THE REPUBLIC OF TURKEY

BAHCESEHIR UNIVERSITY

CUSTOMER SEGMENTATION IN DIGITAL BROADCASTING

Master Thesis

SÜLEYMAN MESUT KEÇECİOĞLU

İSTANBUL, 2016

THE REPUBLIC OF TURKEY

BAHCESEHIR UNIVERSITY

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE COMPUTER ENGINEER

CUSTOMER SEGMENTATION IN DIGITAL BROADCASTING

Master Thesis

SÜLEYMAN MESUT KEÇECİOĞLU

Supervisor: ASSOC. PROF. DR. ALPER TUNGA

İSTANBUL, 2016

THE REPUBLIC OF TURKEY

BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES SCHOOL

COMPUTER ENGINEERING

Name of the thesis : Customer segmentation in digital broadcasting.

Name/Last Name of the Student : Süleyman Mesut Keçecioğlu

Date of the Defense of Thesis: 25.05.2016

The thesis has been approved by the Graduate School of Natural And Applied Sciences.

Assoc. Prof. Nafiz ARICA

Acting Director

I certify that this thesis meets all the requirements as a thesis for the degree of Master Of Arts.

Assist. Prof. Dr. Tarkan AYDIN

Program Coordinator

This is to certify that we have read this thesis and we find it fully adequate in scope,quality and content, as a thesis for degree of Master Of Arts.

Examining Comitte Members	Signature
Assoc. Prof. Dr. Alper TUNGA	_____
Assoc. Prof. Dr. Ahmet KIRIŞ	_____
Assist. Prof. Dr.Cemal Okan ŞAKAR	_____

ABSTRACT

CUSTOMER SEGMENTATION IN DIGITAL BROADCASTING

Süleyman Mesut Keçecioğlu

Computer Engeering

Thesis Supervisor: Assoc. Prof. Dr.Alper Tunga

May 2016, 65 Pages

Many companies have big data about their customers, they need to process these data with various analysis techniques to make meaningful information that provides to take better decisions in the competitive world.One of analysis techniques is Data mining. It determines hidden informations and patterns in a data set to help enterprises to decide. Data mining has several application areas. The most common of them is Customer segmentation. it is the process of splitting customers into sets. Moreover, it is used for developing customized marketing strategy.

The intention of this study is to show relation between viewing habits and customer demographic informations. Results of clustering algorithms are examined with examples. To achieve this goal, clustering methods applied to last one year customers` products and demographic information of Digital Brodcast Company which I worked with. To do that, we can see results using different clustering methods. Process of identifying clusters are based on Euclidean distance.

Firstly we tell about description of data mining, algorithms of data mining and techniques of data preparation and definition of segmentation algorithm. After definitions, we tell about represent of attributes and general information about customer data with graph and tell about preparation of data. After preparation of data, we apply different clustering techniques on this data and show differences between these results.

In the last section, findings of the research are interpreted and discussed to offer strategies for digital broadcasting companies.

Keywords: Clustering, customer segmentation, data mining

ÖZET

DİJİTAL YAYINCILIKTA MÜŞTERİ KÜMELENMESİ

Süleyman Mesut Keçecioğlu

Bilgisayar Mühendisliği

Tez Danışmanı: Doç. Dr. Alper Tunga

Mayıs 2016, 65 Sayfa

Günümüzde çoğu şirket, müşterileriyle ilgili büyük veriler barındırmaktadır. Şirketler bu verileri çeşitli analiz teknikleriyle işleyerek, çıkan sonuçları şirket adına daha iyi kararlar verilmesinde kullanmaktadırlar. Bu tekniklerden biri de veri madenciliğidir. Veri madenciliği, analiz tekniklerinin sonucunda ortaya çıkan anlamlı veriden bir şablonun çıkarılmasını sağlar. Bu şablon şirketlerin müşteri segmentasyonu için yol gösterici durumundadır. Bu segmentasyonun sonucunda müşteri dataları belirli kümelere ayrıştırılmaktadır. Bu kümeler şirketlerin pazarlama stratejilerinin oluşturulmasında etkin rol almaktadır.

Bu çalışmanın amacı çeşitli segmentleme methodlarını kullanarak elimizdeki büyük veriyi işleyip, müşterilerin TV izleme alışkanlıkları ile demografik bilgileri arasındaki ilişkiyi gösteren bir şablon ortaya çıkarabilmektir.Bu işlemler için çalıştığım kurumdaki müşteri datasından örnek bir müşteri kitlesi alınıp, bu kitlenin son 1 yıldaki işlem hareketleri kullanılmıştır. İşlemler sonucunda farklı segmentleme algoritmalarından farklı sonuçlar elde edilip bu sonuçların karşılaştırılması ve yorumlanması yapılmıştır.

Çalışma sırasında ilk olarak veri madenciliğinin tanımı ve veri üzerinde kullanmış olduğum veri hazırlama tekniklerinin ve segmentleme algoritmalarının tanımları yapılmıştır. Daha sonrasında ise çalışmada kullanmış olduğum veriyi tanımak için verinin yapısı ve özellikleri grafiklerle açıklanmıştır. Veri hazırlama işlemi bittikten sonra farklı segmentleme algoritmalarıo sırasıyla veri üzerinde uygulanıp sonuçları gösterilmiştir.

Son bölümde segmentleme algoritmalarından alınan sonuçlarının birbirleriyle karşılaştırılıp yorumlanması yapılmıştır.

Anahtar Kelimeler: Kümeleme, Müşteri Segmentasyonu, Veri Madenciliği

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SYMBOLS

Mean Value	:	μ
Standart Deviation	:	σ
Percent	:	%

1.INTRODUCTION

With developing competitive marketing areas; information about your customers and their needs is primary to develop your service and marketing strategies. That`s why the concept of Customer Relationship Management (CRM), that lets on a firm to develop a different strategy for handling marketing choices about their customers. Even though the primary aim is designed to increase sales figures, this approaches are nowadays being used to develop customer service and upgrade the marketing process.

Knowledge is power and always been power.In new economy era, knowledge has a priority at least labor,raw element and capital.Adam Smith emphasized that importance of saving production factors which make nations to be richer. These production factors are decreasing by sharing and these factors cause wars in last century.oil etc.

Knowledge is a different from production factor. It is increasing by sharing, and unless it is shared, knowledge converts to garbage data.Therefore storing all data is not enough, these data must been processed for converting to meaningful form. Especially if we consider that there will be 8 billion people in the world in 2020 and 30 billion smart things which can connect to internet, data has a more important role.

In today`s, size of data is rapidly increasing day by day because of social media,shopping chart sites,blog sites and sharing digital content.

Of course, with developing capabilities of computers and increasing data on world wide web have useful abilities for companies and marketers.On these abilities, marketers who want to know everything about costumers – what they buy,where they live,how old they are – can store all data without checking they need. Result of this, many companies have garbage data and nobody use this huge data.In this stage some software application appeared which tries to process data for making meaningfull to marketers.

The hardest concept in marketing sector is understanding consumer`s behaviours. The term of consumer behaviour is defined as the interactions of consumers with purchasing,using products and services.Having this knowledge provides great advantages to marketers and results of this they can rapidly evolving their marketplace.

Today`s, billions of people communicate with each other on social networks, make shopping or share pictures,message on internet. All of these interactions produce new data in every second. So size of data is expeditiously increasing in every second. For this data, firms need larger storage area contiunously for keeping huge databases. They store all data of all transactions records which have hidden data. When processing this hidden data, companies can get valuable knowledge about customers. Therefore recently, data and data mining are very essential for firms to gather meaningful information from raw data.

We prefer to work on data of digital broadcasting. This data include demographic information about customer like age, city, gender, favourite team and include preferences of digital broadcast services such as has movie,sport,erotic contents.

The main purpose of the study is clustering customer with content preferences. To do that we make many experiment with different clustering methods. After that many consequences can be produced from these clusters.

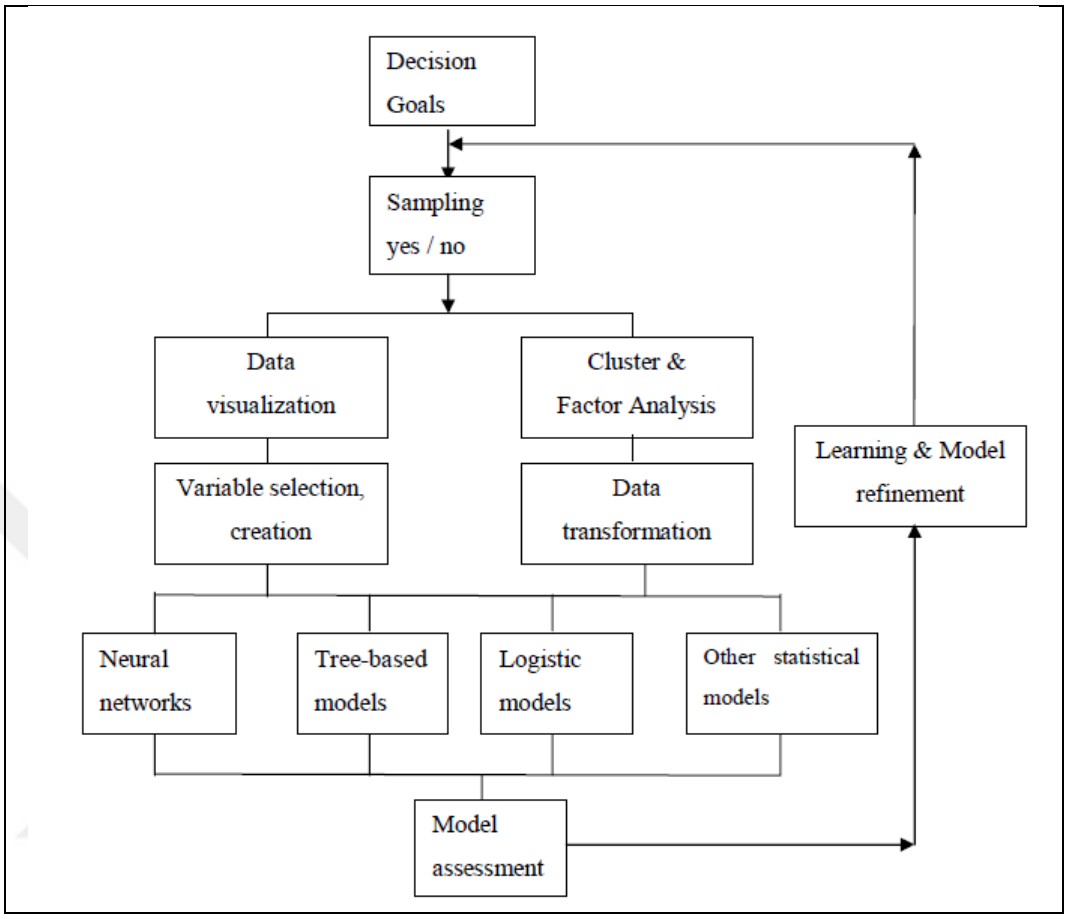
2.LITERATURE SEARCH

The spread of internet use and developments in storage technologies enable companies to meet with big data. In this concepts, the biggest challenge is how to convert ostensibly useless data into useful information. In parallel with advance in technology, techniques of data mining have become more important for solving a variety of industry problems. The importance of these techniques for customer segmentation is becoming indispensable in the field of customer relation. In the literature, many studies handle customer segmentation problem in variety of sectors and various methods are proposed for this purpose. In this section, methodologies and studies of data mining about the customer segmentation problem will briefly be introduced.

