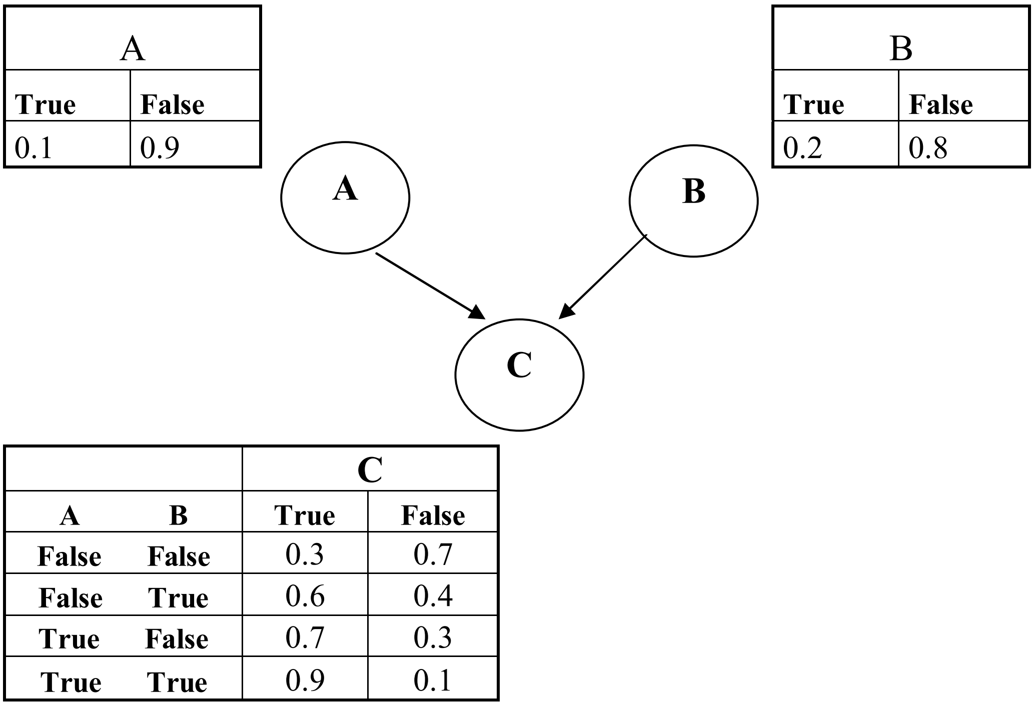
Bayesian networks can be located as a intersection of artificial intelligence, statistics, probability and have an important place among the techniques of data mining and knowledge discovery. The main features of Bayesian networks are being probabilistic graphic model as a general class of model that comes up from the combination of graph and probability theories. The main advantage of the model is to be able to handle complex probabilistic models by separating them into smaller, tractable components. A probabilistic model is characterized in a graph where nodes symbolize stochastic variables and arcs symbolizes among such variables. A probabilistic model is named as Bayesian network when the graph connects its variables is a directed acyclic graph (DAG). The conditional independence assumptions are represented by this graph and theses assumptions are used to resolve the joint probability distribution of the network variables therefore the process of learning from large database is made amenable to computations. A Bayesian network simulated from data can be used to analyze and examine the distant relationships between variables. And also, perform the processes of prediction and explanation by calculating the conditional probability distribution of a variable, provided the values of some others calculating.

The origins of Bayesian networks can be traced back as far as the early decades of the 20th century, when Sewell Wright developed path analysis to aid the study of genetic inheritance (Witten & Frandk, 2005), (Wright, 1923). In their current form, Bayesian networks were introduced in the early 80s as a knowledge representation formalism to encode and use the information acquired from human experts in automated reasoning systems to perform diagnostic, predictive, and explanatory tasks (Pearl, 1988), Charniak, 1991). Their intuitive graphical nature and their principled probabilistic foundations were very attractive features to acquire and represent information burdened by uncertainty. The development of amenable algorithms to propagate probabilistic information through the graph (Pearl, 1988), (Thomas & Spiegelhalter & Gilks, 1992) put Bayesian networks at the forefront of Artificial Intelligence research.

Around same time, the machine learning community came to the realization that the sound probabilistic nature of Bayesian networks provided straightforward ways to learn them from data. As Bayesian networks encode assumptions of conditional independence, the first machine learning approaches to Bayesian networks consisted of searching for conditional independence structures in the data and encoding them as a Bayesian network (Pearl, 1988), (Glymour & Scheines & Spirtes & Kelly, 1987).

Shortly thereafter, Cooper and Herskovitz introduced a Bayesian method, further refined by Heckerman to learn Bayesian networks from data. These results spurred the interest of the Data Mining and knowledge discovery community in the unique features of Bayesian networks (Geiger & Heckerman, 1997).

A network describing the impact of two variables (nodes A and B) on a third one (node C). Each node in the network is associated with a probability table that describes the conditional distribution of the node, given its parents in Figure 5.8 below.



A	
True	**False**
0.1	0.9

B	
True	**False**
0.2	0.8

		C	
A	**B**	**True**	**False**
False	**False**	0.3	0.7
False	**True**	0.6	0.4
True	**False**	0.7	0.3
True	**True**	0.9	0.1

Figure 5.8 A Simple Bayesian network -1
Source: Maimon & Rokach, 2005

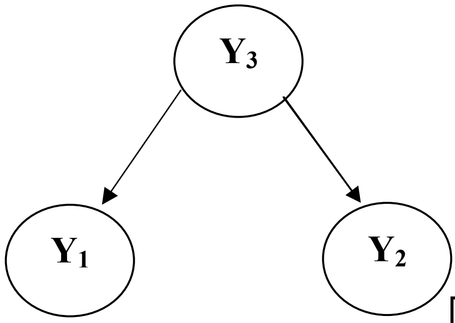
A Bayesian network has two components: a directed acyclic graph and a probability distribution. Nodes in the directed acyclic graph represent stochastic variables and arcs represent directed dependencies among variables that are quantified by conditional probability distributions (Maimon & Rokach, 2005).

A network encoding the conditional independence of YI, Y2 given the common parent Y3. The panel in the middle shows that the distribution of Y2 changes with Yl and hence the two variables are conditionally dependent in Figure 5.9 below.

A network encoding the conditional independence of YI, Y2 given the common parent Y3. The panel in the middle shows that the distribution of Y2 changes with Yl and hence the two variables are conditionally dependent (Maimon & Rokach, 2005).

Conversely, two variables that are marginally dependent may be made conditionally independent by introducing a third variable. This situation is represented by the directed acyclic graph in Figure 5.9 which shows two children nodes (YI and Y2) with a common parent Y3. In this case, the two children nodes are independent, given the common parent, but they may become dependent when we marginalize the common parent out (Maimon & Rokach, 2005).

Y_3	
True	**False**
0.1	0.9



	Y_3	
Y_1	**True**	**False**
False	0.0	1.0
True	0.5	0.5

	Y_2	
Y_3	**True**	**False**
False	0.0	1.0
True	0.5	0.5

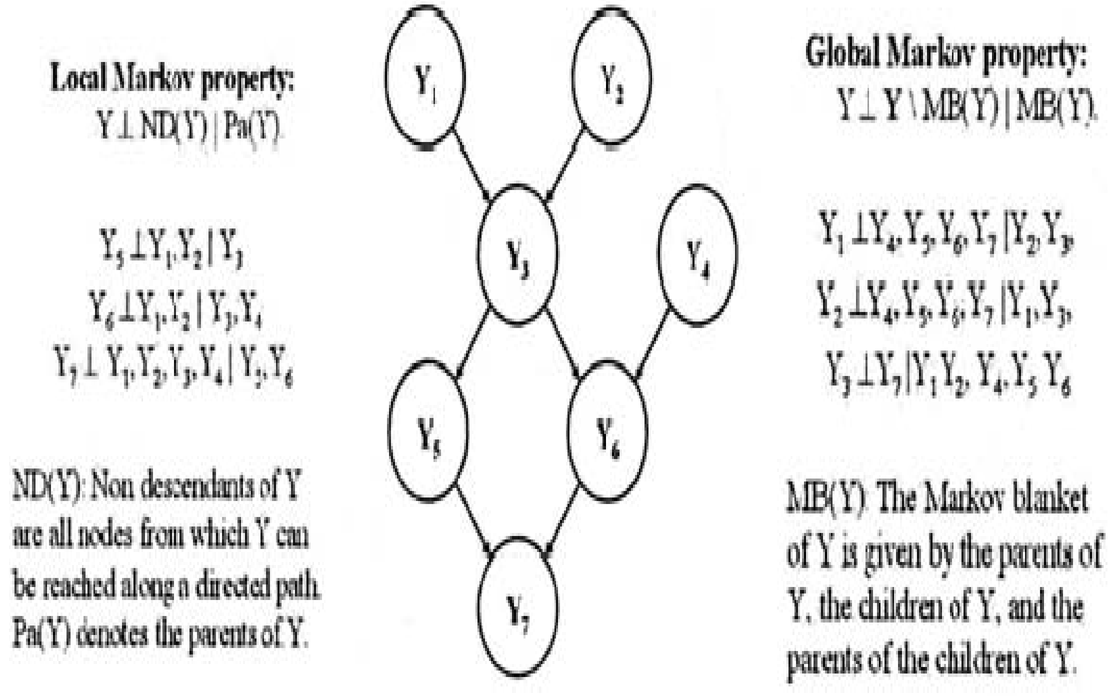
p(Y2 =True I Y, =True) = 0.5

p(Y2 =True I Y, =False) = 0.026

p(Y2 =True) = 0.05

Figure 5.9 A Simple Bayesian network -2
Source: Maimon & Rokach, 2005

In the Figure 5.10 A Bayesian network with seven variables and some of the Markov properties are represented by its directed acyclic graph below. The panel on the left describes the local Markov property encoded by acyclic graph and lists the three Markov properties that are represented by the graph in the middle. The panel on the right describes the global Markov property and lists three of the seven global Markov properties represented by the graph in the middle. The vector in bold denotes the set of variables represented by the nodes in the graph.



Local Markov property:
$Y \perp ND(Y) \mid Pa(Y)$

$Y_5 \perp Y_1, Y_2 \mid Y_3$
$Y_6 \perp Y_1, Y_2 \mid Y_3, Y_4$
$Y_7 \perp Y_1, Y_2, Y_3, Y_4 \mid Y_5, Y_6$

ND(Y): Non descendants of Y are all nodes from which Y can be reached along a directed path. Pa(Y) denotes the parents of Y.

Global Markov property:
$Y \perp Y \setminus MB(Y) \mid MB(Y)$.

$Y_1 \perp Y_4, Y_5, Y_6, Y_7 \mid Y_2, Y_3,$
$Y_2 \perp Y_4, Y_5, Y_6, Y_7 \mid Y_1, Y_3,$
$Y_3 \perp Y_7 \mid Y_1 Y_2, Y_4, Y_5 Y_6$

MB(Y). The Markov blanket of Y is given by the parents of Y, the children of Y, and the parents of the children of Y.

Figure 5.10 A Bayesian network with seven variables

The naïve Bayesian classifier is based on Bayes theory and assumes that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. Bayesian classifiers are useful in that they provide a theoretical justification for other classifiers that they do not explicitly use Bayes theorem. The major weakness of NB is that the predictive accuracy is highly correlated with the assumption of class conditional independence. This assumption simplifies computation. In practice, dependencies can exist between variables. Please see the structure of the NB below (Yeh & Lien, 2007).

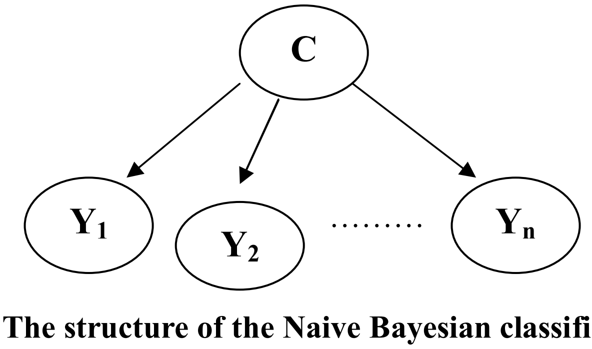


Figure 5.11 The structure of the Naive Bayesian classifier
Source: Yeh & Lien, 2007

6. DATA MINING APPLICATION IN BANKING ENVIRONMENT

6.1 GENERAL INFORMATION ABOUT BANKING IN TURKEY

As of March 2011, there are 44 banks are exist and operating 9581 branches in Turkey according to the database of Banks Association of Turkey (TBB – Türkiye Bankalar Birliği). The registered banks are listed below;

Commercial Banks

i. ***Banks Under the Deposit Insurance Fund:***
Birlesik Fon Bankasi A.S.

ii. ***Foreign Banks Founded in Turkey***
Arap Turk Bankasi A.S., Citibank A.S., Denizbank A.S.,Deutsche Bank A.S., Eurobank Tekfen A.S., Finans Bank A.S., Fortis Bank A.S., HSBC Bank A.S., ING Bank A.S., Millenium Bank A.S.

iii. ***Foreign Banks Having Branches in Turkey***
The Royal Bank of Scotland N.V., Bank Mellat, Habib Bank Limited,
JPMorgan Chase Bank N.A.,Societe Generale (SA), WestLB AG

iv. ***Privately-owned Commercial Banks***
Adabank A.S., Akbank T.A.S., Alternatif Bank A.S., Anadolubank A.S.,
Sekerbank T.A.S., Tekstil Bankasi A.S., Turkish Bank A.S.,
Turkiye Garanti Bankasi A.S., Turkiye Is Bankasi A.S., Yapi ve Kredi Bankasi A.S.

v. ***State-owned Commercial Banks***
Turkiye Cumhuriyeti Ziraat Bankasi A.S., Turkiye Halk Bankasi A.S.,
Turkiye Vakiflar Bankasi T.A.O.

Non-depository Banks

i. ***Foreign Non-depository Banks***
BankPozitif Kredi ve Kalkinma Bankasi A.S. Credit Agricole Yatirim Bank Turk A.S, Merrill Lynch Yatirim Bank A.S., Taib Yatirim Bank A.S.

ii. ***Privately-owned Non-depository Banks***

Aktif Yatirim Bankasi A.S., Diler Yatirim Bankasi A.S.,GSD Yatirim Bankasi A.S.,IMKB Takas ve Saklama Bankasi A.S. , Nurol Yatirim Bankasi A.S., Turkiye Sinai Kalkinma Bankasi A.S.

i. ***State-owned Non-depository Banks***

Iller Bankasi, Turk Eximbank, Turkiye Kalkinma Bankasi A.S.

By the end of 1999, the restructuring process in the banking system of Turkey was started with the disinflation works and continued in 2001 and the following year. The Banking Regulation and Supervision Authority (BRSA) were established as a regulatory and financial authority with administrational and financial autonomy in banking sector. Duties and authorities regarding the supervision and regulation of banks which were previously shared by the Treasury and the Central Bank in the past were transferred to BRSA which started its operations in August 2000.

One of the main objectives of the restructuring in the banking system has been issued the legal and institutional regulations for improvement of audit systems, changing the risk taking and risk-management processes and methods and enhancement of the corporate infrastructure.

All these regulations on banking law, enormous data collection, evaluation and reporting systems such as CRM, legal reporting have become very important issue for whole banks. And they have started to build CRM systems within the banks.

6.2 DATA MINING IN BANKING ENVIRONMENT

A group of customer data has been provided from a bank and applied data mining process by using TANAGRA application. Data mining process in this study consist of Business understanding, data understanding, data preparation, modeling, and evaluation deployment phases. And all the phases of data mining are applied into the selected data in this study.

6.2.1 Business Understanding

As it was expressed in the previous part, because of restructuring of banking environment in Turkey, all the banks have to regulate their organization, business flow and associated banking software applications accordingly. Thus, they have made significant investment on their system in order to follow new banking rules, such as managing customer risks and any kind of legal reporting by implementing CRM and reporting systems.

The bank is foreign capital bank in Turkey and has a small volume in the banking sector. They have a quite limited number of branches and limited customers. The activation of whole customers is also quite low. The main banking products produced for the customers are; loan products, time deposits, demand deposits with overdraft limits, credit cards and debit cards. The banking facilities have been continued thru limited number of branches, internet banking, call center and ATMs.

