DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

DEVELOPMENT AND IMPLEMENTATION OF A PRICE PREDICTION SYSTEM USING MACHINE LEARNING TECHNIQUES

by
Sercan Eren İŞKESEN

September, 2014
İZMİR

DEVELOPMENT AND IMPLEMENTATION OF A PRICE PREDICTION SYSTEM USING MACHINE LEARNING TECHNIQUES

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Master of
Science in Computer Engineering

by

Sercan Eren İŞKESEN

September, 2014

İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"DEVELOPMENT AND IMPLEMENTATION OF A PRICE PREDICTION SYSTEM USING MACHINE LEARNING TECHNIQUES"** completed by **SERCAN EREN İŞKESEN** under supervision of **ASST. PROF. DR. DERYA BİRANT** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. Derya BİRANT

Supervisor

Prof. Dr. Alp Kut

(Jury Member)

Asst. Prof. Dr. Derya EREN AKYOL

(Jury Member)

Prof. Dr. Ayşe OKUR

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

I would like to thank to my supervisor, Asst. Prof. Dr. Derya BİRANT, for her support, supervision and useful suggestions throughout this study.

I would like to offer my special thanks to my instructors for their support and help. Finally, I would like to offer my very special thanks to my friend Çağrı DOĞAN for the motive support. It would not have been able to complete this thesis without their support and help.

Sercan Eren İŞKESEN

DEVELOPMENT AND IMPLEMENTATION OF A PRICE PREDICTION SYSTEM USING MACHINE LEARNING TECHNIQUES

ABSTRACT

Machine Learning and Data Mining techniques are very common usage in recent years and their usages are getting wide in different areas. One of the goals of machine learning is prediction and it can be used to obtain more accurate prediction results. For this reason, in this thesis, a price prediction system was developed using machine learning techniques for finding a value, which is the ideal or closest to ideal, according to the user requirements.

Predicting the price of a product has become an increasingly important area of research. The aim of this study is designing and implementing a price prediction system that helps buyers and sellers to make effective decision.

Clustering and classification are two common tasks of data mining. This thesis proposes hybrid combinatorial method of clustering and classification for predicting the price of a product. In the experimental works, the proposed approach was successfully applied for used-car price prediction as a case study. Experimental results show that using more than one machine learning techniques can give better results than using just only one machine learning technique.

In this study, we provide the following new contributions: (i) collecting data related to the automobile industry, (ii) creating of a data warehouse by applying advanced data processing methods, (iii) clustering data using a clustering algorithm and then classifying the clustered values using a classification algorithm and (iv) comparing the proposed model (K-Means and Naïve Bayes) with the classic model (simple Naïve Bayes) for price prediction.

Keywords: Machine learning, data mining, clustering, classification.

MAKİNE ÖĞRENME TEKNİKLERİ KULLANILARAK FİYAT TAHMİNLEME SİSTEMİNİN YAPILMASI VE GELİŞTİRİLMESİ

ÖZ

Son yıllarda, Makine Öğrenme ve Veri Madenciliği teknikleri yaygın bir şekilde kullanılmaktadır ve farklı alanlarda kullanımları giderek genişlemektedir. Makine öğrenmenin amaçlarından biri tahmin edebilmektir ve daha doğru tahmin sonuçları elde edebilmek için kullanılabilmektedir. Bu nedenle, bu tezde, makine öğrenme teknikleri kullanılarak, kullanıcı gerekliliklere göre en ideal veya ideale yakın sonuç bulan bir fiyat tahminleme sistemi geliştirilmiştir.

Bir ürünün fiyatının belirlenmesi giderek daha önemli olan bir araştırma alanı haline gelmiştir. Bu çalışmanın amacı, alıcıların ve satıcıların etkin karar vermelerine yardımcı olacak bir fiyat tahminleme sistemi tasarlamak ve geliştirmektir.