During the recent years, various data mining techniques containing data visualization, classification, generalization, clustering, association and have applied to several areas. Data mining techniques include a lot of algorithms serving for many aims such as description, description, prediction and association etc. In literature, specifically, classification techniques including regression models, support vector machine (SVM), naïve bayes, decision tree, clustering, neural networks (NN), association rule algorithms are generally preferred to solve the huge data problems.

There is a relationship between data mining and Knowledge discovery in databases (KDD). It is an modeling of large database and explorative analysis and systematically executed. In addition, It is alos KDD is also described as learning process and an iterative discovery and it is a knowledge management framework.

Figure 2.1: Knowledge management process

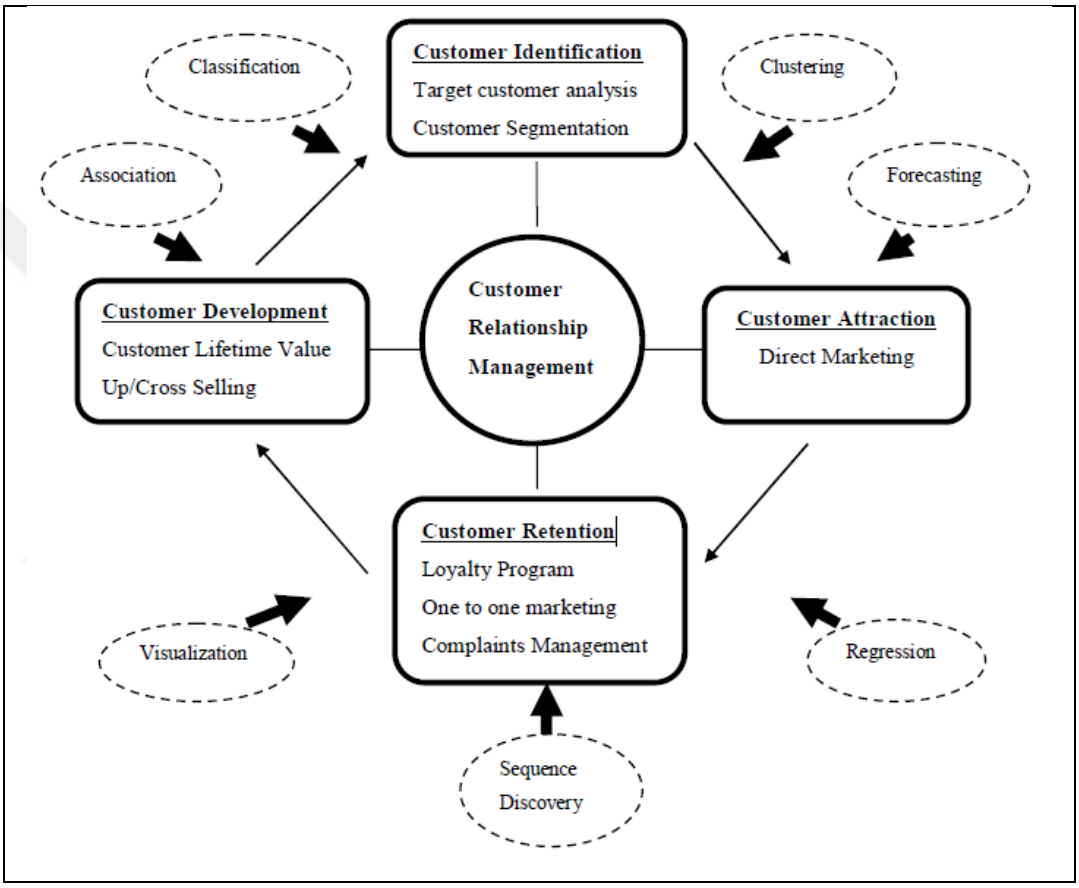


After making decision on using sample or complete database, the next step will be proceeding which is explored the data using the tasks such as visualization. And also appropriate data mining techniques are selected and applied over the data. The outcome of that data mining practice is evaluated to define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

Data mining techniques and tools have a great impact on implementing customer relationship management in various industries. CRM is described as a process of picking up and extensive strategy, identifying and helping to selective customers to develop important value for the firm and the customer (Ngai & Xiu & Chau, 2009).at al, 2009). CRM includes the customer service, sales, integration of marketing and supply chain functions of the enterprise in order to accomplished distributing customer value effectively and efficiently. That shows that CRM is very important instrument for

obtaining customers with the help of data mining techniques. 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 sample flow diagram of classification for data mining techniques is shown in Figure 2.2 below (Ngai & Xiu & Chau,2009)

Figure 2.2: Sample Flow Diagram of Classification Framework for Data Mining



In literature, there is a lot of research effort going on Customer Segmentations in digital broadcasting. Some of them focused relationship between demographic informaton of humans and watched contents.What is more, some of them focused the interactions of members for churn segmentation.

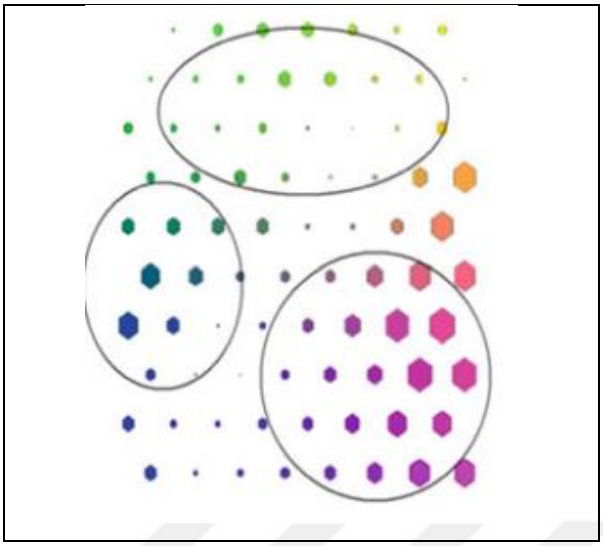
One of the studies (Jonathan Burez at.al February 2007) draws attention the early discovery of possible churners for Pay-TV members. The authors are used watched contents and information of members. This paper shows how a firm performing on a

subscription basis can different binary classification techniques and solution high attrition rates are applied on this customer-churn case: random forests, Markov chains and logistic regression. the most appropriate model for this case is chosen. Then authors are developing churn-prediction model.

Another research (M2 Presswire at.al June 2015) made a research about segmentation for VOD & OTT usage analyzes. It shows parks associates' video viewing segmentation techniques and than explores the comparions between OTT and VOD charge and use within each segment. It explores the interest to access on demand content on different devices. For example, smartphones, tablets and computers. It describes that characterisics of super buyer, consumer segmentation and analysis of OTT vs VOD consumption and expenditure trends.

In antoher survey about segmentation for broadcast industry (Chen, Nai-hua; Huang, Stephen Chi-tsun; Shu, Shih-tung; Wang, Tung-sheng at.al 2013), the authors are examined overall satisfaction, service quality and market segmentation. This paper examines realtions among three overall satisfaction, payment interactions, and service qualities, and customer segment temperance of these relations. Structure of this study is about broadcast for Taiwan. In this research, The demographic variables used (personal monthly income, environmental factors, number of family member, education, viewing hours per day,age, years of installation and). Data is seperated three groups (Figure. 2.3) and then use the K-means method to divide data seperately.

Figure 2.3 : Results of K-Means Clustering



Source: Chen, Nai-hua; Huang, Stephen Chi-tsun; Shu, Shih-tung; Wang, Tung-sheng 2013 Market segmentation, overall satisfaction, and service quality

Other study (Zeng, Fanbin; Liu, Ruini. et, all, 2012) is examined Specialization of Pay Tv and Market segmentation. In this study, focused to segmentation on two categories; First category of object is target audiences that the channel aims in particular to a group of spectators, such as lawyers. and doctors. Take televisions channels which are watched by childrens who aged form 2 years old to 14 in France. It was started in 1990. It takes 13 hours every day to broadcast only the children's programs, founded a children's club, to interviews with experts to fix the problem put forward by their parents or children, and organizes travels of children.

Second category of object is the category of content, that it hones in on just one special type of television programs, for instance, existing channels such as diversity economics, sports and shows, however ultimately these categories will go deep into the segmented trend of television channel and evolve into more specific, like football, basketball and sports channels of go.

3.DATA MINING AND CLUSTERING ALGORITHMS

3.1 DATA MINING AND USAGE AREAS OF DATA MINING

Data are numbers, text or any facts that can be processed. In literature, there are a variety of terms that substitute for data including case, pattern, record, vector, sample, event, pattern, point, observation and so on. Data can be different types such as qualitative or quantitative.

It is possible to say that data mining is all about learning. Data mining algorithm tries to learn the system under analyze. Learning operation is done from collection of data, datasets. A data set can often be viewed as collection of data objects. Data sets contain raw information of the system and the conditions under which system acts like that. Today, companies are storing growing amounts of data which has huge size in different databases and distinct formats.

3.1.1 Definition Of Data Mining?

Data mining is an algorithm for extracting patterns from data. The technologies are very advanced for collecting data and producing. In the past, reaching and recording of information are difficult but analysis and interpretation of this little data are easier and faster. However, collection and storage the information are relatively easier but converting this huge data stack into useful knowledge has become more complicate in recent conditions. Hence, lack of data is no longer a problem; the inability to produce useful information from data is main concern at the current situation. In the light of the developments and needs, data mining has show to benefit from increasing information to better understand the huge amount of data. Data mining is the process of analyzing of large data sets to find unsuspected and unknown relationships and discovering useful information in novel ways that are both understandable and useful to the data. Data mining also provides capabilities to forecast the result of a future observation.

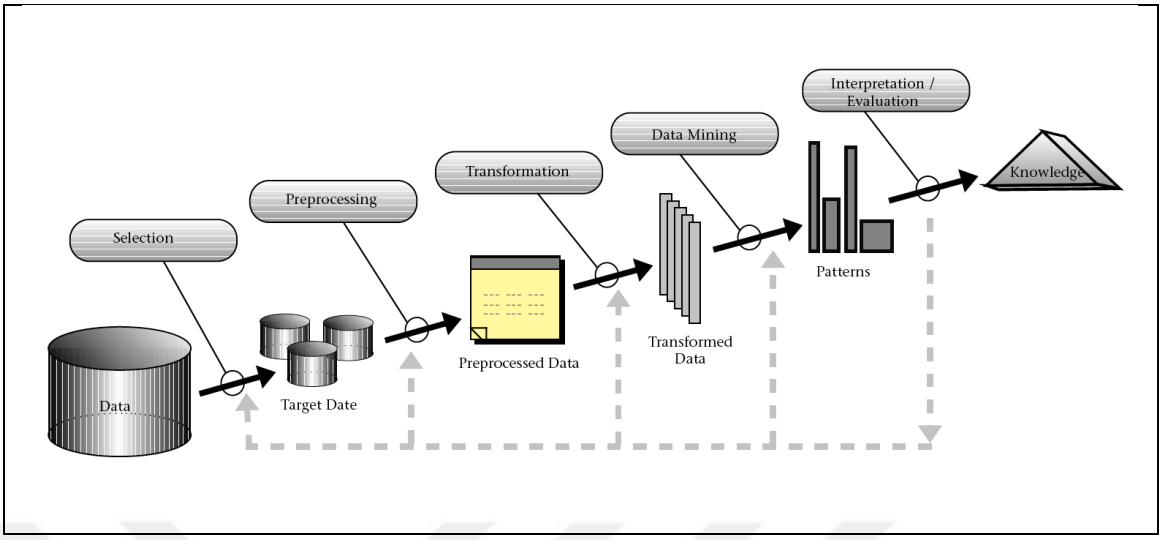
Firms can extract beneficial information from raw data using data mining. Data mining enables to follow customers` behaviors and firms can offer special services for customers so they can increase their benefits. Purpose of data mining is answering questions that cannot be answered through simple query and reporting techniques.