The main purpose of the bank is to improve the probability, effectiveness and decrease the total cost of the bank. Customer profile analysis as financial and nonfinancial point of view provides elegant support for meeting goals of the bank. Also retention of the customers and acquiring the new ones are one of the main activities for improvement on probability.

The data mining problem is to analyze pattern of customer profile as demographic and financial point of view. Therefore, the segmentation of customer profile and financial risks over the selected customer data will be defined. And it will be produced suggestions in order to increase active customers, their probability and decrease the risks according to result of customer profile analysis. The following analysis issues will be tried to perform within the scope of study.

i. Customer profile analysis with the main features of the training data
ii. Customer risk analysis by clustering demographic information of the customers
iii. Customer risk analysis by clustering financial information of the customers

6.2.2 Data Analysis

The data for customer analysis is provided from a foreign capital bank in Turkey. In fact the dataset consist of two databases as customer demographic information and customer risk data. These datasets are joined by customer number and transferred in one database. As totally, there are 1027 customer records selected to be the topic for this study. The final data has been analyzed according to the major features of the database which are gender data, marital status, age in bank, customer types, city codes, risk codes and finance codes.

The 25 percent of the selected customers is female, the 59 percent is male, the 17 percent is unknown

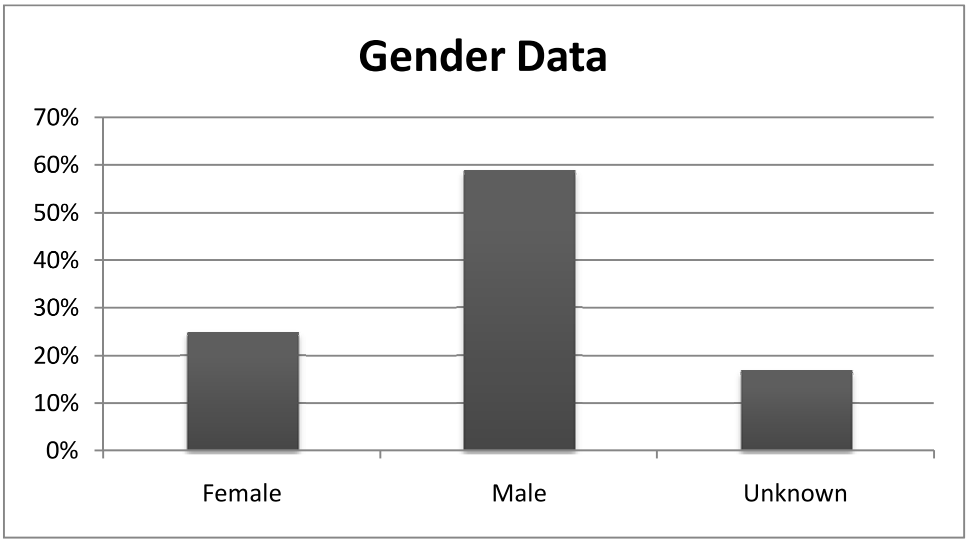


Figure 6.1 Histogram for gender data

The 82 percent of the selected customers is married, the 18 percent is single

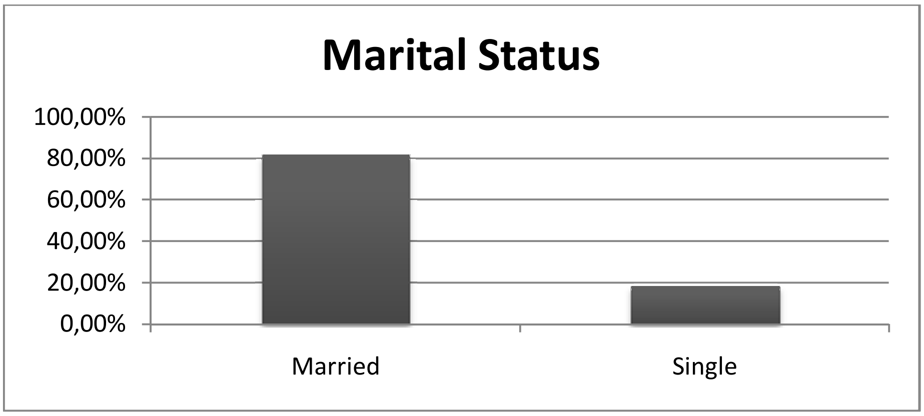


Figure 6.2 Histogram for marital status

Age in bank represents the customer age in the bank means that how many years the person has been as a customer. The 28,30 percent of the customers is 6 years, the 53,60 percent is 5 years, the 13 percent is 4 years, the 5 percent is 3 years, the 0.2 percent is 2 years.

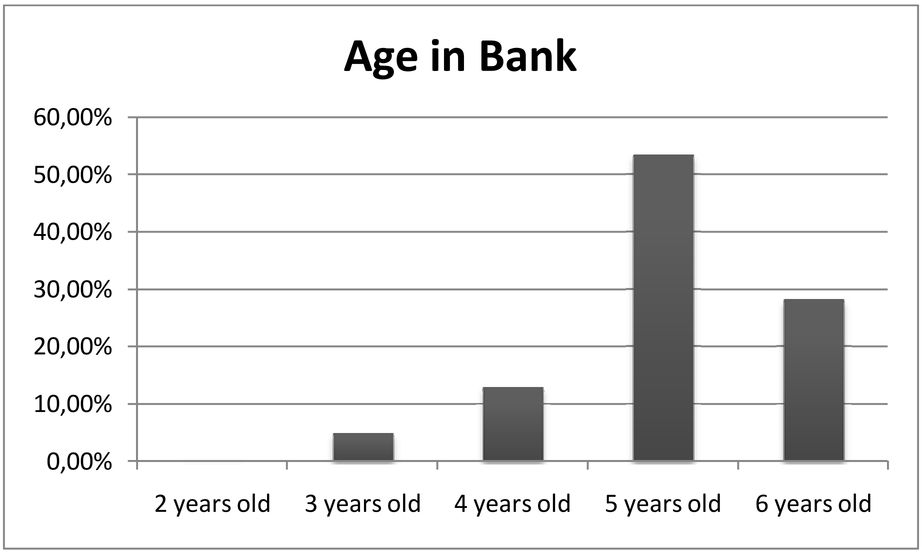


Figure 6.3 Histogram for age in bank

Customer types contain 2 values; retail customer and corporate customer, all the financial structure in the bank are built based on this feature such as loan, time, overdraft processes, interest ratios and the calculation methods. The 85 percent is retail, the 15 percent is corporate customer.

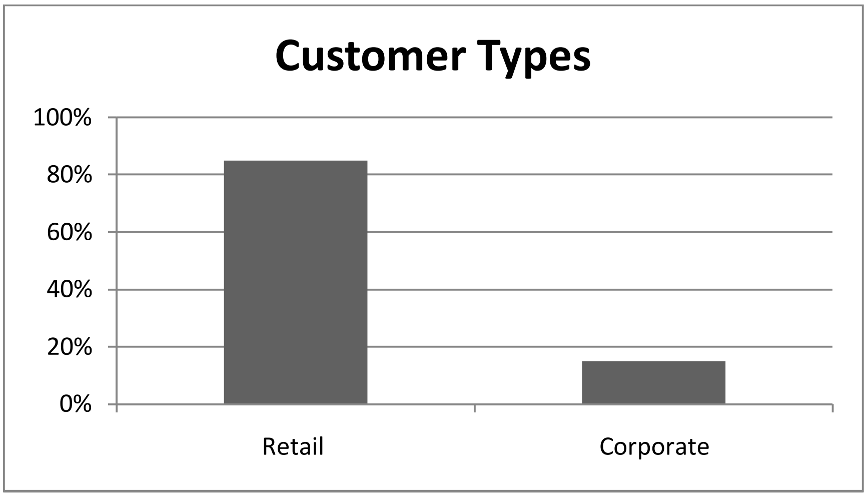


Figure 6.4 Histogram for customer types

City of customers represent the city code of the associated branch of the customers located. It displays city of current branches. The 16,1 percent of the customer from Ankara, the 0,5 percent is from Antalya, the 5 percent from Bursa, the 71,9 percent from İstanbul, the 6,6 percent from İzmir.

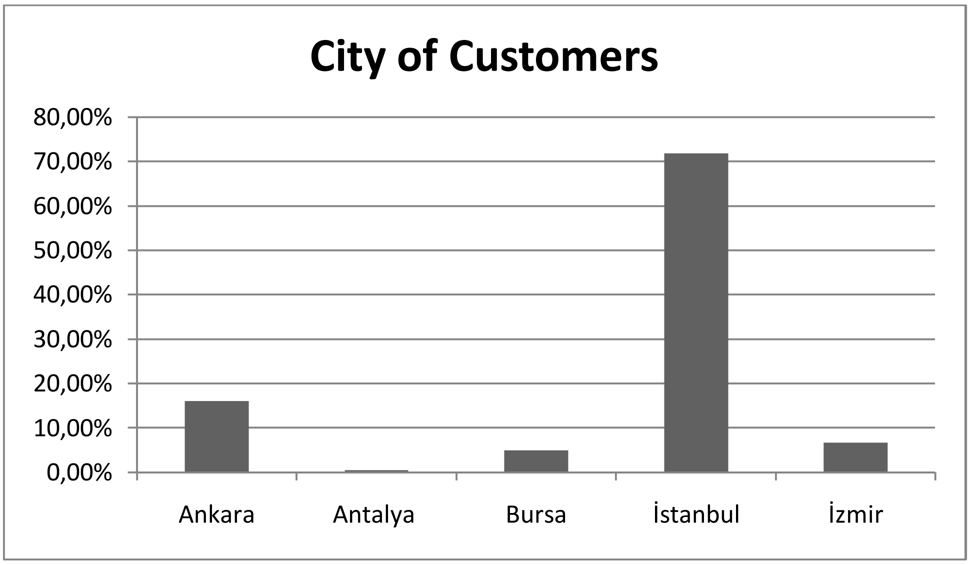


Figure 6.5 Histogram for city of customers

Finance code represents the profession area of customer as retail or corporate in other words; displays working area of the customer, which industry is working for. The value list of finance codes are below;

Table 6.1 Finance code list

Finance code	**Finance code definition**
1	Agriculture, hunting and forestry
2	Mining and Quarrying
3	Manufacturing Industry
4	Electricity gas and water supply
5	Construction -1
6	Construction -2
7	Wholesale and retail trade, motor vehicle services -1
8	Wholesale and retail trade, motor vehicle services -2
9	Wholesale and retail trade, motor vehicle services -3
10	Tourism
11	Maritime Transport, Air Transport, Land Transportation
12	Real estate commission, renting and business activities
13	Education
14	Health and social service
15	Other community, social and personal services

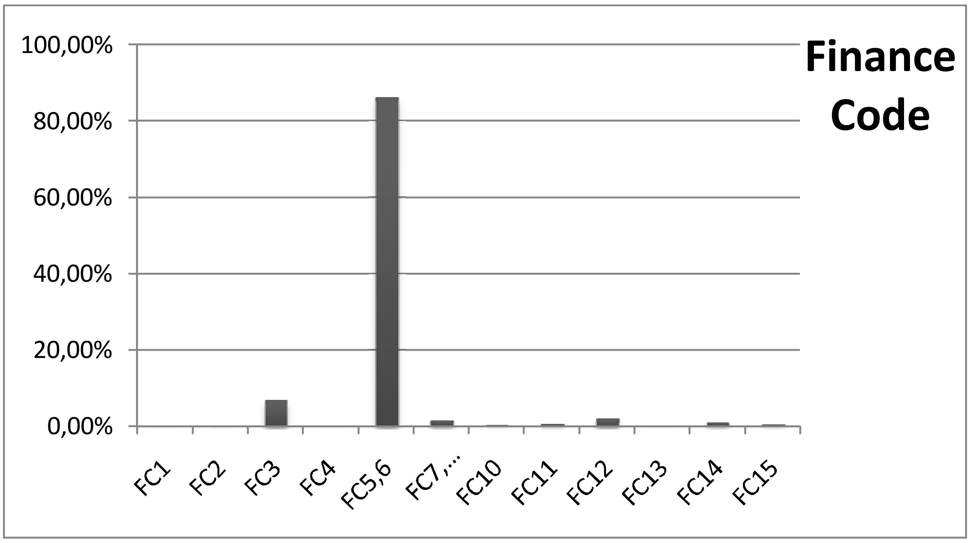


Figure 6.6 Histogram for finance codes

The 0,1 percent of the customers work in Agriculture, hunting and forestry

The 0,19 percent of the customers work in Mining and Quarrying

The 6,82 percent of the customers work in Manufacturing Industry

The 0, 10 percent of the customers work in Electricity gas and water supply

The 86, 37 percent of the customers work in Construction sector

The 1, 56 percent of the customers work in Wholesale and retail trade, motor vehicle services

The 0, 39 percent of the customers work in Tourism

The 0, 68 percent of the customers work in Maritime Transport, Air Transport, Land Transportation

The 2, 04 percent of the customers work in Real estate commission, renting and business activities

The 0, 10 percent of the customers work in Education

The 1, 07 percent of the customers work in Health and social service

The 0, 58 percent of the customers work in other community, social and personal services

Risk code represents the services that are purchased by the customers. Such as loans, time deposits, overdraft accounts…etc. The value list of risk codes are below;

Table 6.2 Risk code list

Risk code	Risk code definition
100	Loan – TL
101	Loan - FX
120	Retail Loan - TL
121	Retail Loan – FX
150	Cash Loan – FX
200	Credit of Guarantee - TL
250	Credit of Guarantee – FX
300	Accounts receivable for liquidation
301	Doubtful Fees, Commissions and Other Account Receivables – TL
302	Uncollectible Loans and Other Account Receivables – TL

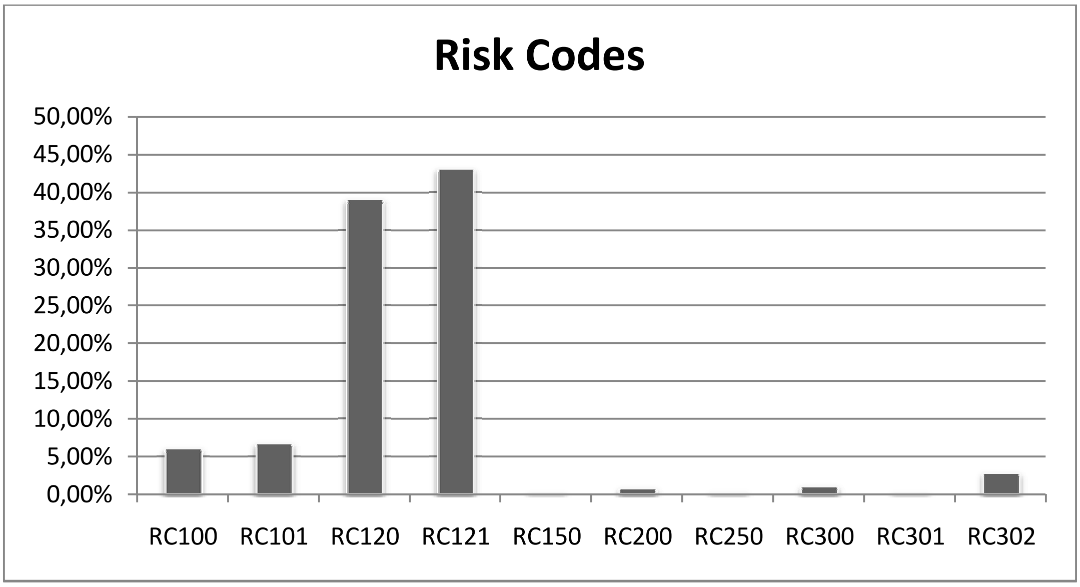


Figure 6.7 Histogram for risk codes

The 6, 04 percents of the customers have Loan in TL

The 6, 72 percents of the customers have Loan in Foreign Currency

The 39, 05 percents of the customers have Retail Loan in TL

The 43, 14 percents of the customers have Retail Loan in Foreign Currency

The 0, 1 percent of the customers has Cash Loan in Foreign Currency
The 0, 78 percents of the customers have Credit of Guarantee in TL
The 0, 19 percents of the customers have Credit of Guarantee in Foreign Currency
The 1, 07 percents of the customers have Accounts receivable for liquidation
The 0, 1 percent of the customers has Doubtful Fees, Commissions and Other Account Receivables in TL
The 2, 82 percents of the customers have Uncollectible Loans and Other Account Receivables in TL

6.2.3 Data Preparation

After analyzing the raw data, it is performed some data cleaning and preparation activities over it. As it was already mentioned above, the raw data consist of demographic and financial information and joined in one table. After that, some fields are deleted, some are renamed, some are revalued, date format are first converted from Julian date to calendar data and then recalculated as year. Approximately 140 fields have been revised and as a result final dataset have been generated in order to be processed in modeling & evaluation phase.