Kümeleme ve sınıflandırma, veri madenciliğinde yaygın olarak kullanılan iki yöntemdir. Bu tezde, bir ürünün fiyatını tahmin etmek için kümeleme ve sınıflandırmanın birlikte kullanımını içeren melez bir yöntem önerilmektedir. Deneysel çalışmalarda, önerilen yaklaşım bir örnek çalışma olarak ikinci el araba fiyat tahmini için başarılı bir şekilde uygulanmıştır. Deneysel sonuçlar birden fazla makine öğrenme tekniği kullanmanın, sadece tek bir makine öğrenme tekniği kullanmaya göre daha iyi sonuçlar verebildiğini göstermektedir.

Bu çalışmada sağladığımız yeni katkılar: (i) otomobil sektörüne ilişkin verilerin toplanması, (ii) gelişmiş veri işleme metotları uygulayarak bir veri ambarı oluşturulması, (iii) bir kümeleme algoritması kullanılarak verilerin kümelenmesi ve daha sonra kümelenmiş değerlerin bir sınıflandırma algoritması ile sınıflandırılması ve (iv) fiyat tahmini için önerilen modelin (K-Means ve Naive Bayes) klasik model (basit Naive Bayes) ile karşılaştırılmasıdır.

Anahtar kelimeler: Makine öğrenmesi, veri madenciliği, kümeleme, sınıflandırma.

CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER ONE
INTRODUCTION

1.1 General

Since the importance of the machine learning and data mining is increased, researchers are working on them. With this importance and needs caused the increase of developing new machine learning tools. With these machine learning tools, data mining applications are getting used commonly. In this study, a new data mining application, which is based on a hybrid combinatorial method of clustering and classification tasks, was developed.

Predicting the price of a product has become an increasingly important area of research because buyers and sellers can make effective decision by this way. Predicting the price depends on many factors, such as item type, model, color, and many more. Automating the (intelligent) evaluation of a product offered for sale and predicting its final price significantly reduces human involvement, ultimately resulting in higher satisfaction for buyers and sellers.

Predictive modeling consists of statistical, machine learning or data-mining solutions including algorithms and techniques to determine future outcomes. A predictive model is made up of a number of predictors, which are variable factors that are likely to influence future behavior or results. In automobile marketing, for example, model, kilometer, engine size, year, gear-type might be used to predict the likelihood of a car price.

Some research works have been done in the area of forecasting the price of a product using machine learning techniques such as Support Vector Machine (SVM), Neural Network (NN), Regression Analysis (RA), Time Series Analysis (TSA) etc. However, differently from these studies, this thesis proposes a hybrid approach and uses different machine learning algorithms.

This thesis shows the improvements in the price prediction by applying clustering and classification techniques one after another. The proposed approach was implemented in automobile industry and experimental results show that applying more than one machine learning technique can give better result than applying just one machine learning technique. With this finding, this study contributes to machine learning researches for their future studies.

1.2 Purpose

The purpose of this study is finding the ideal value related to the user requirements by applying machine learning techniques on collected data. In other words, the aim of study is to develop a system that automatically tries to determine the right value of a product according to the properties of it to help buyers and sellers. Because price prediction is an important topic for market participants to be make right decisions.

In this paper, a price prediction model is proposed using machine learning techniques to determine the price of a product using collected data which includes the previous products' properties and their prices. The proposed model is hybrid combinatorial method of clustering and classification. The proposed approach consists of two main steps. In the first step of the proposed model, the collected data are clustered into groups of similar products based on their characteristics using K-Means clustering algorithm. In the second step of the proposed model, the clustered values are classified using Naive Bayes classification algorithm.

In this thesis, an application was developed to demonstrate the benefits of the proposed model. It finds the ideal value for users toward their requirements and it makes more accurate according to the traditional ones.

1.3 Related Work

In the literature, some studies have been presented for price prediction. One of them is for price-volatility prediction for optimal trading strategy (Choudhury et. al.

2014). They used clustering and support vector machine techniques to obtain better results.

Some studies have been done for the price prediction related to house sales (Bin, 2004; Wang et al 2014; Guo et al 2012). For example, Gacovski et al presented a study about real estate valuation estimation with using data mining for certain city of Skopje and in that study researches has used neural network and decision tree techniques (Gacovski et. al. 2012).