The size of data produced is twice rising every two years. Raw data makes up 90 percent of the digital universe. However more raw data does not mean more information or knowledge. Data mining enables to filter through all repetitive noise, understand kısacawhat is meaning of data and then make new format for meaningfull pattern.

KDD has five steps and these steps are interpretation/evolution of data, selection of data, data transformation, preprocessing of data, data mining.

a) **Preprocessing of Data**: In this stage data are cleaned such as noise and irregularities.
b) **Selection of Data:** In this stage, some data preparations techniques are performed for data. Generalization of data, determination of the data type, visualization of data, data presentation.
c) **Data Transformation:** In this stage data is transmitted to suitable format.
d) **Data Mining:** In this stage, transformed data extract with the support of previous stages.
e) **Evolution / Interpretation:** In this stage, check and discuss results of process.

Figure 3.1: Flow Diagram For Data Process



The Data Mining Process

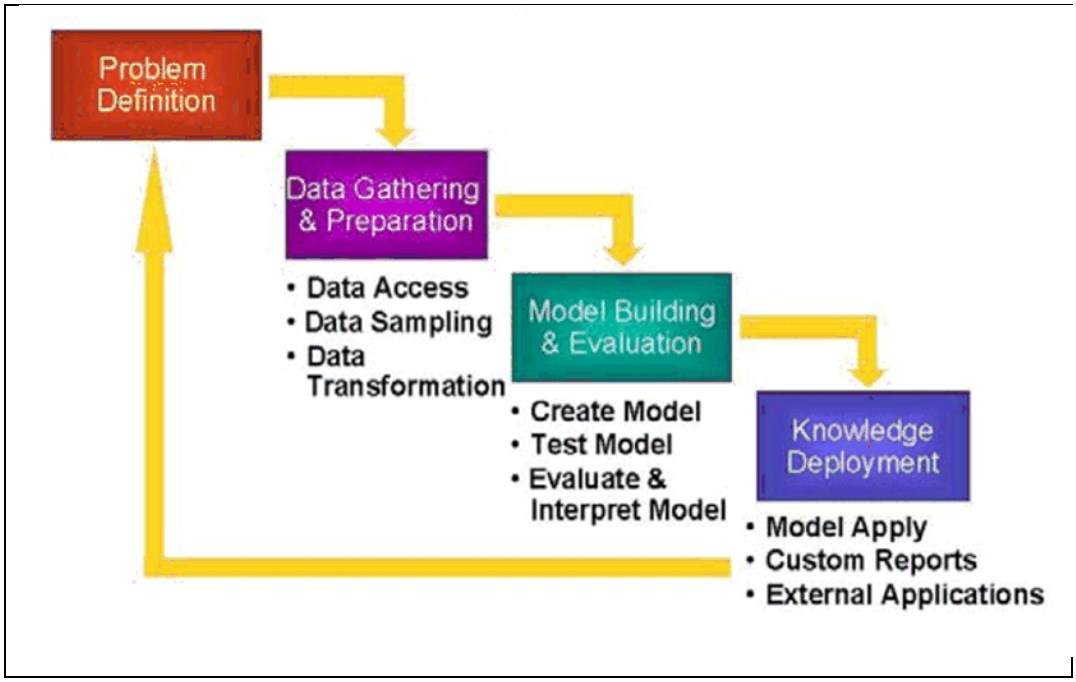
a) Problem definition

b) Data gathering and preparation

c) Model building and evalutaion

d) Knowledge deployment

Figure 3.2: Processes For Data Mining



In complex and huge systems of IT has been including separate analytical and transaction systems, data mining techniques help to link between analytical and transaction systems. Data mining tools analyze relation and patterns in stored raw data.

Data mining includes of five major elements:

a) Extract, transform to specific format to process

b) Manage and store the data.

c) Provide data access to data analysts.

d) Analyze the data by data mining software.

e) Show the data in a meaning format.

3.1.2 Advantages Of Data Mining

The data accumulated from past transactions can be processed into knowledge for future trends. For instance, sales data can be analyzed for deciding to promotional efforts to provide knowledge about consumer buying behavior. Thus, department of marketing could determine which items are most efficient to promotional efforts.

Data mining techniques are used by companies in almost every industry. It provides these companies to determine relationships among customer interactions, customer demographics, product interactions or external factors such as economic indicators, economic performance, competition and market share of others firms. These techniques enable them to decide the new strategy for marketing and provide greater customer satisfaction. What is more, These techniques enable firms to efficient summary information about transactional data.

For businesses, data mining is used to help to make better business decisions and have better strategies. Data mining can enable to produce accurately predict customer loyalty and better marketing stratgies. Some usage areas of data mining; shopping basket analysis, fraud detection, product association minings, customer churn, customer loyalty, market segmentation, and trend analysis

Every day, more and more data is produced and stored. Nearly every interaction produce a data that someone is storing and capturing. Of course, the cause of this stituation is development of internet infrastructure and the increase in the number of devices connected to the internet.

3.2 TYPES OF ATTRIBUTE

An attribute is a property or characteristic of an instance that may vary, either from one instance to another or one time from to another Data objects are described by a number of attributes. Other names for an attribute are variable, characteristic, field, feature, or dimension.

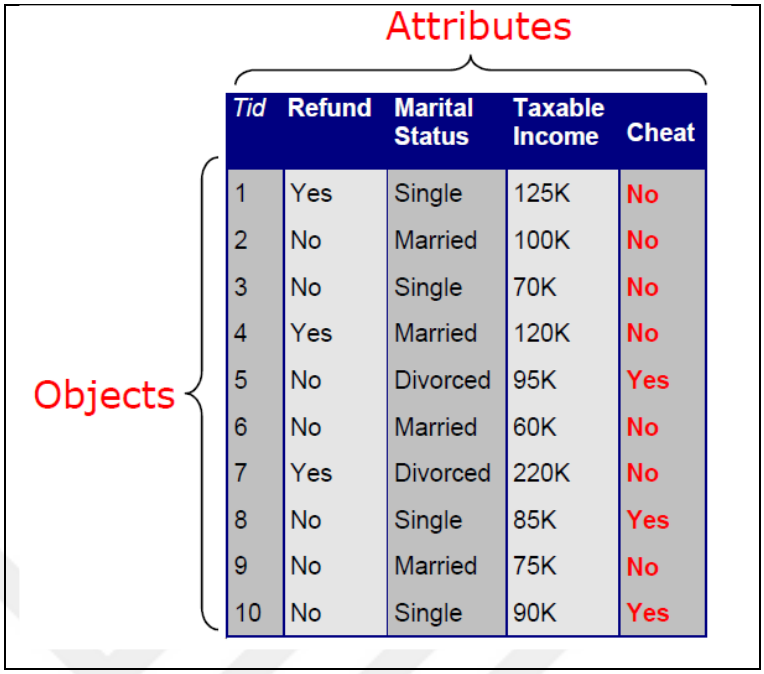
Each individual, independent instance is characterized by its several attribute values. Generally, four types of attributes are defined such as: nominal, ordinal, interval, and ratio. While ordinal and nominal attributes are interval, categorical and ratio attributes are numeric. Types of attribute are summarized in table 3.1.

Table 3.3 : Attribute Types

Attribute Type		Description	Examples
Categorical	Nominal	The values of nominal attribute are just different names (=, ≠)	Employee ID numbers, Zip Codes, Gender
	Ordinal	The values of an ordinal attribute provide enough information to order objects. (<,>)	Hardness of mineral, grades, {good, better, best}
Numeric	Interval	For interval attributes, the difference between values are meaningful. (+,-)	Calendar dates, temperature in Celsius or Fahrenheit
	Ratio	For ratio variables, both differences and ratios are meaningful. (*,/)	Counts, age, mass, length, electrical current

Generally, data is organized in the matrix form an instances and attributes is constructed in row and column. An example is presented in Figure 3.3

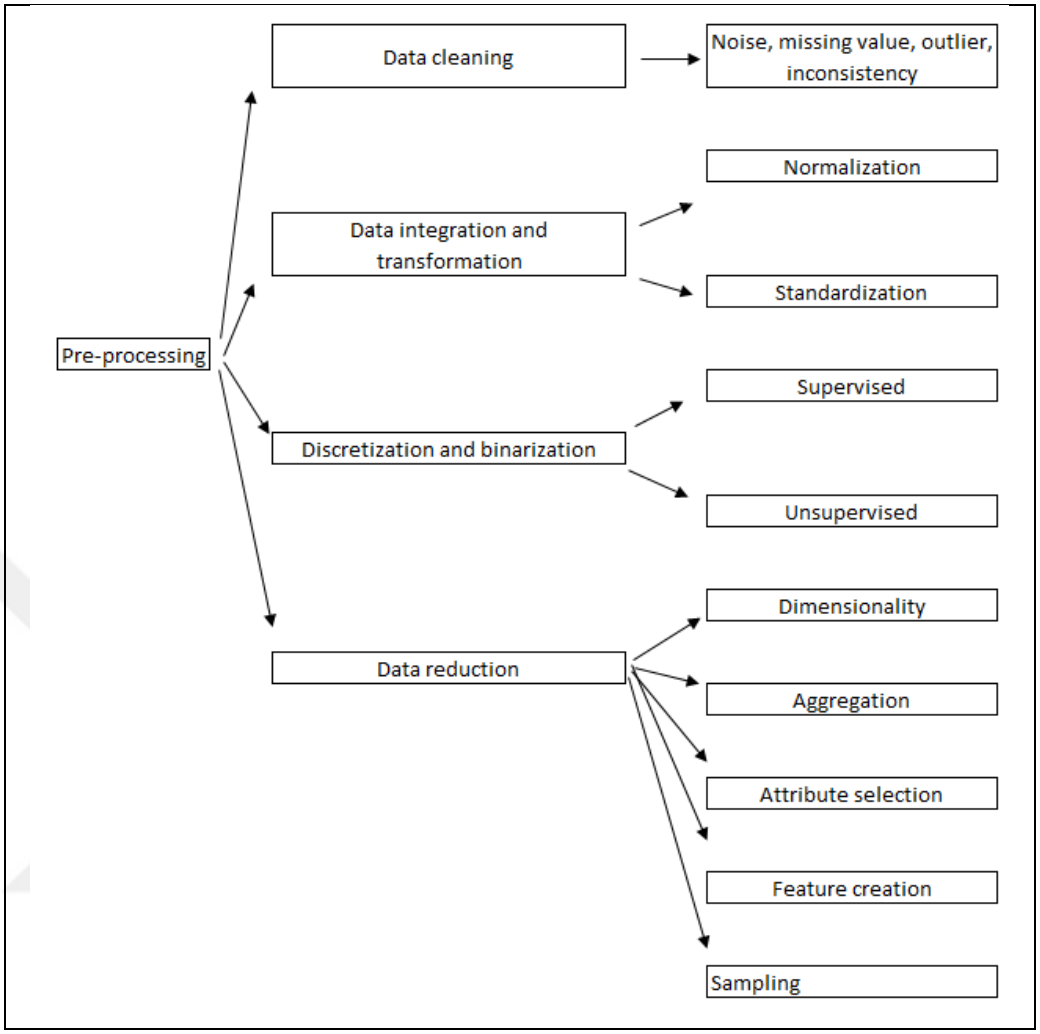
Figure 3.3 : Data Example



3.3 DATA PREPROCESSING

There are a lot of techniques of data preprocessing. One of these is data cleaning. It is used to fix incompatible and delete noise in the data. The other one of techniques is data integration. It merges data from different data sources into a single data center. Other technique is data reduction, it can reduce size of data by eliminating redundant data, aggregating or clustering. Last of techniques is data transformations, it can provide better efficiency and accuracy of data mining algorithms, such as normalization. These data processing techniques can improve the overall results of data mining. Figure 3.4 presents applications of Data Preprocessing Applications.