The field features of raw data, data, and revaluing, renaming facilities are displayed in the following table;

Table 6.3 Customer demographic information

Field No	Field Name	Selected (Y/N)	Data Preparation Activity
1	Bank Number	Y	Deleted
2	Customer Number	Y	OK
3	Record Status	Y	Only active records are selected, after that it is deleted
4	Alternate Address Option	Y	Deleted
5	Name Address Line 1	Y	Deleted
6	Name Address Line 2	Y	Deleted
7	Name Address Line 3	Y	Deleted
8	Name Address Line 4	Y	Deleted
9	Name Address Line 5	Y	Deleted
10	Name Address Line 6	Y	Deleted
11	ZIP Code - First Five Digit	N	
12	ZIP Code Suffix	N	
13	ZIP Code route number	N	
14	ZIP Code check digit	N	
15	Short Name	Y	Deleted
16	Social Security Number	Y	Deleted
17	Tax-ID Number Flag	Y	Deleted
18	Cellular Phone Number	Y	Reassigned as 1(exist) and 0 (not exist)
19	Home Phone Number	Y	Reassigned as 1(exist) and 0 (not exist)
20	Business Phone	Y	Reassigned as 1(exist) and 0 (not exist)
21	Primary Officer Number	Y	Revalued as 1(exist) and 0 (not exist)
22	User Field 3	N	
23	Officer 1	N	
24	Officer 2	N	
25	Customer Opening Date	Y	Converted from Julian date to calendar date. Then recalculated as year and renamed as Age in bank
26	Account Type	N	
27	Customer Type	Y	1 (Personal), 0 (Non Personal)
28	SIC Code	N	
29	Sex / Gender Data	Y	Reassigned as 0 (Female), 1 (Male), 2 (unknown)

Table 6.3 Customer demographic information (continued)

30	Race	N	
31	Own Home	Y	Reassigned as 1(Yes), 0(No), 2(not informed)
32	Year Employed At Current Job	Y	Deleted
33	Income In Multiple (ex. 000	N	
34	Primary Source Of Income	N	
35	Date Of Birth	Y	Converted from Julian date to calendar date. Then recalculated in year and renamed as customer age
36	Number Of Dependents	Y	Deleted
37	Contact Person (Business)	N	
38	Contact Title	N	
39	Withholding Code	N	
40	Highest Used Memo Nbr	N	
41	National ID Number	Y	Reassigned as 1(exist) and 0 (not exist)
42	User Key Field	N	
43	Customer Linkage Field	N	
44	User Field 15	N	
45	Diners Club Holder Flag	N	
46	Diners Club Card Number	N	
47	Diners Club Expiration Date	N	
48	Mastercard Holder Flag	N	
49	Mastercard Number	N	
50	MC Expiration Date	N	
51	VISA Card Holder Flag	N	
52	VISA Card Number	N	
53	VISA Expiration Date	N	
54	ATM Card Holder Flag	N	
55	ATM Card Number	N	
56	ATM Card Expiration Date	N	
57	Language Code	N	

Table 6.3 Customer demographic information (continued)

58	Citizenship Code	Y	Reassigned as 1(Turkish), 0 (Not exist), 2(other nationalities)
59	Legal Residence Code	N	
60	Witholding Percentage	N	
61	Passport Number	Y	Reassigned as 1(exist) and 0 (not exist)
62	Tax-ID Number	Y	Reassigned as 1(exist) and 0 (not exist)
63	Profession Code	Y	Deleted
64	Short Name Derived	N	
65	Intl Dial Code	N	
66	Postal Code	Y	Deleted
67	Accommodation Code	Y	Deleted
68	Branch Number	Y	OK
69	Moved In Date(Cal)	N	
70	Marital Status	Y	Reassigned as 1(exist) and 0 (not exist)
71	Mail Indicator	Y	Deleted
72	Solicitable Code	N	
73	Socio-Economic Group	N	
74	Home Phone Avail	Y	Reassigned as 1(exist) and 0 (not exist)
75	Bus Phone Avail	Y	Reassigned as 1(exist) and 0 (not exist)
76	Personal/NonPersonal	Y	Reassigned as 1(exist) and 0 (not exist)
77	Salutation	N	
78	Facsimile Number	N	
79	Telex Number	N	
80	Telex Answerback	N	
81	Cust DOC Flag	N	
82	Cust DOC Activity Date	N	
83	TIN Cert Flag	N	
84	TIN Activity Date	N	
85	Nbr Certs Sent	N	
86	Customer Document Weight	N	
87	Cash Exclusion Code	N	
88	Cash Exclusion Limit	N	
89	Customer Extract Flag	N	
90	Date Last Maintained	Y	Deleted
91	Curr Country Code	Y	Deleted

Table 6.3 Customer demographic information (continued)

92	Market Seqment	N	
93	Employee Code	N	
94	Inquiry Level	N	
95	Maintenance Level	N	
96	Location Code	N	
97	Last Contact Date(Julian)	N	
98	Date Of Death-CAL	N	
99	Access Code	N	
100	First B-Notice Year	N	
101	Second B-notice Year	N	
102	Cust/Dlr/Dlr Group Flag	N	
103	Preferred Customer Code	N	
104	Source Of Data	N	
105	Customer Classification	N	
106	Level Of Activity	N	
107	Date Moved In Confirmed	N	
108	Telephone Extension	N	
109	Business Phone Ext	N	
110	Address Type	N	
111	Behavioural Type Flag	N	
112	Life Stage Flag	N	
113	Responsiveness Flag	N	
114	Home E-Mail Address	Y	Reassigned as 1(exist) and 0 (not exist)
115	Business E-Mail Address	Y	Reassigned as 1(exist) and 0 (not exist)
116	Batch Release Flag	N	
117	Expected Date	N	
118	Expected Frequency	N	
119	External Originator ID	N	
120	Expected Period	N	
121	Maximum Split Frequency	N	
122	Alias Name	N	
123	Employment Status	N	
124	Multi-Brand Code	N	

Table 6.4 Customer financial information

Field No	Field Name	Selected (Y/N)	Data Preparation Activity
1	Bank Code	N	
2	City Code	Y	OK.
3	Branch Range	Y	OK. Valued as 1,2,3
4	Tax/Citizienship No	Y	Deleted
5	Customer Name	Y	Deleted
6	Financial Code	Y	Reassigned as numeric values starting from 1….. to 15 in financial codes table
7	Risk Code	Y	OK. In the risk code table
8	Limit of Loan	Y	OK
9	Rediscounted Interest Fee, Comm.	Y	Deleted
10	Accured Interest Fee, Comm.	Y	Deleted
11	Branch Code	Y	Deleted
12	Update User	Y	Deleted
13	Cif No	Y	Deleted

The data in Table 6.1 and the data in Table 6.2 are joined by matching customer number. The final set is generated by the fields marked as selected "Y" above.

6.2.4 Modeling & Evaluation and Deployment

After generating final data set, clustering methods are decided to be applied in to it according to meet the goals explained in business understanding section previously. Three models are decided to be built;

i. Customer Profile Analysis in Clustering Tree
ii. Customer Risk Analysis with demographic information in K-Means Clustering
iii. Customer Risk Analysis with financial information in K-Means Clustering

6.2.4.1 Model 1 Customer Profile Analysis – Clustering Tree

It is aimed to define customer profile analysis of the training data according to the major variables exist in it. The target variable is selected as customer number and input variables are selected the following fields;

Age in Bank

Sex

Customer Age

Citizen Code

Marital Status

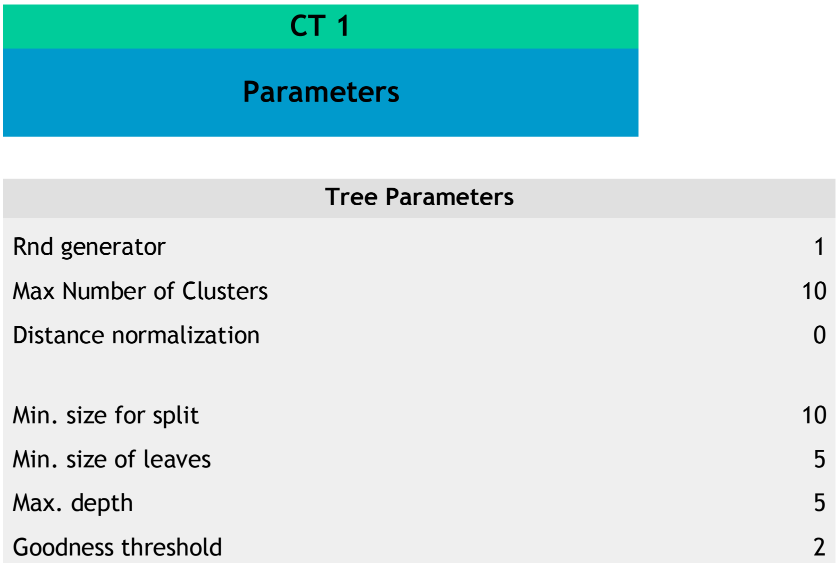
Customer Type

City Code

Finance Code

Clustering Tree method has been run on the training data with the selected parameters above in Tanagra application and the following results have been produced by Tanagra.

Table 6.5 Clustering tree parameters & results

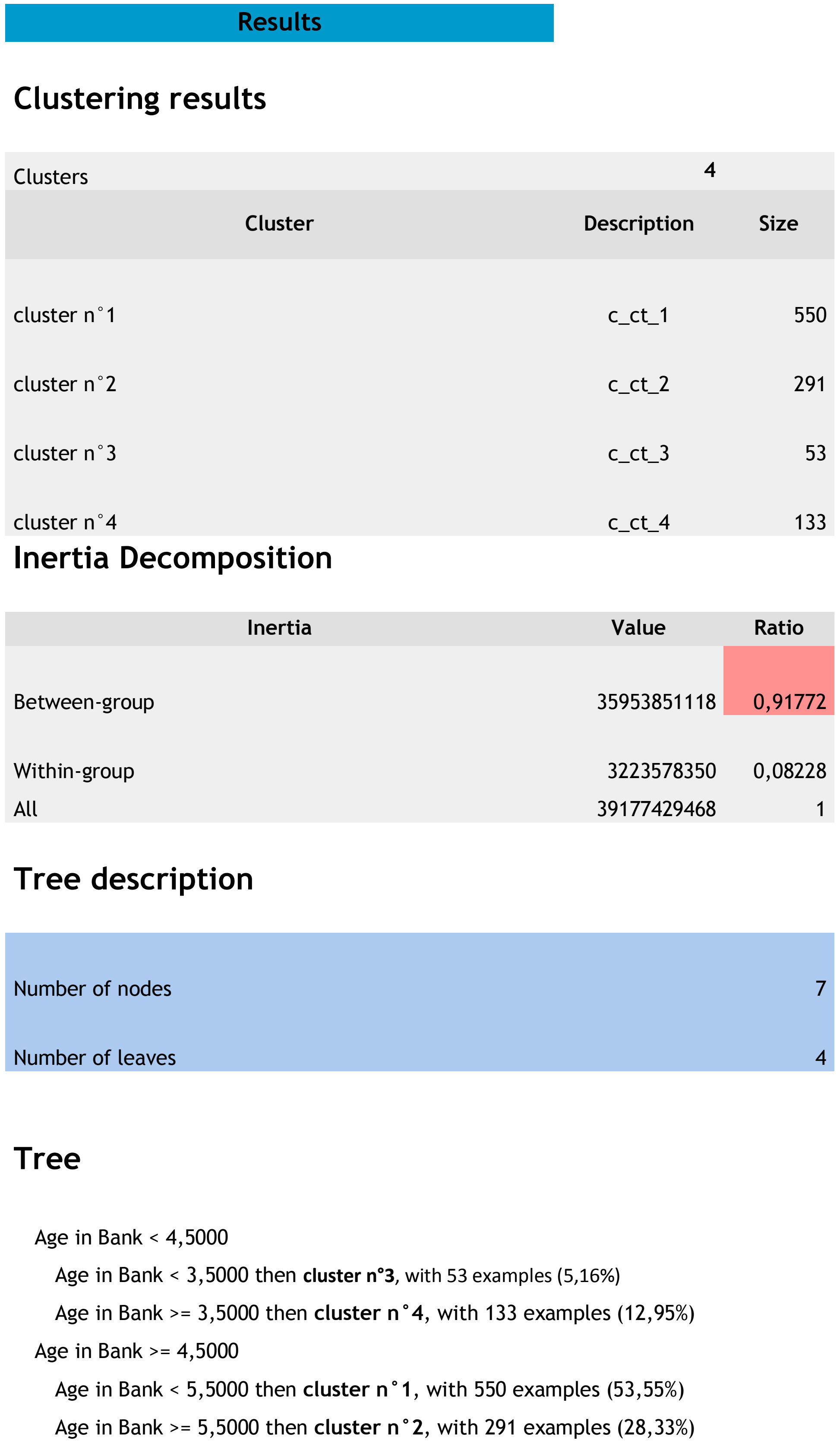


CT 1

Parameters

Tree Parameters	
Rnd generator	1
Max Number of Clusters	10
Distance normalization	0
Min. size for split	10
Min. size of leaves	5
Max. depth	5
Goodness threshold	2

Table 6.5 Clustering tree parameters & results (continued)



Results

Clustering results

Clusters	4	
Cluster	**Description**	**Size**
cluster n°1	c_ct_1	550
cluster n°2	c_ct_2	291
cluster n°3	c_ct_3	53
cluster n°4	c_ct_4	133

Inertia Decomposition

Inertia	Value	Ratio
Between-group	35953851118	0,91772
Within-group	3223578350	0,08228
All	39177429468	1

Tree description

Number of nodes	7
Number of leaves	4

Tree

- Age in Bank < 4,5000
  - Age in Bank < 3,5000 then **cluster n°3**, with 53 examples (5,16%)
  - Age in Bank >= 3,5000 then **cluster n°4**, with 133 examples (12,95%)
- Age in Bank >= 4,5000
  - Age in Bank < 5,5000 then **cluster n°1**, with 550 examples (53,55%)
  - Age in Bank >= 5,5000 then **cluster n°2**, with 291 examples (28,33%)

According to the results, four groups have been generated based on the values of "Age in bank" field.