Figure 3.4 : Data Preprocessing Applications



3.3.1 Data Cleaning

In most of time, There may be missing values in the datasets. Data cleaning techniques can fill in missing values, identifying outliers for noisy attributes, and fix inconsistencies values in the dataset.

One of the common problems is missing values, and these values can break data patterns. In generally, data mining methods remove records which has containing missing values or substitute missing values with the mean, or assume missing values from existing values, or neglect the missing values.

Missing Values Replacement Methods:

a) Ignore the records which have missing values.
b) Replace them with the mean value or the most frequent value.
c) Replace them with a global constant.
d) Use modeling techniques such as nearest neighbors, EM algorithm, Bayes' rule or decision tr
e) Fill in missing values with manually based on your domain knowledge.

Noise is a variance or random error in a measured variable. For instance, for numeric attribute such as price.

Noisy Data Replacement Methods;

a) Binning methods
b) Combined human and computer inspection.
c) Clustering
d) Regression

In the datasets there may be some inconsistencies values. These inconsistencies values can be fixed manually using external references. Some data mining tools can also be used to discover the contravention of known data constraints to block producing incosistenies values.

.

3.3.2 Integration of Data

Data integration is defined as combining data from various data sources into a single data center, For instances, data warehousing. These data sources can contain data cubes, various databases or at files.

There are many problems to about during data integration. For instance, how can be sure that customer id which is in one database is refer to the same entity in another database? Databases systems generally have metadata which can be used to avoid errors for schema integration. Another important problem is redundancy. Correlation anaysis

helps to detect some redundancy cases. This analysis measure how one attribute implies the other for given two attributes. The correlation analysis between x and y can be measured by

$$\frac{P(X \cap Y)}{P(Y)P(Y)} \tag{3.1}$$

X and Y are positively correlated if result value of equation is greater than 1. Attribute which has higher value implies the other lower attributes. Thus, a high value shows that X (or Y) can be removed for redundancy. If result of equation is equal to 1, then there is no correlation between X and Y, they are independent. If result of equation is less than 1, then this shows that each attribute discourages the other. X and Y are negatively correlated.

3.3.3 Transformation of Data

In some stituations, Data may not put data mining techniques with exactly values. When averages and variances of variables have major differences between them, variables which have bigger averages and variance have more impact on results. Therefore we need some transformations for variables to bring the same scale.

3.3.3.1 Min-Max Normalization

The goal of this technique is scaled all values within 0.0 to 1.0 range. Therefore after this technique is processed, largest value in set is 1 and smallest value in set is 0. These numbers except the numbers are assigned values based on the range they are located.
Min-Max normalization is defined as;

$$\text{Xnormal} = \frac{X - Xmin}{Xmax - Xmin} \tag{3.2}$$

This technique preserves the relationships among the original attributes. In this study we use Min-Max Normalization for attributes which have very huge scale for scaling of data to fit into a specific range.

3.3.3.2 Z-Score Normalization

This technique is based on the standard deviation and mean of values. It uses mean and standart deviation for converting new values.
Z-Score Normalization is defined as;

$$X^* = \frac{X - \mu}{\sigma x} \tag{3.3}$$

As seen from above equation (3.3), for standardizating of a value, The average value (μ) on the distance from the standard deviation (σ) is divided. The following formula is used to calculate the standard deviation (σ).

$$\sigma x = \sqrt{\frac{\sum_{i=1}^{n}(Xi-\mu)^2}{n-1}} \tag{3.4}$$

And The following formula is used to find the mean value (μ)

$$\mu = \frac{1}{n}\sum_{i=1}^{n} Xi \tag{3.5}$$

3.3.4 Reduction of Data

Complex data mining techniques on huge amounts of data can take a very long time. Making analysis in this conditions is very inefficient. If you are sure to come same result, when reduce the size of dataset, you can reduce data to much smaller volume. It enables to ensure to be feasible. Data reduction methods include the following;

- **a)** Data cube aggregation
- **b)** Dimension reduction
- **c)** Data compression
- **d)** Numerosity reduction
- **e)** Discretization and concept hierarchy generation.

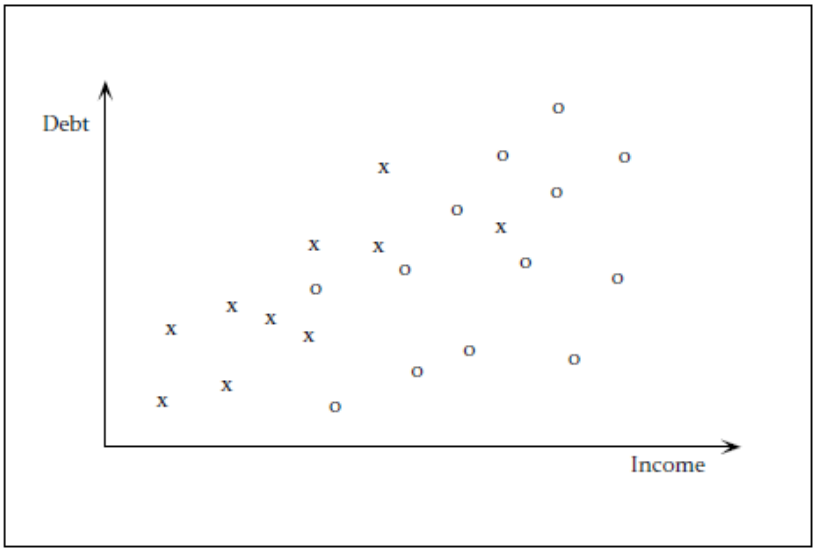
3.4 DATA MINING METHODS

Data mining methods have two forms for extracting models. These forms are prediction and description. In prediction forms, using some variables which are known earlier to predict unknown or future values of other variable. Prediction form is based on the relationship between a thing that you can know and a thing you need to predict. In description forms, using some variables which are known earlier to define models for deciding. The aims of description and prediction can be applied with using data-mining methods.

3.4.1 Classification

Classification techniques can be used to classify new cases into these pre-determined groups. Firstly, algorithm learn from a large data set of pre-classified groups then algorithm can detect exact rules about structure of data using differences between items in each group and finally, apply these rules to new classification problems for classifying. For example; spam mail filters are a good example for classification. This process identify mail as spam by noticing differences in mail text and analyzing words then classify incoming mails according to these rules.

Figure 3.5 : Graph Of Dataset for Two Classes



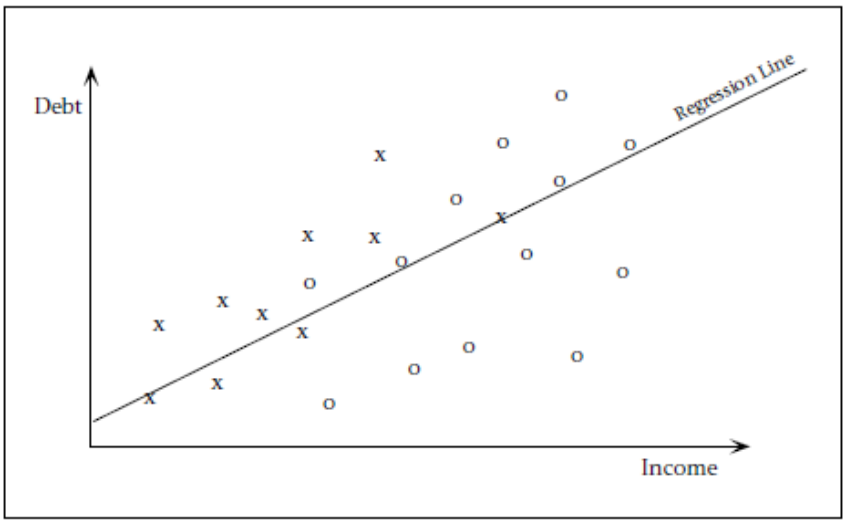
A Simple Data Set with Two Classes Used for Illustrative Purposes.

3.4.2 Regression

Data mining methods can be used to construct predictive models based on many attributes. Regression is one of the data mining methods which is used to analyze the relationship between independent and dependent variables.

In regression model, the independent variables are the variables which you base your prediction on, whereas the dependent variable is the one whose values you want to predict.

Figure 3.6: Simple Linear Regression Graph

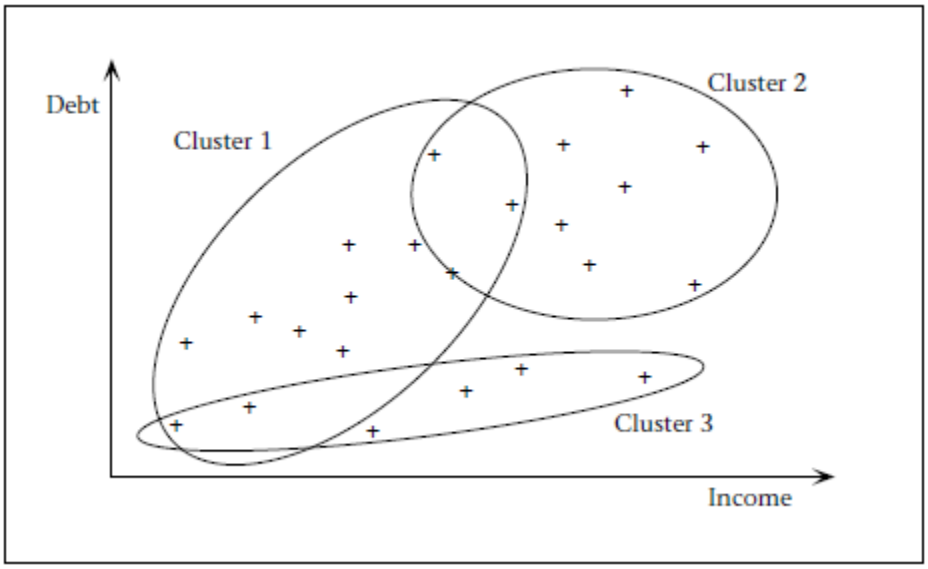


3.4.3 Clustering

Clustering divides dataset into groups which are contained similar objects. Each group consists of objects which are similar between themselves and dissimilar to others groups` objects and each group is called cluster

After results of cluster analysis, if datasets are represented by fewer clusters, you lose details of result but you can achieve simplification. Clustering analysis is commonly used in applications such as image processing, data analysis tools.

Figure 3.7: Simple Clustering of the three clusters



Farley and Raftery (1998) divide the clustering methods into hierarchical and partitioning methods. However, the methods are divided into additional three groups according to Han and Kamber (2001). These are gridbased methods, model-based, density-based methods, clustering. These are;

a) Hierarchial Methods
b) Grid-based Methods
c) Partition-Based Methods
d) Density-based Methods
e) Model-Based Methods

3.4.3.1 K Means Algorithm

K-means is one of the unsupervised learning algorithms that is most commonly used and it`s implementation is very easy. K-means clustering method to a data set consisting of N number of data objects as input parameters K units is to partition the cluster. The goal is realized partitioning process obtained at the end of the cluster, and clusters of maximum similarity between the intra-cluster similarity is to ensure that minimum is..

Implementation is easy. Large-scale data sets can be quickly and effectively. Therefore in this study, k-means clustering algorithm is used for clustering.

The main idea of this algorithm is to define k cluster centers (centroid) then compute the distance between cluster centers and each data item. Assign the data item the center whose distance from the cluster center is the smallest of all the cluster centers. After this step, recalculate the new cluster center and recalculate the distance between each data point and new cluster centers. The algorithm continues to repeat until to no change for centers.

The algorithm essentially consists 4 phases;

a) Determination of cluster centers

b) Clustering data that is outside the center according to distance

c) Determination of new centers according to clustering (or the old center of shifting to the new center)

d) Until the finding new centers which become fixed (stable state), repeating steps 2 and 3.

Here, the two most important aims;

a) Data in cluster must be similar to each other

b) Clusters should different each other as much as possible

Advantages of K-Means Algorithm;

a) It is easy to use and it works very quickly for large dataset.

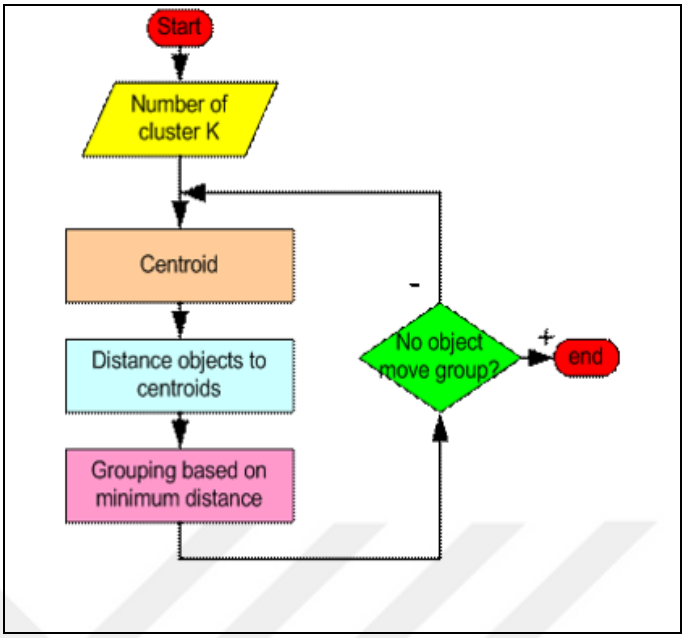
b) In calculating the number of too much data, if number of cluster is small, calculations are faster than hierarchical clustering.

Disadvantages of K-Means Algorithm;

a) K-means algorithm can not be determined the number of clusters k. therefore place a trial and error process until the appropriate number of

b) It is sensitive to noisy data. Clustering is included in the data.

Figure 3.8 : K Means Algorithm



Assume $x^{(1)}, x^{(2)}, ..., x^{(m)}$ are elements of training set and we want to group the data into new clusters. $X(i) \in R^{(n)}$ for each data of groups. Goal of this algroithm is to predict k centroids and a label $c^{(i)}$ for each data. The k-means algorithm is as follows;

Figure 3.9 Pseudocode of K-Mean Algorithm

1. Initialize **cluster centroids** $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$ randomly.
2. Repeat until convergence: {

   For every i, set
   $$c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2.$$

   For each j, set
   $$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}.$$

   }

Source: http://stanford.edu/~cpiech/cs221/handouts/kmeans.html

The notation ||x-y|| means euclidean distance between vectors x and y. In the next chapter, this issue will be discussed.

3.4.3.2 Euclidean Distance

Euclidean distance is the linear distance between two points.

For one dimension, the distance between two points is the absolute value of their numerical difference. If p and q are two points, then the distance between them is given by;

$$\|p\text{-}q\| = \sqrt{(p-y)^2} \tag{3.6}$$

For two dimensipn, assume p=(p_1,p_2) are points in the euclidean plane, then the distance is given by;

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2} \tag{3.7}$$

For n dimension, assume n-dimensional space, the distance is given by;

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} \tag{3.8}$$

3.4.4.4 Association Rules

Association rules is an important place for the analysis of large data and It is one of the main work area of data mining. What is more association rules, is the technique most studies and widely applicale. For using association rule mining techniques, all data must be categorical.

The purpose of the Association Rules model, the realization of the pieces of data (co-occurrence) is to uncover all of the relationships. Classic application area of association rules mining is the market basket data analysis. Market basket analysis is derived from the relationship between the products that customers receive.

Cheese => Bread [Support=%10, Confidence=%75]

The rules above means 10% of all customer has with cheese and bread (Support) and 75% of customers who buy cheese also buy bread (Confidence).

What is more, association rules do not take into care the order of mining products. Support value of rule indicates the probability of transactions. support the value of the rule is calculated by the following equation. Here “n” shows the total number of transaction.

Figure 3.10 : Calculatipn Of Support Number

$$support = \frac{(X \cup Y).count}{n} \tag{3.9}$$

The confidence value of a rule is the proportion of the transactions that contains X which also contains Y. It is defined as;

$$conf = \frac{support(X\ U\ Y)}{support(X)} \tag{3.10}$$

In this study, We use assocation rule mining to determine the attributes which are used in clustering methods. Apriori algorithm is the best known algorithm for association rules. Therefore We use apriori alogrithm in this study. In next chapter, We will provide information about this algorithm.

3.4.4.5 Apriori Algorithm

Apriori algorithm is developed in 1994 by Agrawal and Srikant. This algorithm is used for common item association rule learning over transaction data. It is the most well known algorithm in the association rule mining algorithms and the most widely used algorithm in data mining area.

Apriori uses a “bottom up” approach, where common subsets are extended one item at a time. Apriori designed to perform on database containing transsactions. For example; items of market basket or interactions of customers on website.

Steps of Apriori Algorithm;

a) Determination of the minimum support number and minimum confidence value.
b) Findings support values for each element in the clusters.
c) Removing of items which have lower support value than minimum support value.
d) Create new associations with using the results
e) Removing the cluster elements which have lower support value than minimum support value.
f) Until the no new rules repeating steps d and e.

The pseudo code for the apriori algorithm is given below and this pseudo code supports threshold of ϵ. T is a multiset. k. At each step, the algorithm generates the candidate sets from the large item sets of the prior iterate, observing the descending conclusion lemma. C_k is the candidate set for level . $count[c]$ accesses a field of the data structure that represents candidate set c, which is initially assumed to be zero.

Figure 3.11 Pseudo Code of The Apriori Algorithm

```
Apriori(T, ε)
    L_1 ← {large 1 − itemsets}
    k ← 2
    while L_{k−1} ≠ ∅
        C_k ← {a ∪ {b} | a ∈ L_{k−1} ∧ b ∉ a} − {c | {s | s ⊆ c ∧ |s| = k − 1} ⊈ L_{k−1}}
        for transactions t ∈ T
            C_t ← {c | c ∈ C_k ∧ c ⊆ t}
            for candidates c ∈ C_t
                count[c] ← count[c] + 1
        L_k ← {c | c ∈ C_k ∧ count[c] ≥ ε}
        k ← k + 1
    return ⋃_k L_k
```

Source: ***https://en.wikipedia.org/wiki/Apriori_algorithm***

3.5 DEFINITION OF CUSTOMER SEGMENTATION

Customer segmentation is the process of splitting the customer data into different and internally similar clusters for developing different marketing strategies according to their characteristics. It enables to compaines to use resources efficiently and the maximization of cross and up selling oppotunities. There are a lot of distinct segmentation types based on the attributes of segmentation or techniques of data mining. one of the these segmentation types is behavioral segmentation. According to behavioral segmentation, customers are clustered by usage characteristics and behavioral. it is created with business flows and rules. This concept has some difficulty and behavioral segments can be created by data mining. Clustering alogrithms have many functionality for analyzing behavioral data and identifying the natural groupings of customers. In addition, clustering algorithms provides a solution organized on noticed data patterns and it provides the data mining models which are built well. Clustering algorithms can discover new groups with different profiles and characteristics and lead to various segmentation schemes with business flows and values. What is more, data mining can also be used for the creation of customer segmentation accorind to expceted value of the customers.

4. EXPERIMENTAL SETUP

In this section, we reviewed data model design, structure of data and data preparation steps in order to understand how our algorithm works or what is our algorithm success.

4.1 STRUCTURE OF DATA

Dataset that will be used for segmentation analysis is constructed by using case company`s customer master data.Thesis database is storing information about customer of digital broadcast company. To understand results of works we should get to know structure of data.So in this chapter data which is used is described.

Account Number : It is unique value for a customer.It is numeric value.it is used for determining the customer.

Last Year Billing : This field store data of annual bill for the last one year value.These values have very huge scale so we make normalization for scaling of data to fit into a specific range. In below table present descriptive analysis for billing data. These data gives us insight about purchased products.Table 4.1 shows information about statiscal properties of these field.

Table 4.1 Billing Table

		LastYearBilling	Valid N (listwise)
Statistic	N	100000	100000
	Range	14574,1200000000	
	Minimum	1,3000000000	
	Maximum	14575,4200000000	
	Mean	756,376578600001	
	Std. Deviation	506,9706375276651	
	Variance	257019,227	
Std. Error	Mean	1,6031819214151	

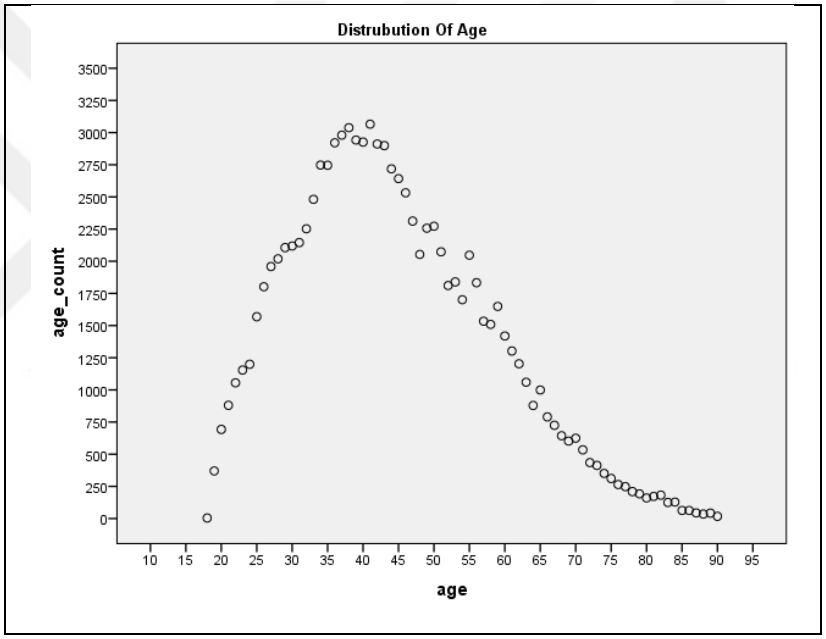
Age : Age of customer is stored in this field.It is a numeric attribute. Table 4.2 presents descriptive statistic for age attribute.

Table 4.2: Age Descriptive Table

	Mean	Maximum	Minimum	Range	Standard Error of Mean	Standard Deviation	Valid N
AGE	44	90	18	72	0	14	100000

As is seen from Figure 4.1, The majority of data is from 30 and 55.

Figure 4.1: Distribution Of Age



Membership Age : This field shows time of between contract date and now with freeze and suspend periods.So it gives us total membership age of customer. Table 4.3 presents descriptive statics.

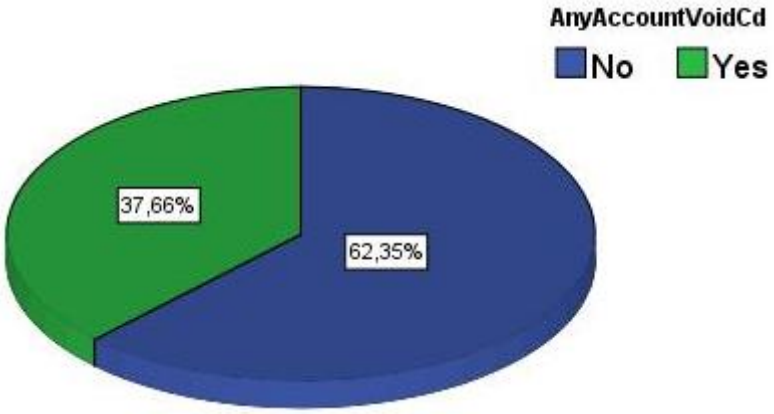
Table 4.3 Age Of Membership Table

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
Membership_Age	100000	14	1	15	6,30	,015	4,888	23,893
Valid N (listwise)	100000							

Account Void Count : This field stores count of void request in last a year. It is a important information for analyzing pleasure of customer. Similarly, Void attribute is used as binary attribute. If account has at least one void order then attribute is set 1, otherwise 0. Figure 4.2 presents of distribution.