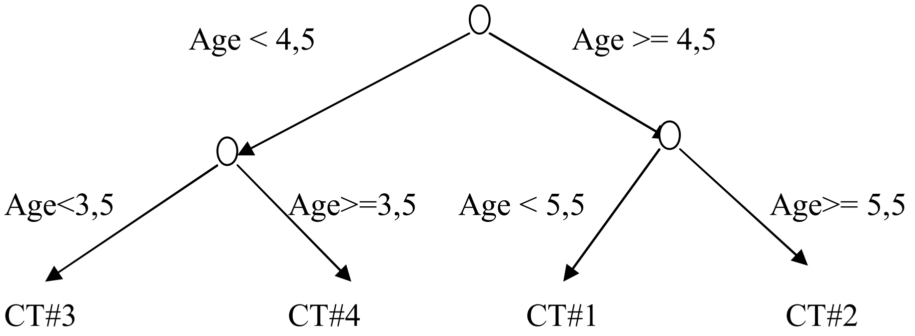


Figure 6.8 Clustering tree of age in bank

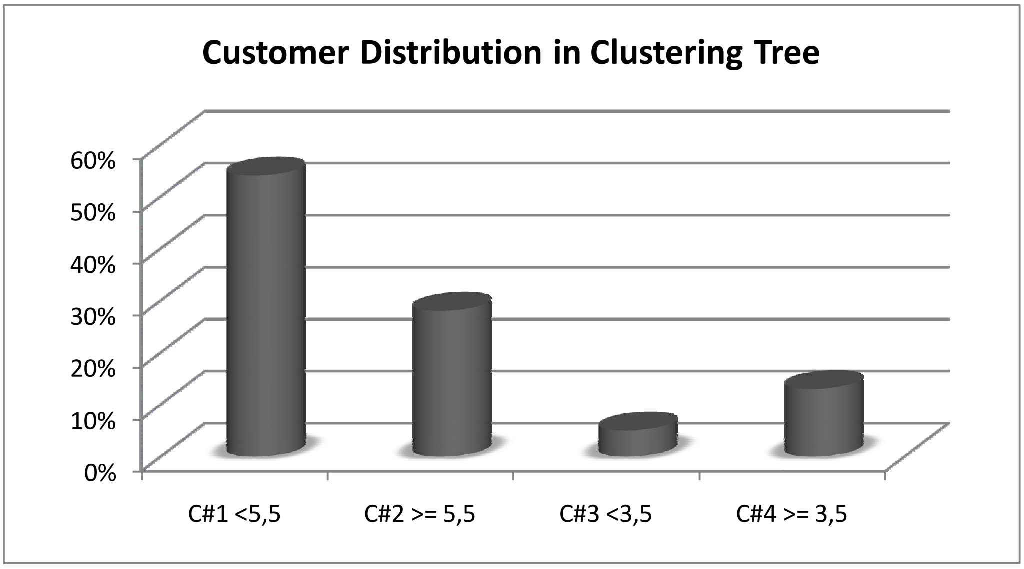


Figure 6.9 Customer distribution in clustering tree

There are 550 records in **Cluster#1** and the following profile analysis is observed;

i. The 60,18 percent is male, the 24,36 percent is female and the 15,46 percent of it has not got any sex information

ii. The 14,72 percent has not got age information, the 3,8 percent is less and equal to 30 years old, the 29,27 percent is between 31-40 years including them. The

44, 9 percents is between 41-60 years old, including them. The 7, 3 percents is over 60 years old.

iii. The 80,72 percent is married, the 19,28 percent is single

iv. The 85,27 percent is personal (retail) customer, the 14,72 percent is non personal (corporate) customer

v. The 68,90 percent is in İstanbul, the 31,09 percent is in out of İstanbul

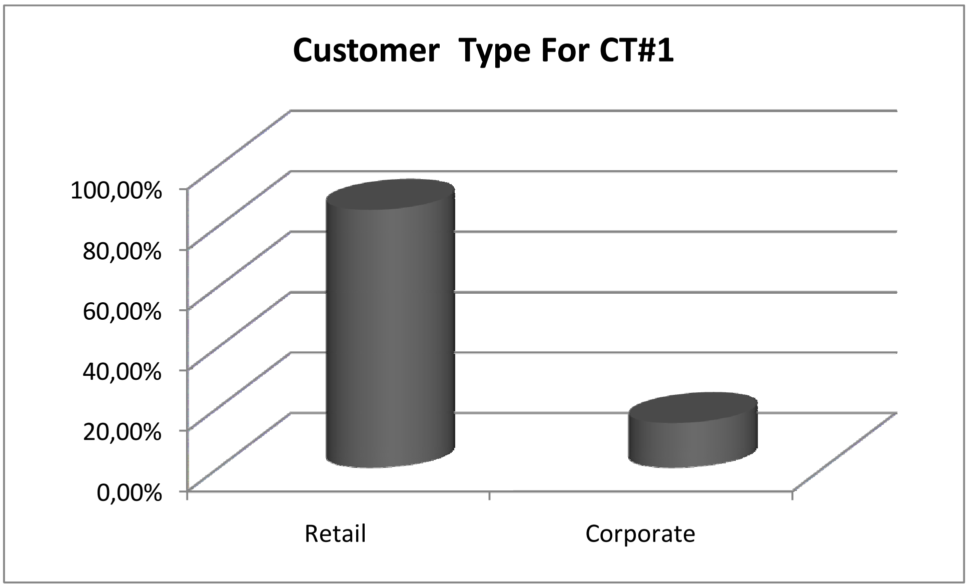


Figure 6.10 Customer type for CT#1

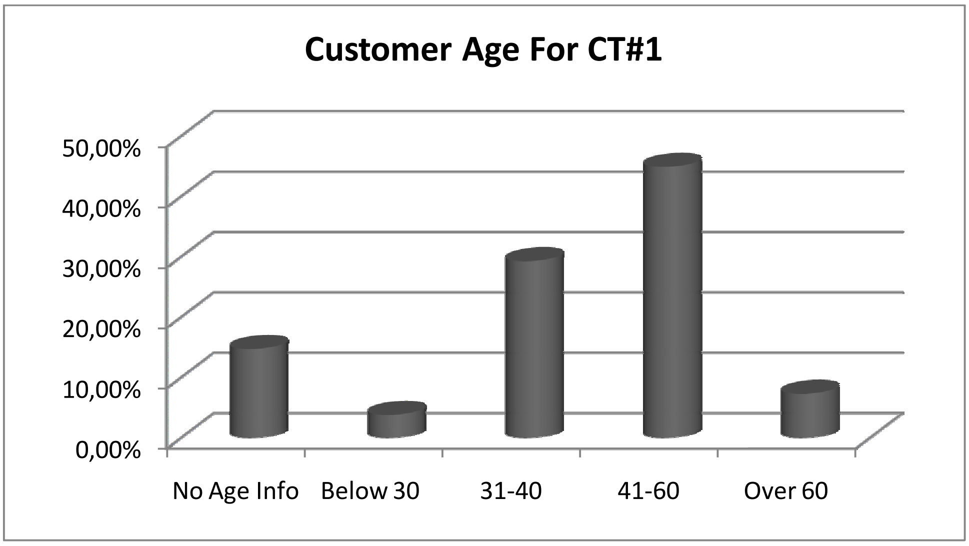


Figure 6.11 Customer age for CT#1

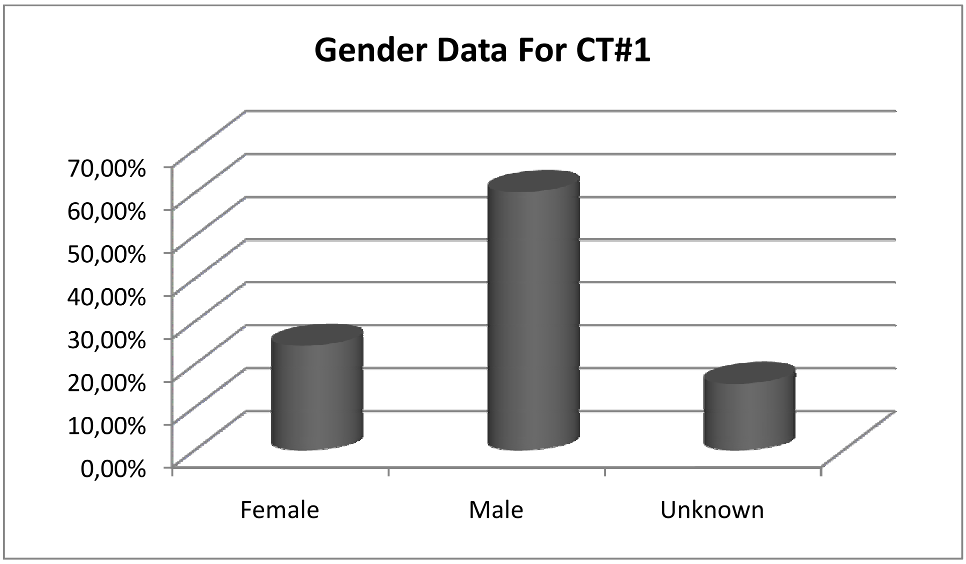


Figure 6.12 Gender data for CT#1

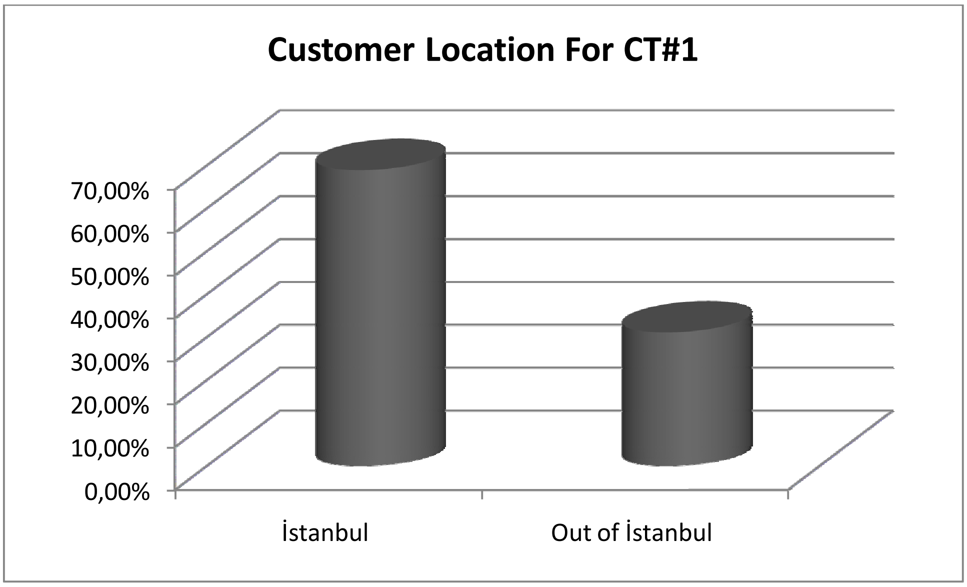


Figure 6.13 Customer age location for CT#1

There are 291 records in **Cluster#2** and the following profile analysis is observed;

i. The 67,01 percent is male, the 27,49 percent is female and the 5,5 percent of it has not got sex information

ii. The 4,47 percent has not got age information, the 2,4 percent is less and equal to 30 years old, the 36,42 percent is between 31-40 years including them, the 52, 23 percent is between 41-60 years old including them, the 4, 48 percent is over 60 years old

iii. The 80,76 percent is married, the 19,24 percent is single

iv. The 95,53 percent is personal (retail) customer, the 4,47 percent is non personal (corporate) customer

v. The 83,16 percent is in İstanbul, the 16,84 percent is in out of İstanbul

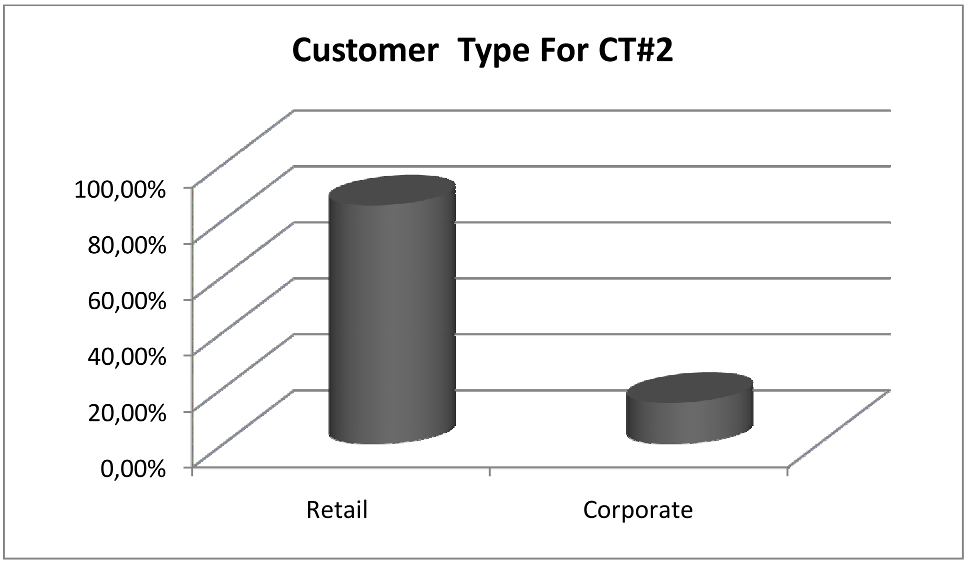


Figure 6.14 Customer type for CT#2

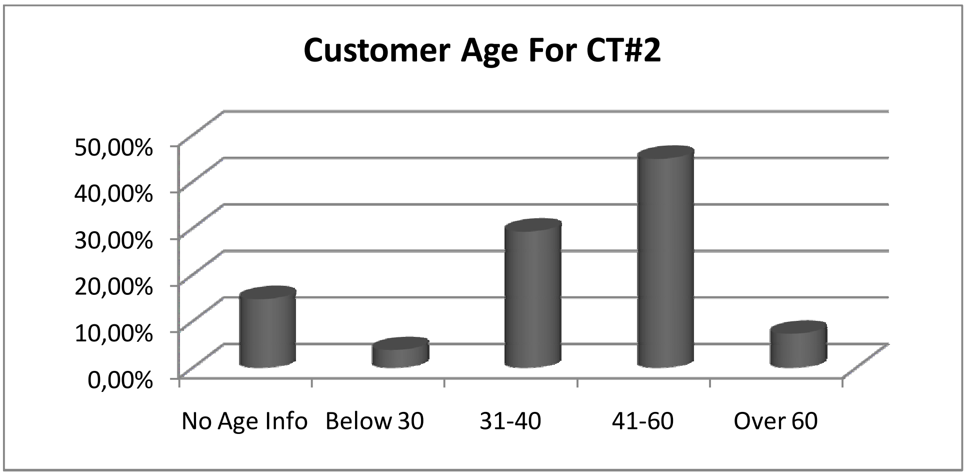


Figure 6.15 Customer age for CT#2

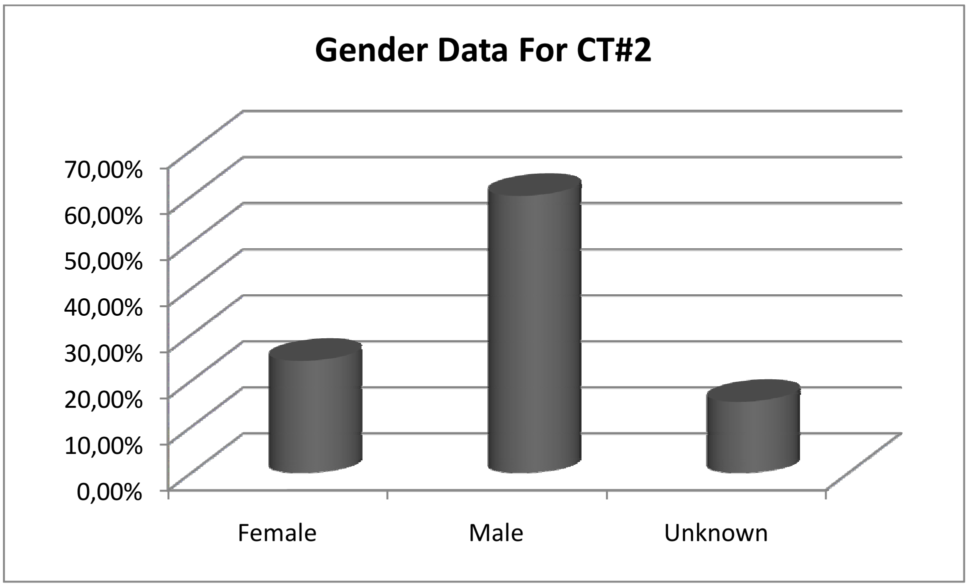


Figure 6.16 Gender data for CT#2

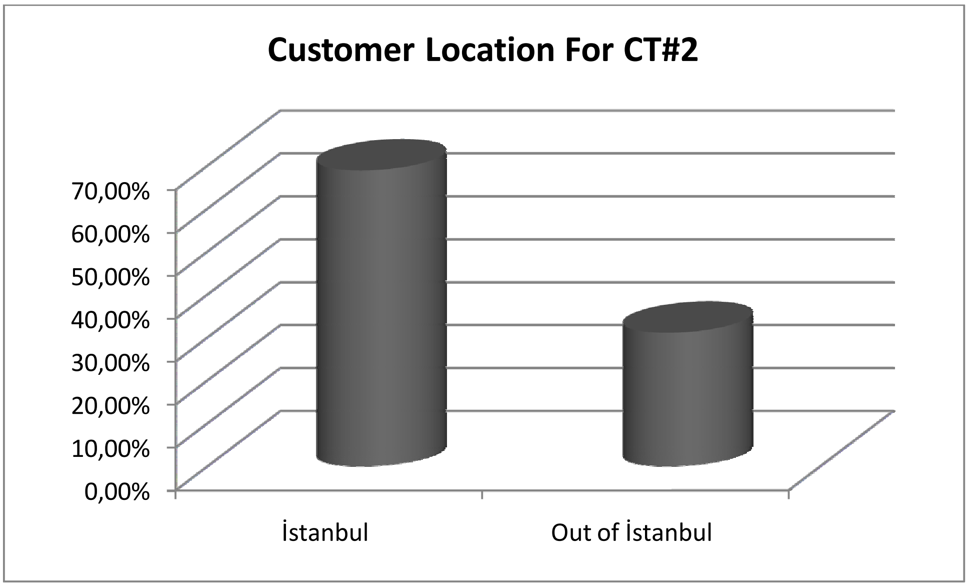


Figure 6.17 Customer locations for CT#2

There are 53 records in **Cluster#3** and the following profile analysis is observed;

i. The 20,75 percent is male, the 11,32 percent is female and the 67,93 percent of it has not got sex information

ii. The 67,92 percent has not got age information, the 1,89 percent is less and equal to 30 years old, the 9,43 percent is between 31-40 years including them, the

15,09 percent is between 41-60 years old including them, the 5,17 percent is over 60 years old

iii. The 80,76 percent is married, the 19,24 percent is single

iv. The 32,07 percent is personal (retail) customer, the 67,93 percent is non personal (corporate) customer

v. The 83,16 percent is in İstanbul, the 16,84 percent is in out of İstanbul

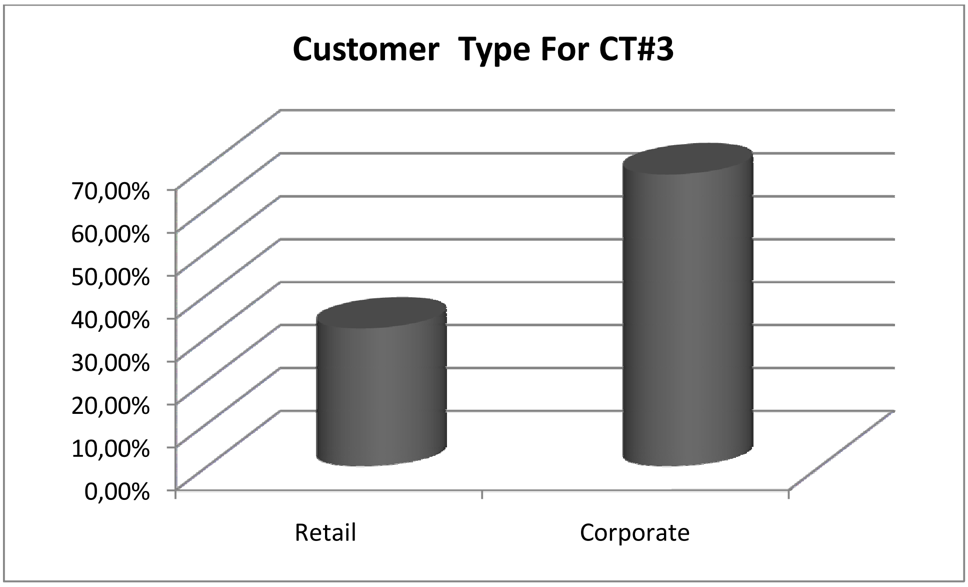


Figure 6.18 Customer type for CT#3

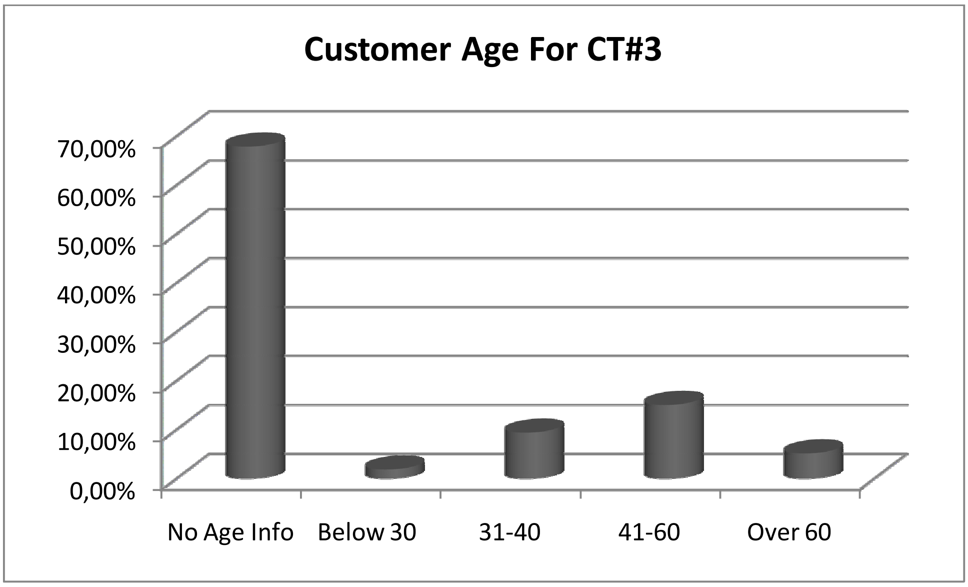


Figure 6.19 Customer age for CT#3

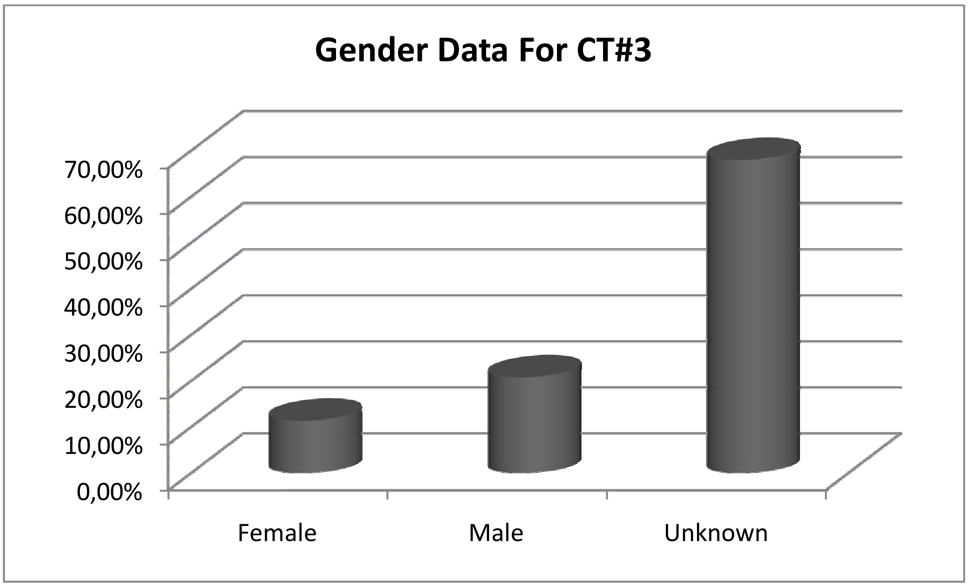


Figure 6.20 Gender data for CT#3

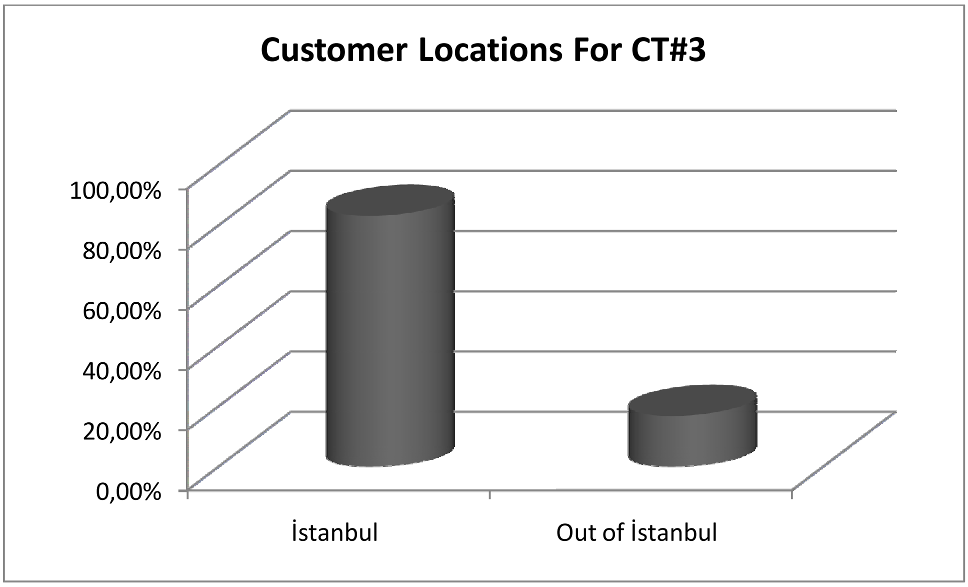


Figure 6.21 Customer locations for CT#3

There are 133 records in **Cluster#4** and the following profile analysis is observed;

i. The 51,13 percent is male, the 24,06 percent is female and the 24,81 percent of it has not got sex information

ii. The 20,30 percent has not got age information, the 3,76 percent is less and equal to 30 years old, the 27,82 percent is between 31-40 years including them, the 42,86 percent is between 41-60 years old including them, the 5,26 percent is over 60 years old

iii. The 81,95 percent is married, the 18,05 percent is single

iv. The 78,95 percent is personal (retail) customer, the 21,05 percent is non personal (corporate) customer

v. The 64,66 percent is in İstanbul, the 35,34 percent is in out of İstanbul

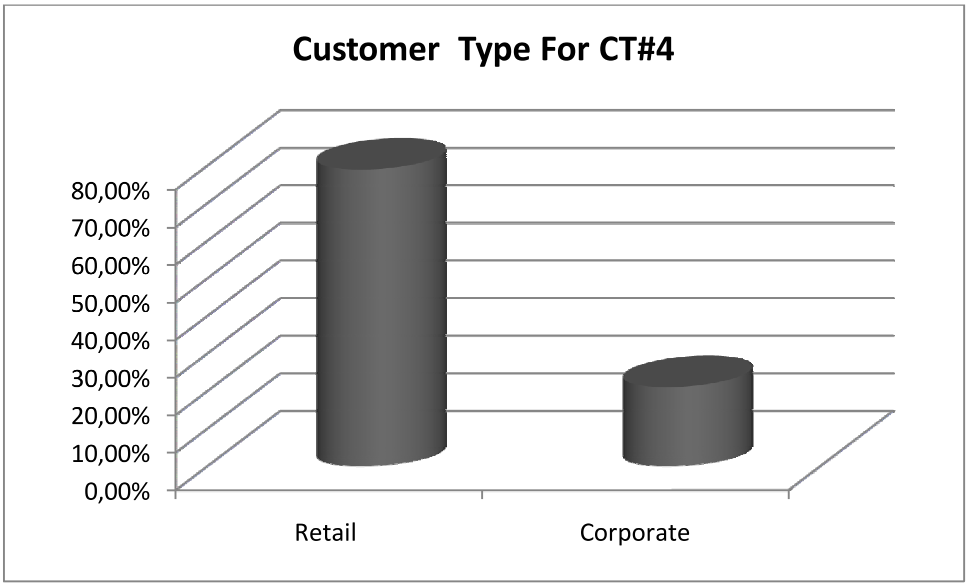


Figure 6.22 Customer type for CT#4

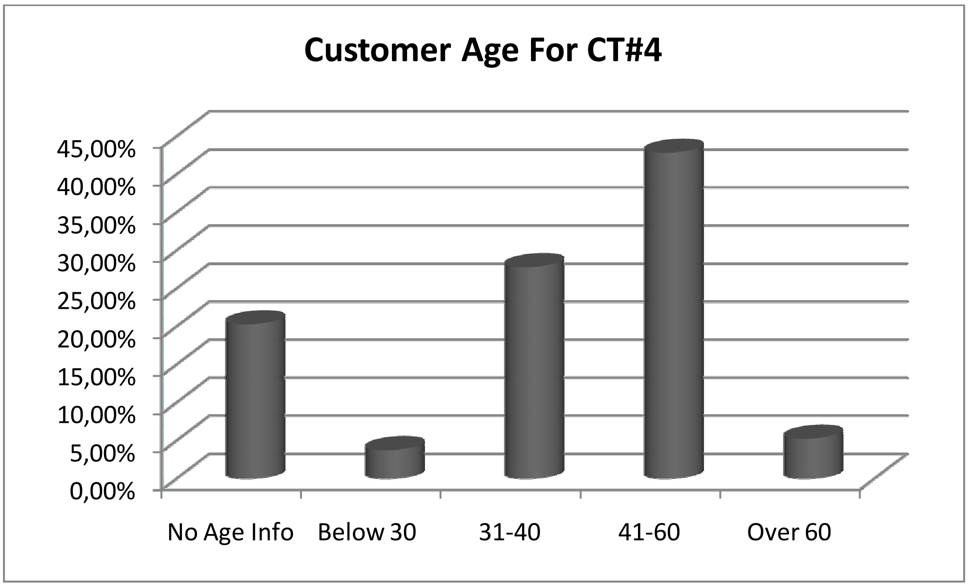


Figure 6.23 Customer age for CT#4

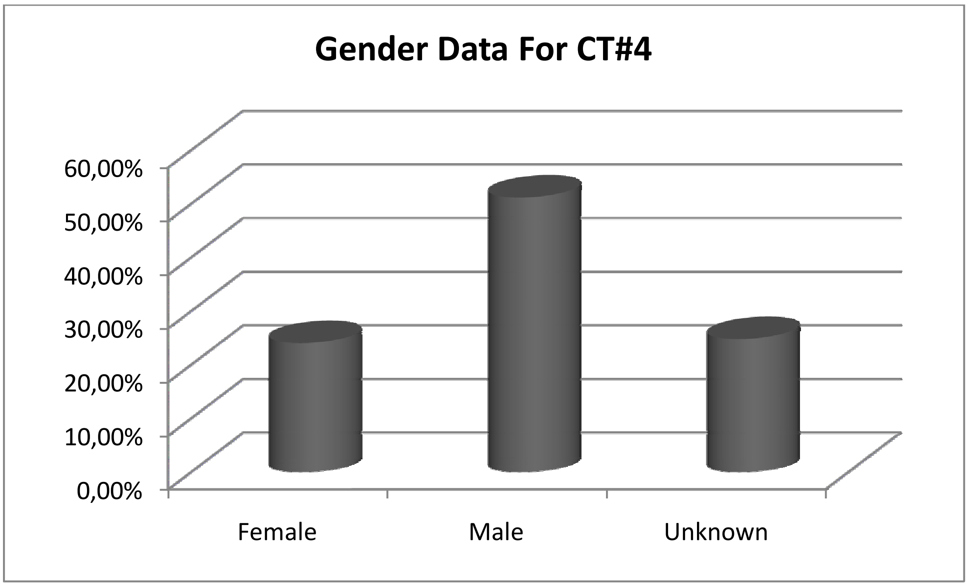


Figure 6.24 Gender data for CT#4

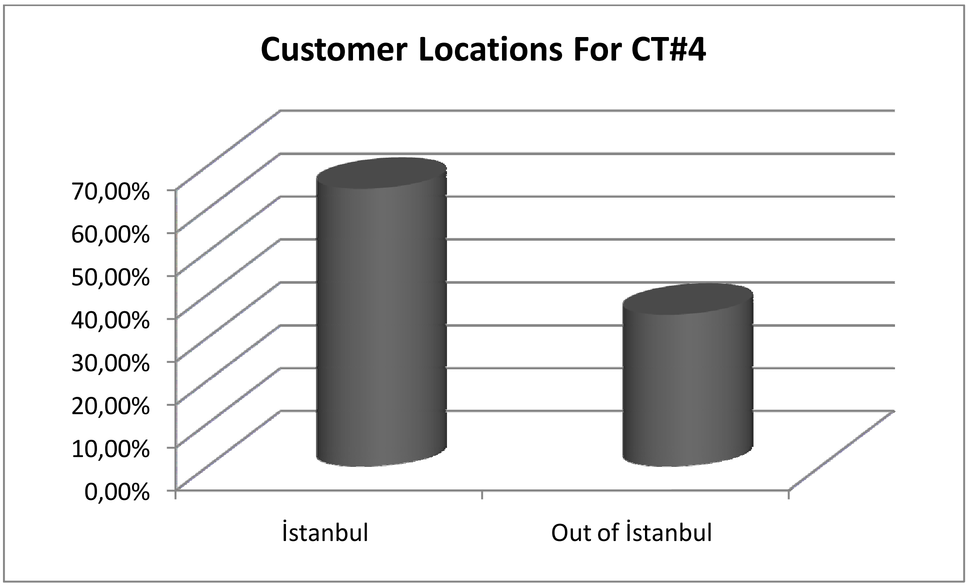


Figure 6.25 Customer locations for CT#4

6.2.4.2 Model 2 Customer Analysis – K- Means Clustering

It is aimed to define customer risk analysis of the training data according to the major variables exist in it. The input variables are selected the following fields;

Customer Age

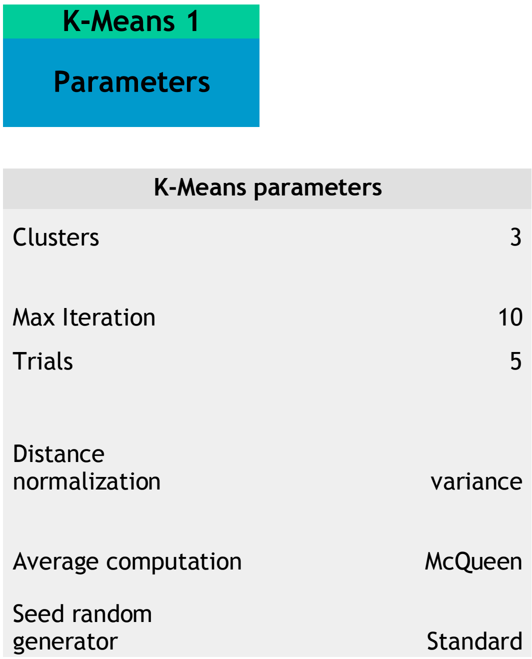
Sex

Marital Status

Customer Type

K-Means clustering method has been run on the training data with the selected parameters above in Tanagra application and the following results have been produced by Tanagra.