Figure 4.2 Graph of Void Order Distribution



Content Fields : Content fields are binary value.If value is 0 then member can not receive broadcast for this content, else member can receive broadcast.

Content fields are;

Sport : It contains football matches in Turkey Super League,

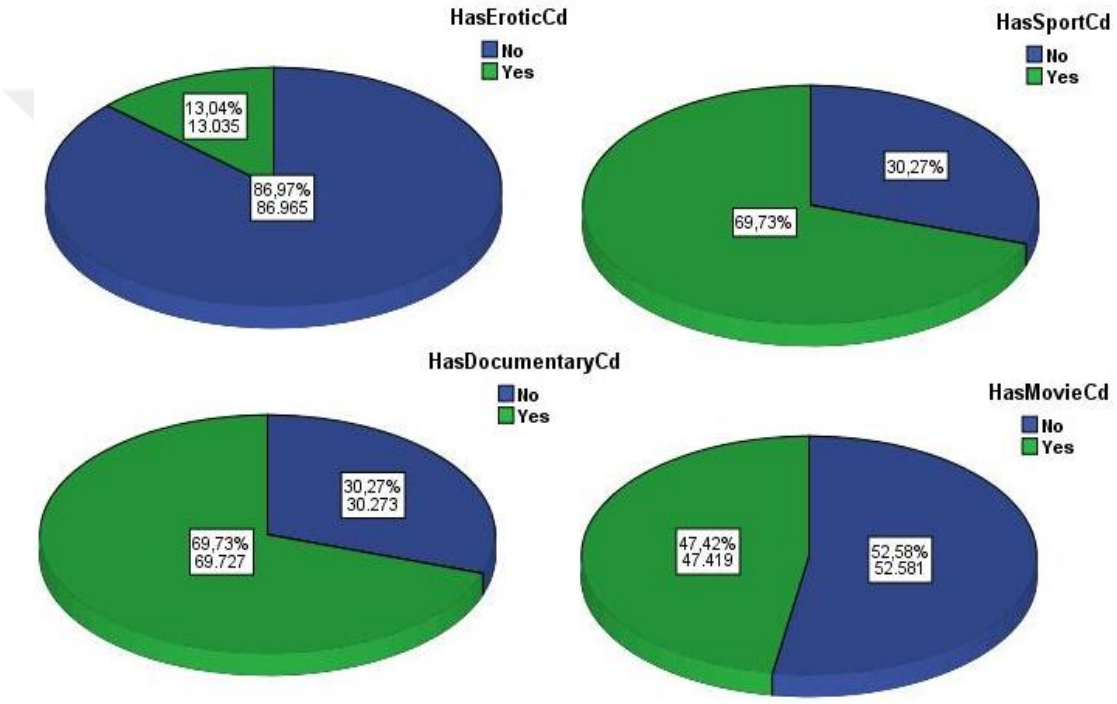
Erotic : It contains erotic content,

Documentary : It contains documentary content,

Movie : It contains movie content,

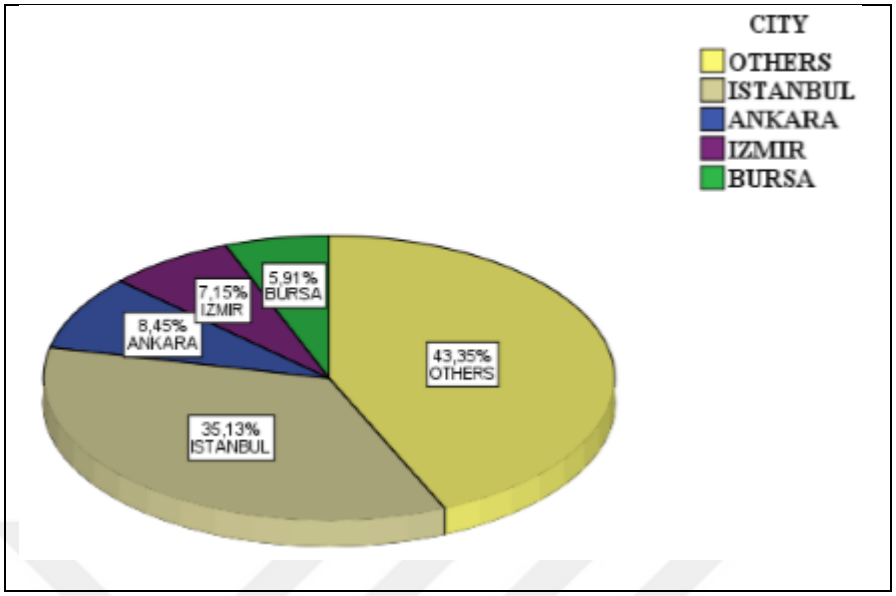
This attribute is one of the most important attributes in our dataset.It gives us insight about habits of watching television for accounts. We have found this content data by reviewing the authorized channels of packets of accounts. Figure 4.3 presents of content distribution.

Figure 4.3 Graph of Content Distribution



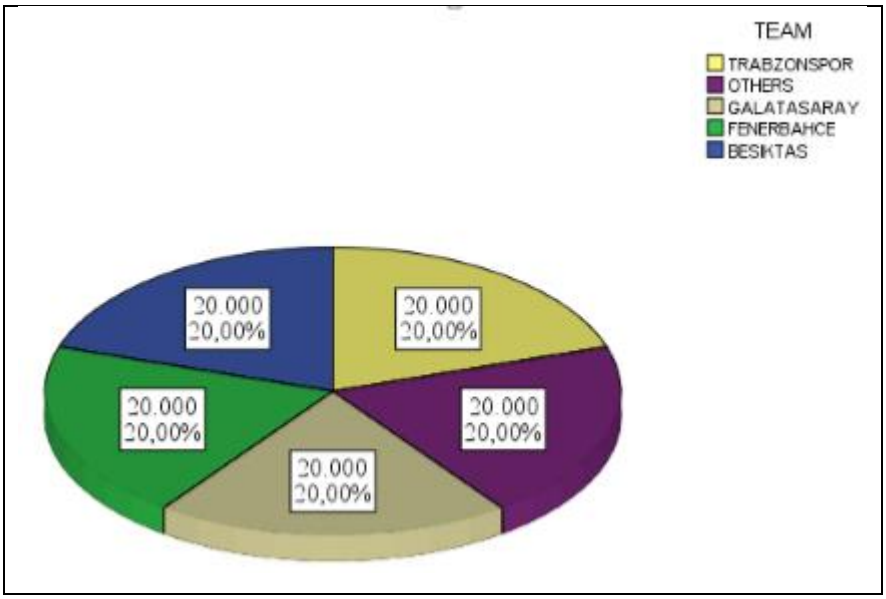
Location City Field : In this field, values are stored as binary value. Location of member is stored in this field. five cities which have most members are chosen and other cities are showed in Anatolian Cities group and each city is shown in a separate column. City fields are Istanbul, Izmir, Ankara, Izmir, Bursa and Other Cities. Account of the city that is located in is set 1, others are set 0. Figure 4.5 presents of city distrubution.

Figure 4.4 Graph of City Distribution



Favourite Team Field : In this field, values are stored as binary value. Four teams which have most fans are chosen and other teams are showed in Other Teams column. Teams are Fenerbahce, Galatasaray, Besiktas, Trabzon and Other Teams. Each team is shown seperate column. In order to We can see the relationship with each other, equal number of data for Fenerbahce, Galatasatay, Besiktas, Trabzonspor and others are taken. Figure 4.5 presents distrubution of teams.

Figure 4.5 Distribution Of Teams



4.2 DATA PREPARATION

Data which is used in my project has many different attribute. There is a huge scale difference between some attributes like last year billing, favourite team and watched contents etc. Before data is used, numeric attributes are transformed into same numeric scale or grouping according to the data distrubution. The goal of this process is to make an total set of values which have a specific property.

Data which is used in my project is taken from company which is I worked and then variety of techniques are applied to prepare data. Throughout this process data is stored in MySQL server. Using RDBMS has accelerated so my job because different attributes for different scenarios are used. So MySQL has a lot of features to manipulate data easily.

After that apply techniques are started to prepare data. Custom application is written to manipulate data. .NET enviroments are used when writing code and the C# language is used. This application is basically do the following; the seperating data according to the specified values, finding max, min, std, variance values, calculating normalization values according to the specified algorithms, the seperating products are purchased by members according to the contents, (For instance; Erotic, Film, Documentary, Movie) , saving the results file, creating arff file for WEKA and creating csv file for SPPS.

After data prepared, analysis is started for data. SPSS 23 and WEKA are used for this process. Especially for clustering algorithms, WEKA is better choice. Because WEKA has more clustering algorithms to implement than SPSS. Also SPSS has many features useful for me. all tables and figures are created using SPSS and It gives me more detailed statistical information.

Different ways are tried to prepare data for each attribute. Because each attribute has different features. These ways are listed below.

Last Year Billing : These values have very huge scale so firstly normalization is made for scaling of data to fit into a specific range and table is created as Table 4.4 Min/max normalization is used for this transformation.

Table 4.4 Normalized Values of Billing

AccountNumber	LastYearBilling
8799980	1
631679	0,819344839
3471591	0,678200811
565440	0,595358759
1019322	0,563108442
1003805	0,52203838
2675756	0,490021353
2902620	0,476283302
30984	0,46865128

But using the data in this format is not efficient for segmentation. Because this form is very difficult to interpret the data. So elements which is in the set are separated into groups. Also the maximum and minimum values are determined for each group. Table 4.5 presents distribution of groups.

Table 4.5 Distribution of Billing Groups.

ID	Min Value	Max Value
1	1	868
2	869	1827
3	1828	4203
4	4249	14576

To determine max and min values for groups, max and min values are found for each cluster. Also K-means cluster algorithm is used to separate into cluster. Table 4.6 presents centers of clusters.

Table 4.6 Cluster Centers

Final Cluster Centers

	Cluster			
	1	2	3	4
LAST_YEAR_BILLING	505,46	1175,36	2316,89	5423,46

And Table 4.7 presents distribution of cases in each cluster

Table 4.7 Number Of Item For Each Cluster

Number of Cases in each Cluster		
Cluster	1	69299,000
	2	26963,000
	3	3655,000
	4	83,000
Valid		100000,000
Missing		,000

Age : This attribute with values between 18 and 90. In order to separate into groups, K-means algorithm is used. Firstly 5 clusters are created, but as is seen from Table 4.8, clusters` cases are not distrubuted uniformly. Especially cluster 3 and cluster 4 has less items then other clusters.

Table 4.8 Number of Cases in Each Cluster

Number of Cases in each Cluster		
Cluster	1	19073,000
	2	28319,000
	3	17925,000
	4	6584,000
	5	28099,000
Valid		100000,000
Missing		,000

As a solution, cluster 3 and cluster 4 are combined and new cluster is entitled cluster 4. Also maximum and minimum values of each groups is shown in Table 4.9.

Table 4.9 Max and Min Values of Age Groups

ID	Min Value	Max Value
1	18	31
2	32	41
3	42	53
4	54	90

Membership Age : Similar to age attribute, for membership age, K-means cluster algorithm is used for this attribute. Table 4.10 presents number of cases in each cluster.