Table 6.6 K-means clustering-A parameters & results

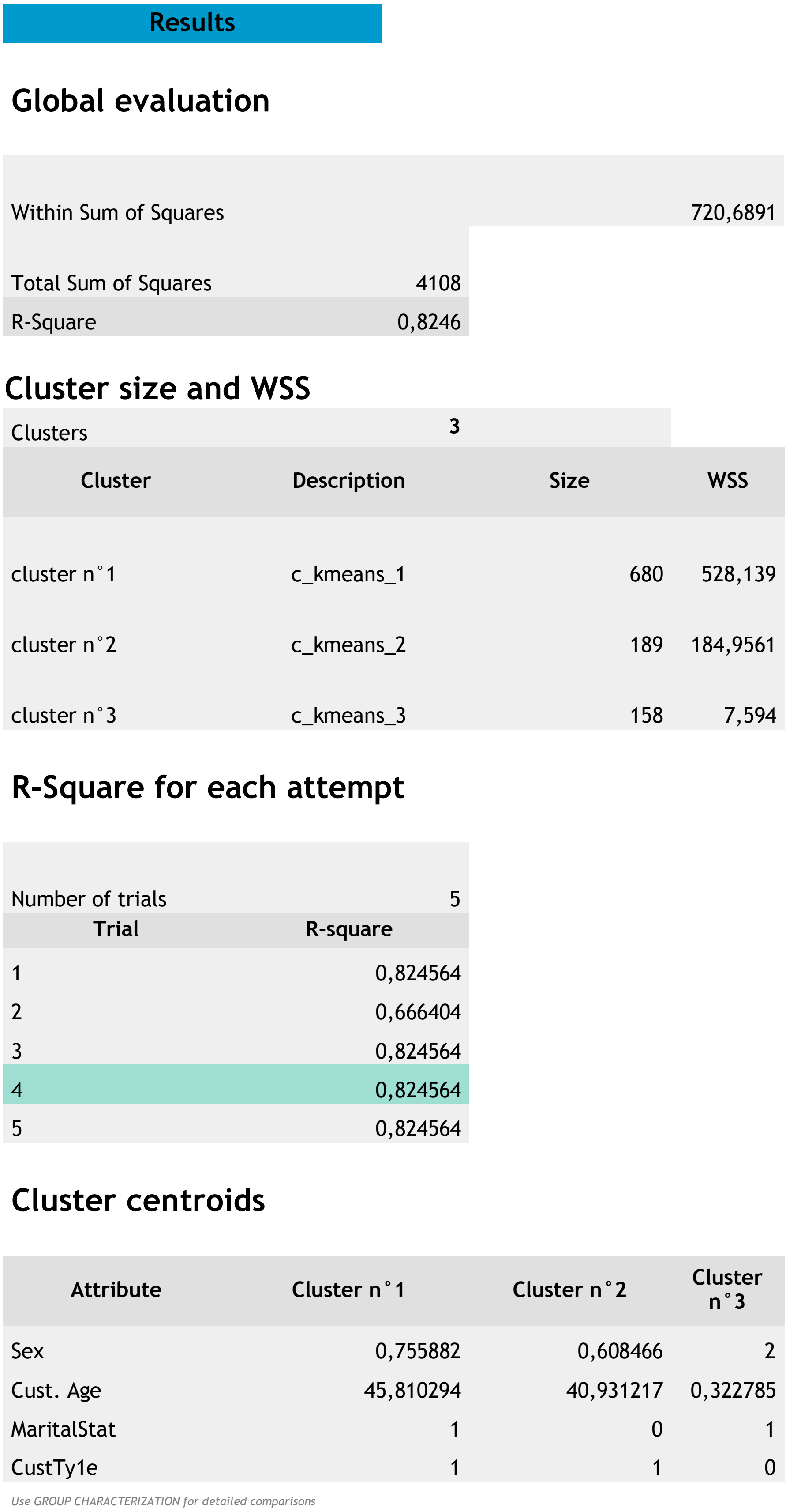


K-Means 1

Parameters

K-Means parameters	
Clusters	3
Max Iteration	10
Trials	5
Distance normalization	variance
Average computation	McQueen
Seed random generator	Standard

Table 6.6 K-means clustering-A parameters & results (continued)



Results

Global evaluation

Within Sum of Squares	720,6891
Total Sum of Squares	4108
R-Square	0,8246

Cluster size and WSS

Clusters 3

Cluster	Description	Size	WSS
cluster n°1	c_kmeans_1	680	528,139
cluster n°2	c_kmeans_2	189	184,9561
cluster n°3	c_kmeans_3	158	7,594

R-Square for each attempt

Number of trials 5

Trial	R-square
1	0,824564
2	0,666404
3	0,824564
4	0,824564
5	0,824564

Cluster centroids

Attribute	Cluster n°1	Cluster n°2	Cluster n°3
Sex	0,755882	0,608466	2
Cust. Age	45,810294	40,931217	0,322785
MaritalStat	1	0	1
CustTy1e	1	1	0

Use GROUP CHARACTERIZATION for detailed comparisons

According to the results, five trials have been performed and the 4.trial has been found the best solution for this model. Finally, three groups have been generated based on the input variables;

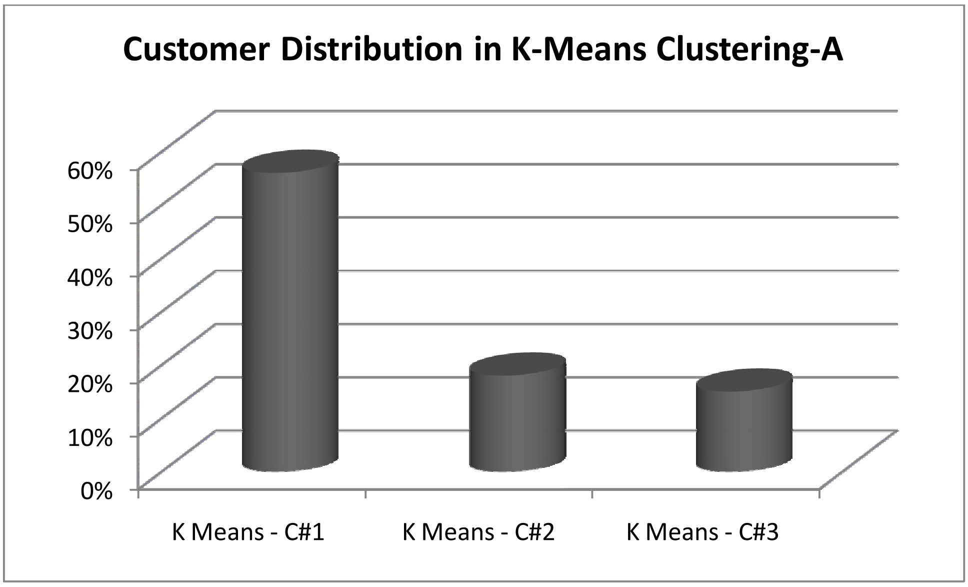


Figure 6.26 Customer distribution in k-means clustering-A

There are 680 records in **Cluster#1** and the following profile analysis is observed;

i. All the selected customers are personal customers and married

ii. The average age is 46 and the 72, 94 percent of the group is male

iii. The 99, 56 percents of the customers are working in construction sector (finance code 5, 6)

iv. The 91, 47 percents of the customers have retail loan products in the bank (risk code 120,121)

v. Total risk is 60.002.082 TL out of 193.002.082 TL

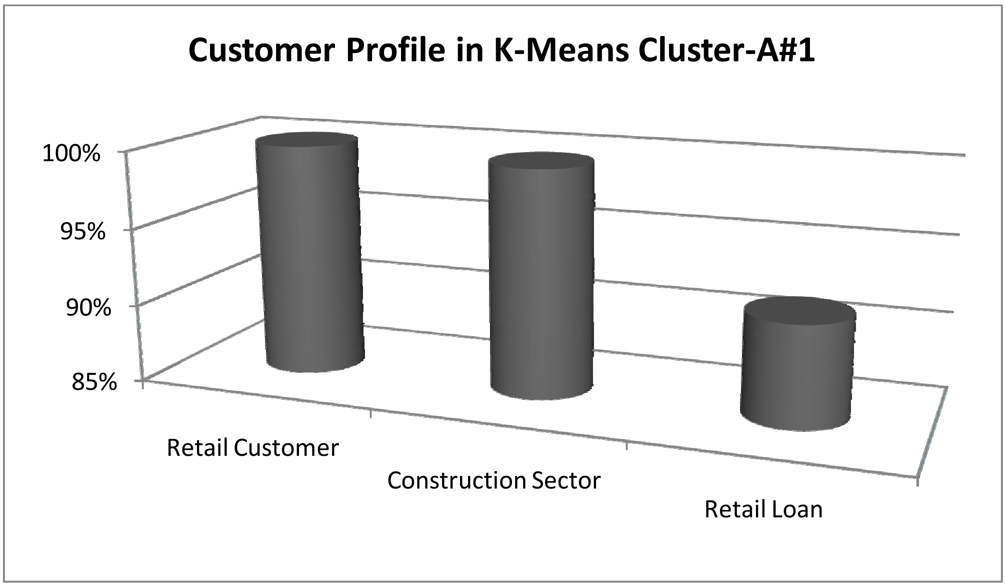


Figure 6.27 Customer profile in k-means clustering-A#1

There are 189 records in **Cluster#2** and the following profile analysis is observed;

i. All the selected customers are personal customers and single. The average age is 41 and the 60, 84 percent of the group is male

ii. All the customers are working in construction sector (finance code 5, 6)

iii. The 95, 77 percents of the customers have retail loan products in the bank (risk code 120,121)

iv. Total risk is 11.646.847 TL out of 193.002.082 TL

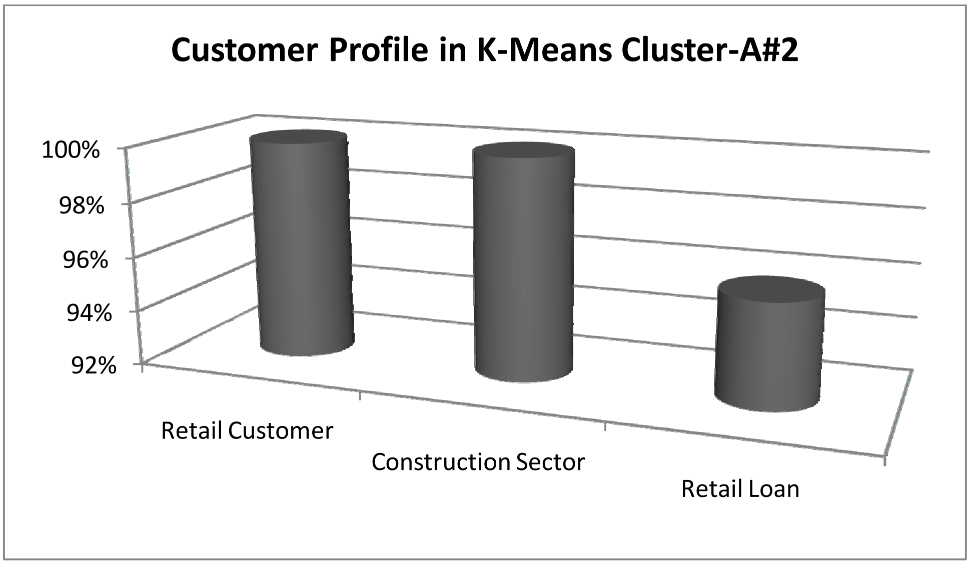


Figure 6.28 Customer profile in k-means clustering-A#2

i. All the selected customers are non personal (corporate) customers and the marital status of them have been set as married and there is no sex information. The average age is 32, 2

ii. The 44, 32 percents of the customers are working in manufacturing industry (finance code 3)

iii. The 12, 65 percents of the customers are working in construction sector (finance code 5, 6)

iv. The 13, 29 percents of the customers are working in Real Estate commission, renting and business (finance code 12)

v. The 81, 65 percents of the customers have loan products in the bank (risk code 100,101)

vi. Total risk is 120.533.227 TL out of 193.002.082 TL

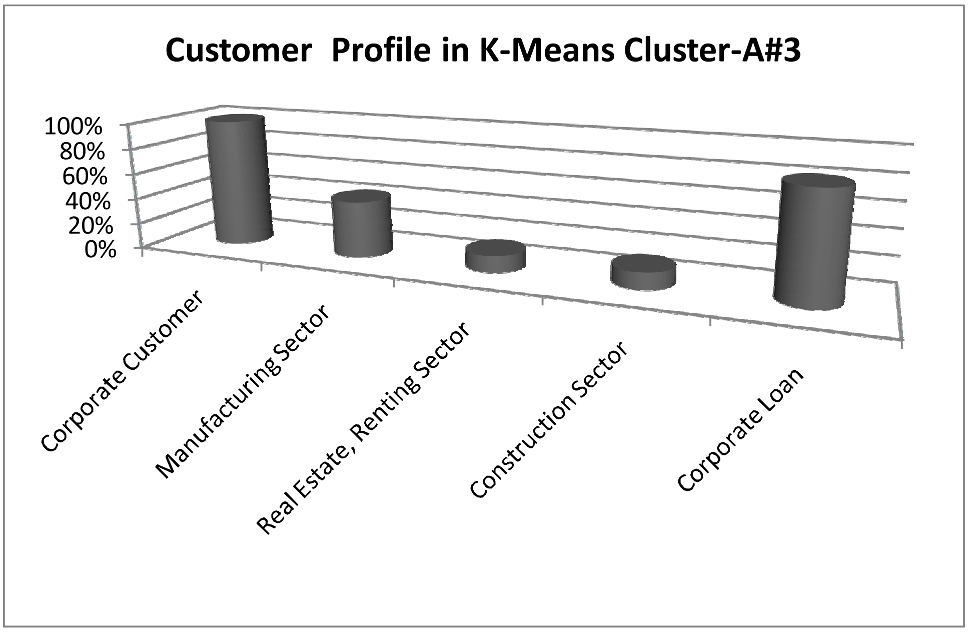


Figure 6.29 Customer profile in k-means clustering-A#3

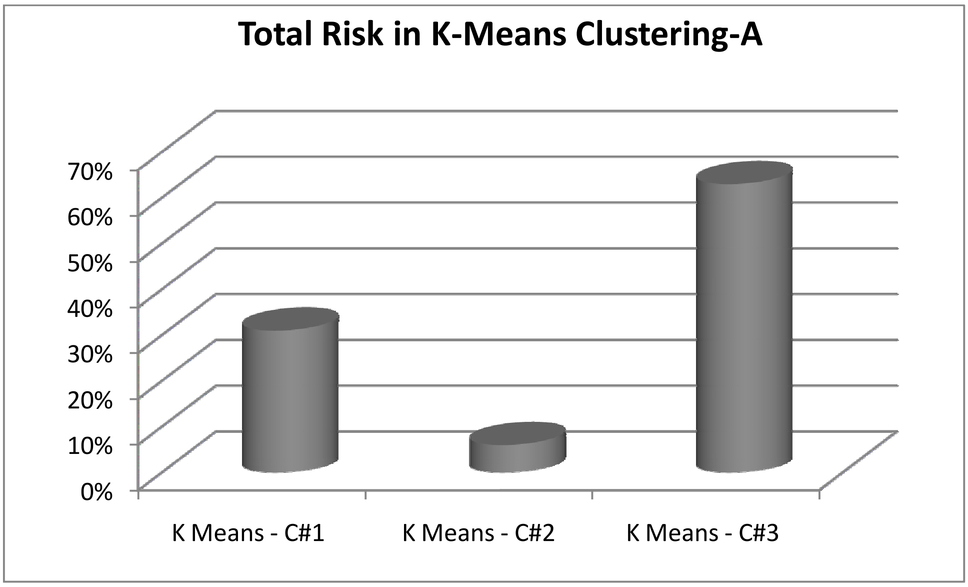
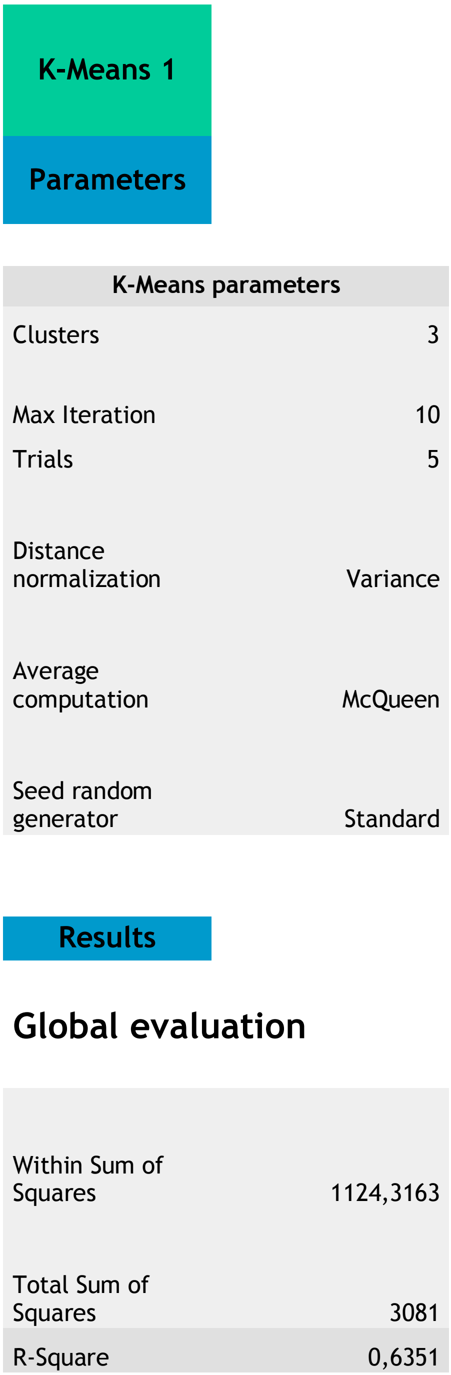


Figure 6.30 Total risk in k-means clustering-A

6.2.4.3 Model 3 Customer Risk Analysis – K- Means Clustering

It is aimed to define customer risk analysis of the training data according to the major variables exist in it. The target input variables are selected the following fields;
Cust. Type
City Code
Age in Bank
K-Means clustering method has been run on the training data with the selected parameters above in Tanagra application and the following results have been produced by Tanagra.

Table 6.7 K-means clustering-B parameters & results



K-Means 1

Parameters

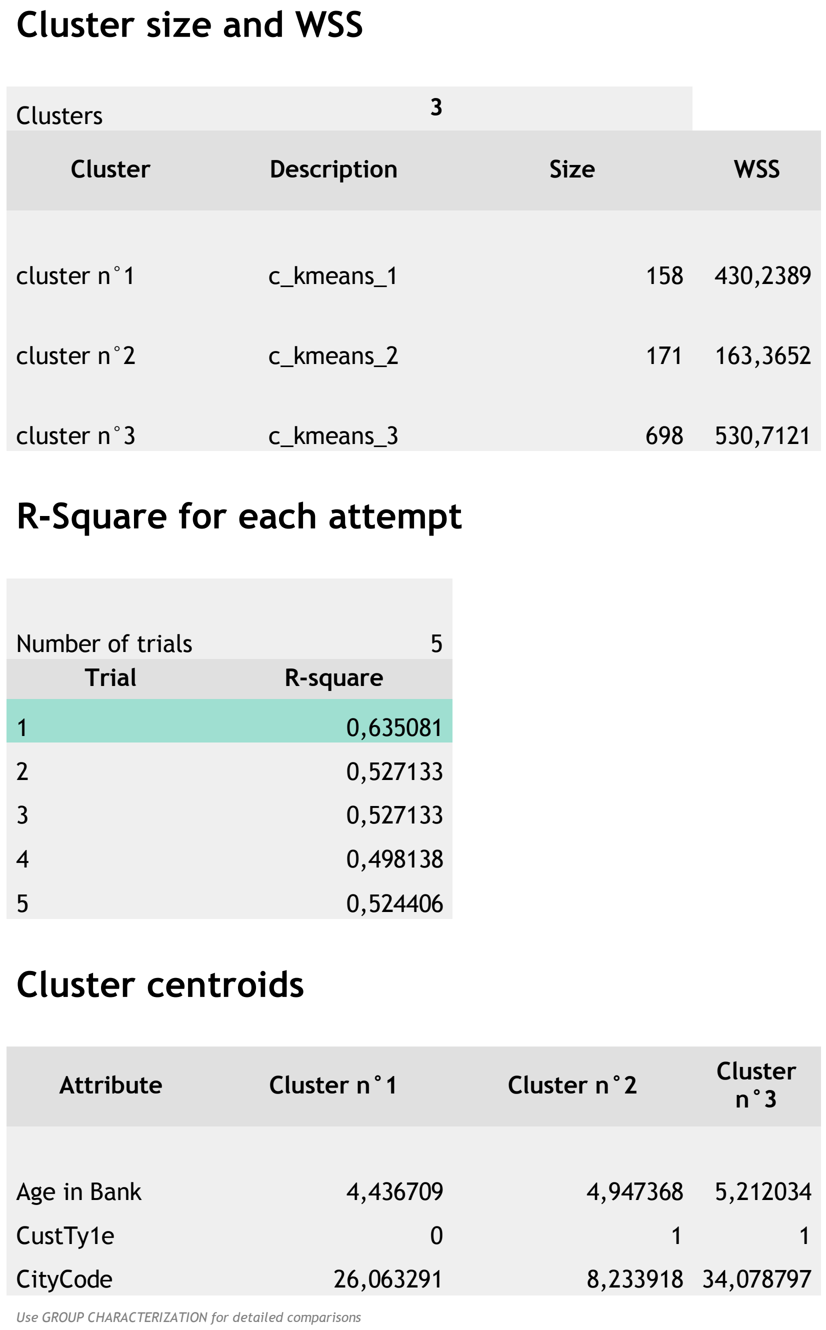
K-Means parameters	
Clusters	3
Max Iteration	10
Trials	5
Distance normalization	Variance
Average computation	McQueen
Seed random generator	Standard

Results

Global evaluation

Within Sum of Squares	1124,3163
Total Sum of Squares	3081
R-Square	0,6351

Table 6.7 K-means clustering-B parameters & results (continued)



Cluster size and WSS

Clusters	3		
Cluster	**Description**	**Size**	**WSS**
cluster n°1	c_kmeans_1	158	430,2389
cluster n°2	c_kmeans_2	171	163,3652
cluster n°3	c_kmeans_3	698	530,7121

R-Square for each attempt

Number of trials	5
Trial	**R-square**
1	0,635081
2	0,527133
3	0,527133
4	0,498138
5	0,524406

Cluster centroids

Attribute	Cluster n°1	Cluster n°2	Cluster n°3
Age in Bank	4,436709	4,947368	5,212034
CustTy1e	0	1	1
CityCode	26,063291	8,233918	34,078797

Use GROUP CHARACTERIZATION for detailed comparisons

According to the results, five trials have been performed and the 1.trial has been found the best solution for this model. Finally, three groups have been generated based on the input variables;

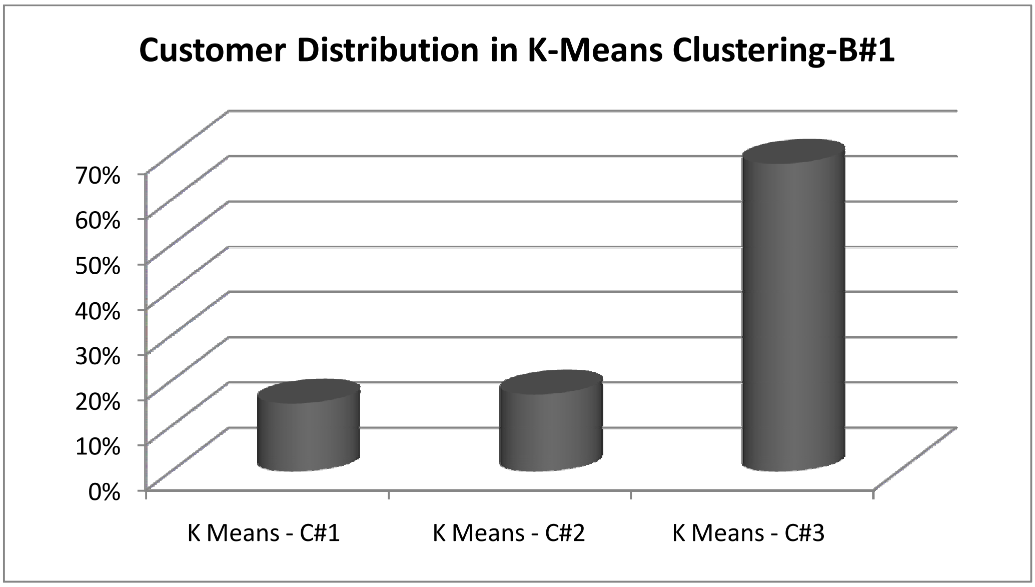


Figure 6.31 Customer profile in k-means clustering-B#1

There are 158 records in **Cluster#1** and the following profile analysis is observed;

i. All the selected customers are non personal customers

ii. The 60, 12 percent of the customers locate in İstanbul

iii. The 77, 21 percents of them are the customer of the bank is over 4 years

iv. The 81, 65 percents of the customers have loan products in the bank (risk code 100,101)

v. Total risk is 120.533.227 TL out of 193.002.082 TL

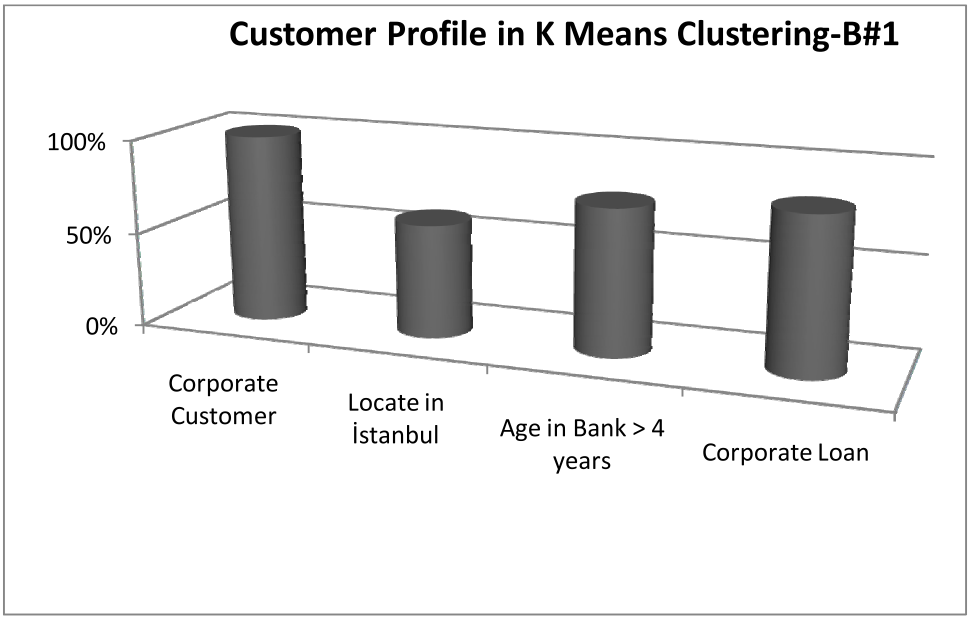


Figure 6.32 Customer profile in k-means clustering-B#1

There are 171 records in **Cluster#2** and the following profile analysis is observed;

i. All the selected customers are personal customers

ii. All the customers locate in out of Istanbul and Izmir

iii. The 81, 28 percent of them is the customer of the bank over 5 years

iv. The 94, 74 percents of the customers have retail loan products in the bank (risk code 120,121)

v. Total risk is 12.089.966 TL out of 193.002.082 TL

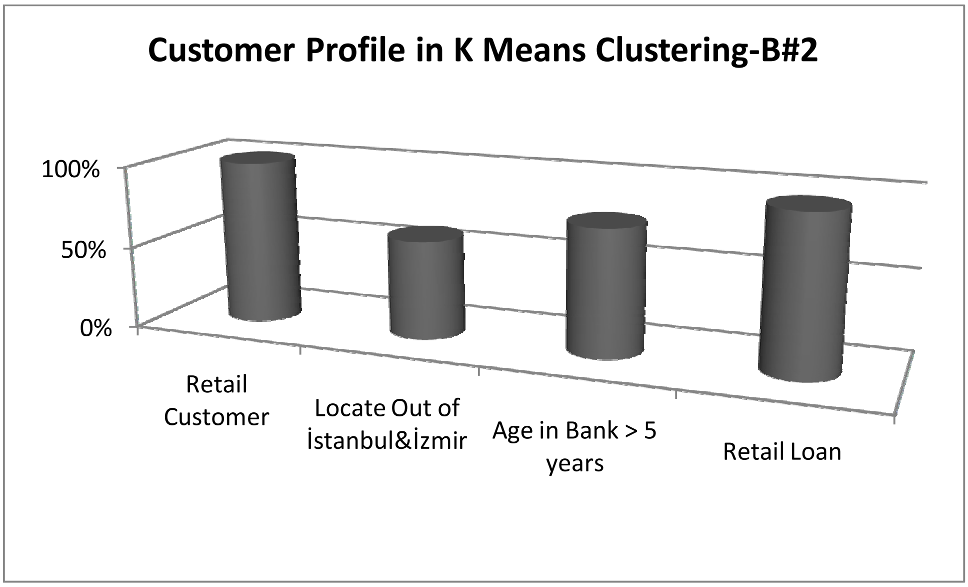


Figure 6.33 Customer profile in k means clustering-B#2

There are 698 records in **Cluster#3** and the following profile analysis is observed;

i. All the selected customers are personal customers

ii. All the customers locate in Istanbul and Izmir

iii. The 87, 10 percents of them are the customer of the bank is over 5 years

iv. The 97, 56 percents of the customers have retail loan products in the bank (risk code 120,121)