Table 4.10 Number of cases in each Cluster

Number of Cases in each Cluster

Cluster	1	25154,000
	2	39097,000
	3	35749,000
Valid		100000,000
Missing		,000

As is seen from Table 4.10, clusters are separated uniformly. Groups are created using these clusters. In Table 4.11, Groups` max and min values are shown.

Table 4.11 Max and Min Values of Membership`s Groups

ID	Min Value	Max Value
1	1	4
2	5	10
3	11	15

Also Table 4.12 presents centers of clusters.

Table 4.12 Cluster Centers

Final Cluster Centers

	Cluster		
	1	2	3
Membership_Age	13	1	7

Any Account Void : Values are distrubuted in huge range so it was needed to apply the technique of data preparation.

Firstly the K-means alogrithm is applied to find centers of clusters thus they are grouped according to these centers. But clusters did not have uniformly distrubution so this attribute is used as binary values. It can have 0 or 1 values and shows that members have any account void order or not.

In this process, some attributes are ignored for breaking results. For example; there are also attributes such as Gender, Complaint Count, Suspend Count, Satellite Type attributes but these attributes do not give us the results properly.

5.FINDINGS

Apriori algorithm is used for creating association rules on WEKA. In this chapter rules which are found are examined.

Association Rules for parameters; Support Min Threshold = 0.05 and Support Max Threshold = 0.2 and Confidence Treshold = 0.9 and Count of Item = 100K (For all data)

a) 19,86 percent of members; have no any account void order, live in other cities and watch documentary and sport contents.

b) 19,79 percent of members; live in other cities, watch documentary and sport contents and do not watch movie contents

c) 19,67 percent of members; live in other cities, have (1-868) annual billing, watch documentary and sport contents and do not watch erotic contents

d) 19,63 percent of members; have (5-10) membership age, watch documentary and sport contents and do not watch erotic contents

e) 19,51 percent of members; have no any account void order, have (1-868) annual billing, watch documentary and sport contents and do not watch movie contents

f) 19,5 percent of members; have no any account void order, watch movie,sport and documentary contents.

g) 19,48 percent of members; have (42-53) age , watch documentary and sport contents

h) 19,26 percent of members; have (1-4) membership age, watch documentary and sport contents, do not watch erotic and movie contents

i) 19,04 percent of members; have (32-41) age, watch documentary and sport contents

j) 19,04 percent of members; have no any account void order, have (1-4) membership age, watch sport and documentary contents, do not watch erotic contents, have (1-868) annual billing

k) 18,96 percent of members; do not watch erotic contents, watch sport and documentary contents, have (868-1827) annual billing

l) 18,82 percent of members; have no any account void order, do not watch documentary and sport contents.

m) 18,69 percent of members; have (1-4) membership age, watch documentary and sport contents, do not watch movie contents, have (1-868) annual billing.

n) 18,67 percent of members; live in Other city, watch documentary and sport contents, do not watch movie and erotic contents

o) 18,36 percent of members; have no any account void order, watch documentary and sport contents, do not watch erotic and movie contents, have (1-868) annual billing

p) 17,53 percent of members; have (1-4) membership age, watch sport and documentary contents, do not watch erotic and movie contents, have (1-868) annual billing.

q) 17,32 percent of members; watch movie contents, do not watch sport and documentary contents.

r) 17,27 percent of members; do not have any account void order, do not watch documentary,erotic and movie contents.

s) 17,23 percent of members; do not watch documentary and sport contents, watch movie contents

t) 17,05 percent of members do not have any account void order, live in othercity, watch documentary and sport contents, do not watch erotic contents.

u) 17,05 percent of members; have (1-4) membership age, watch documentary and sport contents, do not watch movie contents.

v) 15,25 percent of members; watch sport, movie and documentary contents and have (868-1827) annual billing.

w) 14,82 percent of members; watch movie, do not watch sport,documentary and erotic contents.

x) 14,46 percent of members; watch movie and sport contents, do not watch documentary and erotic contents, do not have any account void order.

y) 10,25 percent of members; watch sport,documentary and erotic contents.

z) 11,96 percent of members; watch documentary contents and do not watch erotic and movie contents, live in Istanbul

aa) 11,96 percent of members; watch sport contents, do not watch movie and erotic contents.

bb) 11,72 percent of members; watch sport, movie, documentary contents, do not watch erotic contents have (868-1827) annual billing.

cc) 10,38 percent of members; watch movie and documentary contents, have (11-15) membership age.

dd) 12,56 percent of members; watch sport and documentary contents, do not watch movie and erotic contents.

ee) 16,27 percent of members; favourite team is Fenerbahce, watch documentary and sport contents.

ff) 16,91 percent of members; favourite team is Fenerbahce, watch documentary and sport contents, have (1-4) membership age.

gg) 13,36 percent of members; watch documentary and sport contents, do not watch erotic contents.

hh) 10,25 percent of members; watch erotic,documentary and sport contents.

ii) 8,16 percent of members; watch erotic,movie,sport and documentary contents.

jj) 6,4 percent of members; watch documentary and sport contents, do not have any account void order.

Association Rules for parameters; Support Min Threshold = 0.2 and Support Max Threshold = 0.4 and Confidence Treshold = 0.9 and Count of Item = 100K (For all data)

a) 39,54 percent of members; watch documentary and sport contents, do not watch movie contents.

b) 38,04 percent of members; watch dcoumentary and sport contents, do not watch erotic contents, have (1-868) annual billing.

c) 37,44 percent of members; watch documentary and sport contents, do not watch erotic and movie contents.

d) 37,11 percent of members; do not have any account void order, watch documentary and sport contents, do not watch erotic contents.

e) 31,79 percent of members; have (1-868) annual billing, watch documentary and sport contents, do not watch movie contents.

f) 31,59 percent of members; live in OtherCity, watch documentary and sport contents.

g) 31,11 percent of member; have (1-4) mmebership age, watch sport and documentary contents.

h) 30,27 percent of members; do not watch documentary and sport contents.

i) 30,18 percent of members; watch documentary, movie and port contents

j) 30,06 percent of members; watch documentary and sport contents, do not watch movie and erotic contents.

k) 27,49 percent of members; do not watch documentary,erotic and sport contents.

l) 27,46 percent of members; do not have any account void order, have (1-868) annual billing, watch sport and documentary contents.

m) 27,24 percent of members; live in OtherCity, watch documentary and sport contents, do not watch erotic contents.

n) 27,23 percent of members; have (1-4) membership age, watch documentary and sport contents, do not watch erotic contents.

o) 26,68 percent of members; do not have any account void order, have (1-4) membership age, watch documentary and sport contents.

p) 26,2 percent of members; have at least one account void order, watch documentary and sport contents.

q) 26,11 percent of members; have (1-868) annual billing, do not watch documentary and sport contents.

r) 25,73 percent of members; have (1-4) membership age, have (1-868) annual billing, watch sport and documentary contents.

s) 30,24 percent of members; do not have any account void order, do not watch movie and erotic contents.

t) 27,1 percent of members; do not have any account void order, do not watch movie and erotic contents, have (1-868) annual billing.

u) 23,89 percent of members; do not watch sport,movie and erotic contents, have (1-868) annual billing.

Association Rules for parameters; Support Min Threshold = 0.4 and Support Max Threshold = 0.6 and Confidence Treshold = 0.8 and Count of Item = 100K (For all data)

a) 59,46 percent of members; watch sport and documentary contents, do not watch erotic contents.

b) 43,52 percent of members; watch documentary and sport contents, do not have any account void order.

c) 43,24 percent of members; watch documentary and sport contents, have (1-868) annual billing.

d) 50,12 percent of members; do not watch movie and erotic contents.

e) 44,56 percent of members; do not watch movie and erotic contents, have (1-868) annual billing.

f) 59,46 percent of members; watch sport contents but do not erotic contents.

g) 59,46 percent of members; watch sport and documentary contents, do not watch erotic contents.

Association Rules for parameters; Support Min Threshold = 0.6 and Support Max Threshold = 1.0 and Confidence Treshold = 0.7 and Count of Item = 100K (For all data)

a) 69,72 percent of members; watch sport and documentary contents

b) 61,93 percent of members; do not watch erotic contents and have (1-868) annual billing.

Association Rules for parameters; Support Min Threshold = 0.05 and Support Max Threshold = 0.1 and Confidence Treshold = 0.7 and Count of Item = 100K (For all data)

a) 9,35 percent of members; have (1-4) membership age, watch documentary contents, do not watch movie contents, favourite team is Galatasaray.

b) 9,45 percent of members; favourite team is Galatasaray do not watch erotic, movie contents, watch sport and documentary contents.

c) 9,69 percent of members; have (1-4) membership age, watch documentary contents, do not watch movie contents, favourite team is Fenerbahce.

d) 9,45 percent of members; favourite team is Fenerbahce do not watch erotic, movie contents, watch sport and documentary contents, have (1-868) annual billing

e) 8,99 percent of members; have at least one account void order, have (868-1827) annual billing, watch documentary contents.

In addition to these results, In the following result list, we use only the data of members who live in Istanbul. Thus I can compare with members of living in Istanbul and all members.

Association Rules for parameters; Support Min Threshold = 0.1 and Support Max Threshold = 0.8 and Confidence Threshold = 0.9 and Count of Item = 35133 (For only data of members who live in Istanbul)

a) 20,13 percent of members; do not have any account void order, watch documentary and sport contents amd do not watch erotic contents.

b) 20,1 percent of members; do not have any account void order, watch documentary and sport contents amd do not watch erotic contents, have (1-868) annual billing.

c) 20,11 percent of mebers; have (42-53) age,watch documenatary and sport contents.

d) 7,94 percent of members; have (1-4) membership, watch documentary and sport contents, have (1-868) annual billing, do not watch movie contents.

e) 33,12 percent of members; watch movie, documentary and sport contents.

f) 24,15 percent of members; watch movie, documentary and sport contents, do not watch erotic contents.

g) 20,42 percent of members; watch movie, documentary and sport contents, do not have any account void order.

h) 18,51 percent of members; watch movie contents, do not watch sport and documentary contents.

i) 17,72 percent of members; watch movie, documentary and sport contents, have (868-1827) annual billing.

j) 15,44 percent of members; watch movie, documentary and sport contents, have (11-15) membership age.

k) 15,15 percent of members; watch movie, documentary and sport contents, do not watch erotic contents, do not have any account void order.

l) 13,3 percent of members; watch movie, documentary and sport contents, favourite team is Besiktas.

m) 12,82 percent of members; watch movie contents, do not watch documentary and sport contents, have (1-868) annual billing.

n) 12,7 percent of members; watch movie, documentary and sport contents, have at least one account void order.

o) 11,8 percent of members; watch movie and documentary contents, have (11-15) membership age, favourite team is Besiktas

p) 11,38 percent of members; watch movie contents, do not watch documentary and sport contents, do not have any account void order.

q) 9,6 percent of members; watch movie, documentary and sport contents, have (54-90) age.

r) 10,69 percent of members; watch documentary,sport and erotic contents.

s) 8,96 percent of members; watch erotic, documentary, movie and sport contents.