v. Total risk is 60.378.889 TL out of 193.002.082 TL

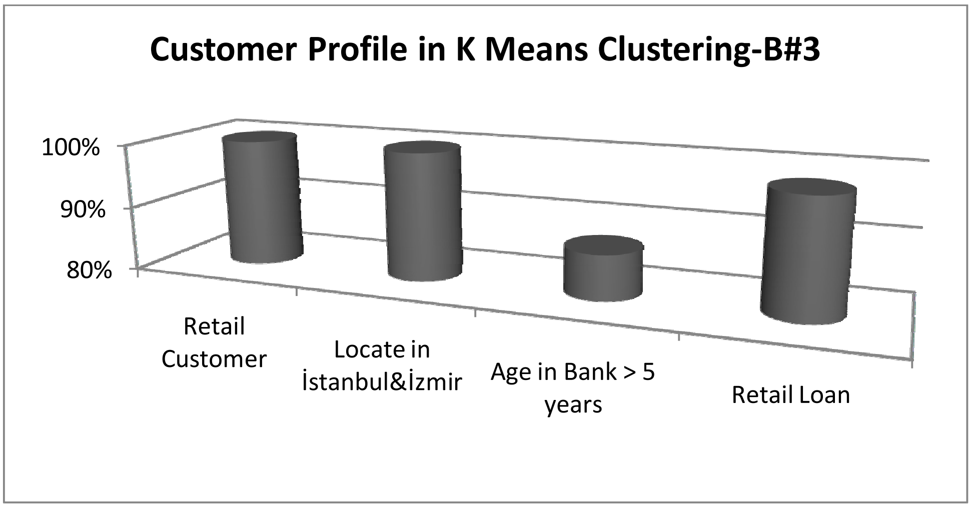


Figure 6.34 Customer profile in k-means clustering-B#3

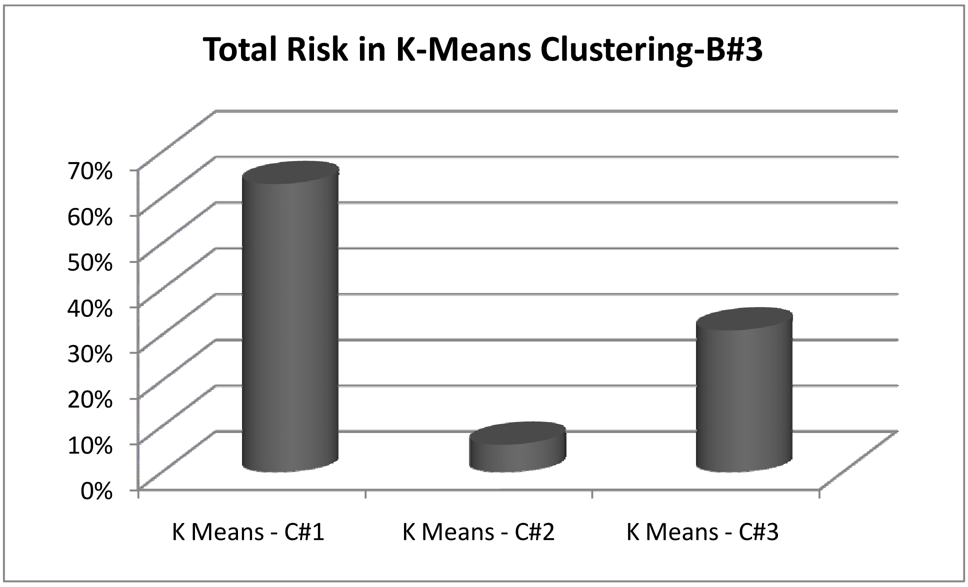


Figure 6.35 Total risk in k-means clustering-B#3

6.2.5 Suggestions

All main characteristics of the customer records have been inserted as input variables with the target of customer number. Therefore, customer pattern has been defined clearly based on "age in bank" of customers. According to the results, the customers are mainly male, between 41- 60 years old, retail customers and located in İstanbul. It can be suggested that, marketing should more focus on the customers with following features;

i. Female customers
ii. Customer located out of İstanbul
iii. Corporate customers
iv. Customers younger than 40 years old.

Additionally, it can be prepared and launched new marketing campaigns accordingly. Also, it is observed that active customer whose age in bank is less than 4,5 years old is the quarter of the customer whose age in bank higher that 4,5 years old. That result displays, customer loyalty is very high among the customer whose age in bank over 4,5 years old. But customer loyalty should be also improved in the group of customer whose age in bank lower than 4,5 years old.

The second model is build in order to discover customer risk pattern by using customer demographic information. Results of the model reveal the major customer group which is having high amount of loans in the bank is located in İstanbul and working for construction sector. The bank should definitely develop new marketing campaigns for gaining new customers out of İstanbul and in different sector as well as improving the loyalty of exiting customer.

Results of the third model clearly reveal that the bank is focused on quite limited customer groups and possible has not very dynamic and various marketing strategies. Even though, corporate customers have high volume loans and very profitable, the bank has very low volume of corporate customers. It can be suggested that the bank should try to gain new corporate customer not only in İstanbul but also out of İstanbul. The bank also should improve CRM and MIS systems for creating new marketing campaigns and informing customers.

7. CONCLUSION

This research is based on providing profound information about knowledge discovery and data mining disciplines as well as introducing well-known and most widely used data mining processes and methods by applying appropriate techniques into customer data supplied from a bank. The aim of the study is to display the importance of data management and evaluation mechanism in business life.

Data mining has firstly started with the data collection activities early in 1960s and been improved depended on the developments in information technologies and quality management as newly invented in 1980s. When it was reached to 1990s, data mining was became data warehouse and decision support systems were applied in areas of marketing and customer relationship management issues. Finally, it was started called as data mining with speedy changes and improvements in knowledge discovery and information technologies.

Data mining has an organic link to Knowledge Discovery in Database (KDD), data warehouse methodologies and is the major topic for artificial intelligence.

Data mining has very close relationship between statistics, database technologies, machine learning, visualization and information science. With the support of these disciplines, data mining produce very effective, efficient and useful results in decision making, customer relationship management, marketing and sales areas. Data mining is commonly used in finance, banking, retail, marketing, manufacturing, transportation and health care.

Data mining is commonly used in banking industry not only customer relationship management but also providing legal reporting for the associated organs of government periodically.

According to the literature research was performed within the scope of this study, data mining produce very useful results for health care such as cancer survivability, fraud detection, customer relationship management, financial decision making in banking, finance industry, knowledge base mining in internet, quality metrics in manufacturing industry, knowledge management for marketing.

Customer relationship management is one of the main application areas of data mining. In every steps of CRM such as customer identification, development, retention and attraction, data mining techniques can be efficiently used as the task of description, estimation, prediction, classification, clustering and association.

The most important step of data mining is initial phase called business understanding. If the business goals in other words the purposes of data mining project are not clearly defined, rest of the phases will not be able to proceed to complete properly. Also the following phases, data understanding and preparation must be carefully analyzed and performed in order to produce more realistic and appropriate results. Data cleaning and exploratory techniques must be selected according to business goals and data mining methods are supposed to applied. Another important issue is the selection of most appropriate modeling method that is also most dependent to business goals and current data features. As a result, all phases of data mining are relatively connected to each other and the selections in phases have directly affect the following phase and to be affected from previous phase. Thus, data mining phases should be evaluated as whole system and the parameter assignments must be performed taking into consideration of all sources, circumstances and constraints in the data mining project.

In the last section of the study, a customer profile analysis in terms of customer demographic and risk information are performed and clustering technique has been selected. There were two methods as clustering tree and k-means clustering have been applied into banking customer data. In this practice, it has been observed that, setting business goals was easily defined but the data preparation phase took a long time. Data cleaning was the time consuming phase of the whole project. Raw data can contain huge mass of dirty data. In this case, data miner should be very carefully used data cleaning and preparation techniques in order to protect the original concept of the data characterization; otherwise some important features of the data can be spoilt. After completing data preparation and modeling phases, the results of all three models are evaluated within ratio of the model and also evaluated by comparing to each other. It is noticed that, they are quite realistic models as R is over 0,6 and 0,8. The model will be more realistic and appropriate if the ratio is more close to the value of 1.

In conclusion, this study gives detail information about data mining including their processes, methods, application areas and evaluation of it. Also provides an application of data mining in banking environment in order to show all phases of a data mining project. Therefore, that practice will be useful for the readers who want to work in data mining projects.

REFERENCES

Books

Breiman Leo, Friedman Jerome, Olshen Richard, and Stone Charles(1984), Classification and Regression Trees, Chapman&Hall/CRC Press,Boca Raton,USA

Cios Krzysztof J., Pedrycz Witold, Swiniarski Roman W., Kurgan Lukasz A.(2007), Data Mining, A Knowledge Discovery Approach, Springer

Cooper G. F. and Herskovitz E. (1992) A Bayesian method for the induction of probabilistic networks from data. Mach Learn, Volume 9, Pages 309-347.

Delen D., Walker G., Kadam A. (2005), Predicting Breast Cancer Survivability: A comparison of Three Data Mining Methods, Artificial Intelligence In Medicine, Volume 34, Pages 113-127.

Fausett Laurene (1994), Fundamental of Neural Networks, Prentice Hall, Upper Saddle River, NJ

Feeders A., Daniels H., Holsheimer H.M. (2000), Methodological and Practical Aspects of Data Mining ,Information&Management,Volume 37, Pages 271 – 281

Geiger D. and Heckerman D. (1997) A characterization of Dirichlet distributions through local and global independence. Ann Stat Volume 25, Pages 1344-1368,1997.

Glymour C., Scheines R., Spirtes P. and Kelly K.(1987) Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling. Academic Press, San Diego, CA

Groth Robert (1999), Data Mining :Building Competitive Advantage,Prentice Hall, New Jersey

Han J. and Kamber M.(2001), Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco,USA

Hand David , Mannila Heikki and Smyth Padhranic (2001), Principles of Data Mining,MIT Press, Cambridge,MA

Haykin Simon (1990), Neural Networks: A Comprehensive Foundation, Prentice Hall, Upper Saddle River, NJ

Kohonen Tuevo (1982), Self-organized formation of topologically correct feature maps, Biological Cybernetics, Vol.43, Pages.59-69

Kohonen Tuevo (1989), Self-Organization and Associative Memory, 3rd ed.,Springer-Verlag, Berlin

Larose,D.T.(2005), Discovering Knowledge in Data: An Introduction to Data Mining, John Wiley and Sons,Ltd.,Hoboken,N.J.,USA

Maimon O. and Rokach L. (2005), Data mining and knowledge discovery handbook, Springer, NY,USA

Olson D.L., Delen,D.(2008), Advanced Data Mining Techniques, Springer Science+Business Media, Inc., NY,USA

Pearl J.(1988), Probabilistic Reasoning in Intelligent Systems: Networks of plausible inference. Morgan Kaufmann, San Francisco, CA

Quinlan J.Ross (1992),C4.5:Programs for Machine Learning,Morgan Kaufmann, San Francisco,USA

Ritter Helge (1995), Self-organizing feature maps: Kohonen maps, in M.A.Arbib,ed.,The Hand-book of Brain Theory and Neural Networks, Pages 846-851, MIT Press, Cambridge,MA

Thomas A., Spiegelhalter D. J. and Gilks W. R.(1992), Bugs: A program to perform Bayesian inference using Gibbs Sampling. In J. Bernardo, J. Berger, A. P. Dawid, and A. F. M. Smith, editors, Bayesian Statistics 4, pages 837-42, Oxford University Press, Oxford, UK

Tiwana A. (2001), The Essential Guide to Knowledge Management :E-buseinessand CRM Applications, Prentice Hail PTR, Saddle River

Witten ,I.H. And Frandk,E.(2005), Data Mining Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers, California,USA.

Wright S. (1923), The theory of path coefficients: a reply to niles' criticism. Genetics, Pages 239-255

Periodicals

Anderson Joan L., Jolly Laura D., Fairhurst Ann E. (2007), Customer relationship management in retailing: A content analysis of retail trade journals. Journal of Retailing an Consumer Services, Volume 14, Pages 394 – 399

Berry M. J.A., Linoff G.S.(1997) Data Mining Techniques For Marketing, Sales and Customer Support, John Wiley & Sons Inc.,USA

Berry,M.J.A and Linoff,G.S. (2004),Data Mining Techniques : For Marketing, Sales and Customer Relationship Management, Wiley Publishing,Inc.Indianapolis, USA.

Charniak E. (1991)Belief networks without tears. AI Magazine, Page 5042

Daskalaki S., Kopanas I., Goudara M., Avouris N. (2003), Data Mining For Decision Support on Customer Insolvency in Telecommunications Business, European Journal Of Operational Research, Volume 145, Pages 239-255

Hong T., Han I. (2002), Knowledge-based data mining of news information on the Internet Using Cognitive Maps and Neural Networks,Systems with Application, Volume 23, Pages 1-8

Hui S.C., Jha G. (2000), Data Mining For Customer Service Support, Information & Management, Volume 38, Pages 1 - 13

Konrad Rachel (February 2001), Data Mining: Digging user info for gold,ZDNET News

Malone J., McGarry K., Bowerman C. (2006), Automated Trend Analysis of Proteomics Data Using an Intelligent Data Mining Architecture, Expert Systems with Applications, Volume 30, Pages 24-33

Metayashuck Jennifer (1999), The National IT Salary Survey: Pay up, Information Week

Musaoglu, C., (2003) Customer acquisition and retention modeling in consumer finance sector using data mining,Thesis Study, Bogazici Press, İstanbul

Ngai E.W.T., Xiu Li, Chau D.C.K.(2009), Application of data mining techniques in customer relationship management : A literature review and classification, Expert systems with application, Volume 36, Pages 2592 – 2602.

Nie G., Zhang L., Liu Y., Zheng X.,Shi Y. (2009), Decision Analysis of Data Mining Project Based on Bayesian Risk, Expert Systems with Application, Volume 36, Pages 4589 - 4594

Olafsson S. (2006), Introduction to operations research and data mining, Computer & Operations Research, Volume 33, Pages 3067-3069

Olafsson S., Li X., Wu S. (2008) Operations research and data mining, European Journal Of Operational Research, Volume 187, Pages 1429-1448

Sadıç Şenay (2008), Master Thesis: Data Mining Including Application of Cognitive Maps and Decision Tree Algorithm, İTU, İstanbul

Shaw M. J., Subramanian C., Tan Gek Woo, Welge Michael E.(2001), Knowledge management and data mining for marketing. Decision Support System, Volume 31, Pages 127- 137

Shu Y. (2007), Inference of Power Plant Quake-Proof Information Based on Interactive Data Mining Approach, Advanced Engineering Informatics, Volume 21, Pages 257-267

Sun Jie, Li Hui (2008), Data Mining Methods for Listed Companies' Financial Distress Prediction, Knowledge Based System,Volume 21, Pages 1 - 5

Tayi G., Davidson I. (2009), Data Preparation Using Data Quality Matrices For Classification Mining, European Journal of Operational Research, Volume 197, Pages 764-772

The Technology Review Ten (January/February 2001.), MIT Technology Review

Unal Osman Onat (2008), Master Thesis: Data Mining Applications on Web Usage Analysis & User Profiling, İTU, İstanbul

Wang H., Weigend A.S. (2002), Data Mining For Financial Decision Making, Decision Support Systems, Volume 32, Pages 417-418

Yeh I-Cheng, Lien Che-hui (2007), The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients, Expert systems with application, Volume 36, Pages 2473-2480

Zikmund W.G., McLeod R. Jr. and Gilbert F. W.(2003), Customer Relationship Management: Integrating Marketing Strategy and Information Technology., Leyh Publishing, Danvers, MA., USA.

Other Publications

www.gartner.com

www.jstor.org/pss/2685468

CURRICULUM VITAE

Name Surname: Ecehan ÇETİN

Address :. Acıbadem Mah. Beyazleylaklı Sk. Gülen Apt.No:1/17 KADIKÖY / İSTANBUL
Date and Place of Birth: Samsun, 1971

Foreign Language: English

Undergraduate Program: İstanbul Technical University Industrial Engineering,1992

Graduate Program: Bahçeşehir University 2011

Name of Institute: The Graduate School of Natural and Applied Sciences

Name of Program: Industrial Engineering Graduate Program

Publication: -

Professional Experience:

Millennium Bank A.Ş.	
Project / Quality Manager in IT Department	May’2011& Aug.2007
Oyakbank A.Ş.	
Supervisor in IT Department	Jul.2007& Nov.2006
Provus Software Company	
Senior System Analys in IT Development Dept.	Oct.2006 & Nov.2003
Hsbc/Demirbank A.Ş.	
Expert Business Analyst in IT Department	Oct.2003& Jun.1999
Logo Business Solutions	
System Analyst on Software Department	May.1999& Nov.1997
İstanbul Çorap Sanayii A.Ş.	
Production Planning Expert	Sep.1995& May 1993