Association Rules for parameters; Support Min Threshold = 0.001 and Support Max Threshold = 0.02 and Confidence Threshold = 0.7 and Count of Item = 35133 (For only data of members who live in Istanbul)

a) 1,99 percent of members; watch movie, sport and documentary contents, do not have any account void order, have (1-4) membership age, favourite team is Fenerbahce.

b) 1,99 percent of members; watch movie, sport and documentary contents, do not have any account void order, have (1-868) annual billing, have (32-41) age.

c) 1,99 percent of members; watch movie, sport and documentary contents, do not have any account void order, have (1827-4203) annual billing, favourite team is Besiktas, have (11-15) membership age

d) 1,99 percent of members; watch movie, sport and documentary contents, do not watch erotic contents, have (1-868) annual billing, have at least one account void order.

e) 1,99 percent of members; watch movie, sport and documentary contents, do not watch erotic contents, have (868-1827) annual billing, do not have any account void order, have (32-41) age.

f) 1,98 percent of members; watch movie, sport and documentary contents, do not watch erotic contents, have (868-1827) annual billing, have at least one account void order., have (54-90) age.

g) 1,98 percent of members; watch movie contents, do not watch sport, documentary and erotic contents, have (1-868) annual billing, have (54-90) age.

h) 1,97 percent of members; watch erotic, documentary and sport contents, have (868-1827) annual billing, have at least one account void order.

i) 1,95 percent of members; watch erotic, movie, documentary and sport contents, have (11-15) membership age, have (54-90) age.

j) 1,94 percent of members; watch erotic, movie, documentary and sport contents, favourite team is Trabzonspor.

Association Rules for parameters; Support Min Threshold = 0.1 and Support Max Threshold = 0.8 and Confidence Threshold = 0.9 and Count of Item = 20000 (For only data of members whose favourite team is Fenerbahce)

a) 27,66 percent of members; watch documentary, movie and sport contents.

b) 26,57 percent of members; watch documentary, movie and sport contents, have (1-4) membership age.

c) 24,79 percent of members; watch documentary, movie and sport contents, do not have any account void order.

d) 24,29 percent of members; watch documentary, movie and sport contents, have (1-4) membership age, do not have any account void order.

e) 20,44 percent of members; watch movie, sport and documentary contents, do not watch erotic contents.

f) 24,98 percent of members; watch movie contents, have (1-868) annual billing, have (1-4) membership age, do not have any account void order.

Association Rules for parameters; Support Min Threshold = 0.001 and Support Max Threshold = 0.02 and Confidence Threshold = 0.7 and Count of Item = 20000 (For only data of members whose favourite team is Fenerbahce)

a) 1,97 percent of members; watch erotic, movie, sport and documentary contents, live in Istanbul, do not have any account void order, have (1-4) membership age.

b) 1,97 percent of members; watch erotic, movie, sport and documentary contents, have (42-53) age, do not have any account void order, have (1-4) membership age.

c) 1,94 percent of members; watch erotic, documentary and sport contents, do not watch movie contents, live in OtherCity.

d) 1,94 percent of members; watch erotic contents, live in OtherCity, have (1-4) membership age, have (18-31) age, do not have any account void order.

e) 1,94 percent of members; watch erotic, documentary and sport contents, live in OtherCity, have (18-31) age.

f) 1,94 percent of members; watch erotic, documentary and sport contents, live in OtherCity, have (18-31) age, have (1-4) membership age.

g) 1,67 percent of members; watch erotic, movie, sport and documentary contents, have (1-4) membership age, have (32-41) age.

h) 1,87 percent of members; watch erotic, documentary and sport contents, have (18-31) age, have (1-4) membership age, have (42-53) age, have (1-868) annual billing.

i) 1,87 percent of members; watch erotic, documentary,movie and sport contents, have (1-4) membership age, have (868-1827) annual billing.

Association Rules for parameters; Support Min Threshold = 0.1 and Support Max Threshold = 0.8 and Confidence Threshold = 0.9 and Count of Item = 20000 (For only data of members whose favourite team is Galatasaray)

a) 27,09 percent of members; watch documentary, sport and movie contents.

b) 26,3 percent of members; watch movie, documentary and sport contents, have (1-4) membership age.

c) 26,97 percent of members; watch movie contents, do not have any account void order, have (1-868) annual billing, have (1-4) membership age.

d) 35,97 percent of members; watch movie contents, do not have any account void order, have (1-4) membership age.

6.DISCUSSION

a) Erotic contents are watched less than sport,documentary and movie contents.

b) The most watched contents together are sport and documentary.

c) Number of members who do not watch any contents is more than number of members who watch all contents.

d) The lowest bill-paying members in favor of Galatasaray.

e) The highest bill-paying members in favor of Besiktas.

f) Members of the most watching sports contents in favor of Fenerbahce.

g) Mmebers of the most watching sports contents live in Other Cities.

h) Members of at least watching sports contents in favor of Other Teams. Fans of Besiktas are at least wacthing sports contents in four big team (Fenerbahce, Galatasaray, Trabzonspor, Besiktas)

i) Members of at least watching sport contents live in Izmir

j) Members of the most watching erotic contents live in Bursa.

k) Members of at least wacthing erotic contents live in Ankara.

l) Mebers of the most watching erotic contents in favor of Besiktas.

m) Members of at least watching erotic contents in favor of Galatasaray.

n) Members of the most watched movie contents live in Izmir.

o) Members of at least watching movie contens live in Other Cities.

p) Members of the most watched movie contents in favor of Besiktas.

q) Members of at least watching movie contents in favor of Fenerbahce.

r) Members who have the oldest membership age are in favor of Besiktas.

s) Members who have the oldest membership age live in Istanbul.

t) Members who have the youngest membership age in favor of Galatasaray.

u) Members who have the youngest the membership age live in Other Cities

v) The youngest members live in Other Cities.

w) The youngest members` favourite team is Galatasaray

x) The oldest members live in Ankara.

y) The oldest members` favourite team is Besiktas.

7.CONCLUSION

This study aims to divide the customers into small groups using clustering techniques and also find relative importance of these groups for decision. For clustering, some data preparation techniques are processed. After these period, processed data is clustered by Weka and we can achieve some results. In this thesis, real customer data which is in 2015 is used.

According to this results we can derive information from customer data which contains the application of descriptive and predictive analytics to support the martketing. Many statistical and AI techniques are used in data mining. Apriori algorithm is one of the fastest and earliest tools for Association Mining. In this study, it is intended to use data mining methods to derive conclusions from a large set of real data to be used in strategy setting and decision making process in highly competitive environment.

When clusters are examined, the most watched contents together are sport and document. All members watching sports contents watching the documentary contents. And %50 of these customers` age are between 18 and 31. What is more, %27 of these members whose age is between 18 and 31 live in OtherCity. So generally, young members and living in othercity watching more sports contents.

The members who watch no contents live in the Bursa. Bursa have 5914 members and 1054 of them do not watch any contents so their ratio is the highest value and it is %17,82. For all members; %15 of the members whose age is between 54 and 90 do not watch any contents. But this result is different for members who are living in Bursa. In Bursa, the age of members who do not watch any contents are between 32 and 41. According to these results; Members watched at least content live in Bursa and members who are in the oldest age group are the watched at least content. In addition; the age of most members of the movie watching is between 54 and 90.

According to another results; Members who are at least watching movie contents fan of Fenerbahce and Besiktas fans the most watched movie contents. As you can see favourite team of the members who pay most bills is Besiktas and live in Istanbul, What is more favourite team of the members who pay the least bills is Galatasaray and live in Bursa. Members whose favourite team is Besiktas have the oldest membership age.

When we examined results for erotic contents; Besiktas fans the most watched erotic contents and Galatasaray fans watched at least member of erotic contents. What is more, most watching erotic contents members live in Bursa and at least watching erotic contents members live in Ankara. Most watching erotic contents members` age between 32 and 41. At least watching erotic contents members` age is between 18 and 31.

According to another results; %37,66 of members have least one account void order and %62,34 of members do not have any account void order. %29,58 of this customers of favourite team is Trabzonspor and it is the highest value for this set. %8,08 of this customers of favourite team is Galatasaray. It is the smallest value fot this set. So Fans of Trabzonspor have more potential risk for churn and fans of Galatasaray have less potential risk for churn.

Another results; Members of %24,5 between the ages of 54 and 90. %46,59 of this customers have membership age which is between 11 and 15. It is the highest ratio for this membership age group. %48,79 of this customers live in Istanbul. %75,62 of this customers` favourite team is Besiktas. It is the highest value for this set. So members who have the oldest membership age and oldest age live in Istanbul and they are fans of Besiktas.

Finally according to another results; %13,03 of members watch Erotic contents. Members who live in Bursa most watch erotic contents. According to count of members who live in Bursa. Ages of members of %15,64 are between 32 and 41. For all members, %13,78 of members whose age are between 32 and 41. It is the highest ratio for this set. So members who live in Bursa more watching erotic contents. In addition members whose age are between 32 and 41 more watching erotic contents.

As you can see in our results, using these results, broadcaster companies can improve their relationship with their customers. When the results are combined with reporting environment that allow create creative strategy for CRM studies in future. Continue of this study, similar studies will be done for OTT broadcasting. Thus we can compare customer profiles between Pay-TV broadcast and OTT broadcast or these results can be combined with the results of different process. Such as; churn prediction, special offers for customers. Another study can be done as a continuation of this work, larger dataset

can be used. Because in this study we take 25K members for each team. If we can use larger dataset, we can take distribution which is closer to reality.

REFERENCES

Books

Xianfeng Z., Chen L. & Wang J., 2009. *Web Information Systems and Mining*

Sumathi S. & Sivanandam S., 2006. *Introduction to Data Mining and its Applications*

Periodicals

Jonathan Burez, Dirk Van den Poel, 2007. *CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services.* (*Pages 277–288)*

M2 Presswire, 2015. *Segmentation VOD & OTT Usage - Computers, Tablets, and Smartphones Analysis.*

Nai-hua C., Huang, Chi-tsun S., Shu, Shih-tung, Wang & Tung-sheng, 2013. *Market segmentation, service quality, and overall satisfaction: self-organizing map and structural equation modeling methods (pages 969-987)*

Zeng, Fanbin, Liu & Ruini 2012. *Specialization & Personalization of Pay TV Channel*

Journal of Business Administration Research

Other References

Basic Definitions: Data Mining, http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm [accessed 18-12-2015]

Data statics and size, http://uk.emc.com/about/news/press/2011/20110628-01.htm [Accessed 20-12-2015]

Type of Attributes, https://users.cs.fiu.edu/~taoli/class/CAP4770-F10/class021.ppt [Accessed 10-01-2016]

Chakrabarti, Cox, Frank, Güiting, Han, Jiang, Kamber *Data Mining Know It All*, https://books.google.com.tr/books?id=WRqZ0QsdxKkC&pg=PT74&lpg=PT74&dq=Incomplete,+noisy,+and+inconsistent+data+are+commonplace+properties+of+large&source=bl&ots=PUCPkBUCm2&sig=lkyCSd2K-0JmAt_rM9oop4AkYWM&hl=tr&sa=X&ved=0ahUKEwjIj5qc7YPMAhUFLhoKHX45CGMQ6AEIJjAB#v=onepage&q=Incomplete%2C%20noisy%2C%20and%20inconsistent%20data%20are%20commonplace%20properties%20of%20large&f=false [Accessed 12-01-2016]

Data Integration & Data Transformation In Data Mining Science, http://www.slideshare.net/farshadbadie/data-integration-and-data-transformation [Accessed 18-01-2016]

Cluster Analysis and Clustering Algorithms, https://en.wikipedia.org/wiki/Cluster_analysis [Accessed 08-09-2015]

K-Means Clustering Algorithm, https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm [Accessed 10-09-2015]

http://stanford.edu/~cpiech/cs221/handouts/kmeans.html [Accessed 27.03.2016]

Apriori Algorithm, https://en.wikipedia.org/wiki/Apriori_algorithm [Accessed 03-04-2016]

Euclidean Distances, https://en.wikipedia.org/wiki/Euclidean_distance [Accessed 24.03.2016]

Association Rules, http://w3.gazi.edu.tr/~akcayol/files/WM_L2AssociationRules.pdf [Accessed 12.04.2016]