Chan developed a system for predicting the digital camera prices with using several machine learning techniques (Chan, 2008). Furthermore, Nasira and Hemageetha have worked to find a forecasting model for vegetable prices using neural network technique (Nasira & Hemageetha 2012). A study about stock price prediction has been done using k-nearest neighbor algorithm (Alkhatib et. al. 2013).

Moreover, another study uses a data set for used cars on sale during summer of 2004 in Netherlands and in that study they use 38 attribute including price, age, kilometer, horsepower and other features. In addition, this study made for only Toyota automobiles and researcher use regression trees for predict price (Shmueli et. al. 2005).

Another study also uses automobile dataset to predict car prices. In that study, researchers use multiple regression, multi-layer perception, support vector machine and neural network approaches (Cortez, 2012).

Furthermore, a car price prediction research has been made with car leasing application as a master thesis with using some machine learning algorithms (Listiani, 2009).

In 2012, a GIS based decision support system for hotel rate estimation and temporal price prediction study has been made for hotel brokers. In that study,

regression and support vector machine algorithms are used for prediction (Kisilevich et. al. 2012).

In addition, product price prediction is made with using k-nearest algorithm and in that study 20 fold cross validation has been made with using 10521 instances (Raykhel & Ventura 2009).

The study in this thesis differs from the previous studies on the following methodological grounds. First, it automatically collects used-car data from a website and constructs a data warehouse. Second, another novelty of this study is that it uses a hybrid combinatorial method of clustering and classification for price prediction. Third, it uses an estimator to determine the number of clusters.

1.4 Organization of the Thesis

This thesis includes nine chapters and the remaining of this thesis is organized as follows.

In Chapter 2, general information about machine learning and data mining, main tasks of data mining, basic concepts, definitions and explanations related to the thesis topic are given.

In Chapter 3, basic definitions about clustering, explanation about a clustering validation technique, the description of a clustering algorithm and examples are given.

In Chapter 4, basic definitions about classification, explanations about classification validation techniques, the description of a classification algorithm and examples are given.

In Chapter 5, data preprocessing techniques and how data was processed in this study are explained with examples. Each step of data preprocessing is explained: data

collection, data reduction, data transformation (data discretization, data normalization) and data integration.

In Chapter 6, the design of the system is given in detail. It presents UML diagrams such as class diagram and sequence diagram. It explains database in detail by giving Entity Relation (E/R) diagram, database tables and view lists.

In Chapter 7, the implementation of a price prediction system is explained. Development tools and database system used to implement the system is given. In addition, the details about the proposed model are given.

In Chapter 8, experimental work and obtained results are given. It presents the comparison between two approaches on the test set in the experiment.

Finally in Chapter 9, the conclusion remarks and future works are given.

CHAPTER TWO
MACHINE LEARNING AND DATA MINING

2.1 Machine Learning

2.1.1 Basic Definition of Machine Learning

Machine learning is a study about artificial intelligence which concerns about construction of systems that can learn from specific data. The main focus of machine learning researches are the detection of complex patterns to computer issues and give users the ability to make decisions about the data. In addition, many research area uses or closely linked to machine learning techniques, some of them are statistics, probability theory, data mining, pattern recognition and of course artificial intelligence (Zhang & Ji, 2013). Most field of computer science is applied machine learning techniques (Lima & Machado, 2012). Moreover some of the area which uses machine learning is machine perception, natural language processing, search engines, stock market analysis, speech and handwriting recognition, computer games, some of websites (Devi et. al. 2007). There are some kinds of machine learning techniques. These are supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning (Li, Guo, & Elkan 2011). Most of time, these are perplex with machine learning algorithms (Decherchi et. al. 2012). Some approaches to machine learning algorithms can be listed like this; decision tree learning, association rule learning, artificial neural networks, support vector machines and supervised clustering (Thuy et. al. 2008).

2.1.2 Supervised Learning

Supervised Learning is one of the machine learning techniques in which input data is labeled. A model is constructed by training process, which requires making predictions and corrections if those predictions wrong. The training process continues to execute until model succeeded an intended level of accuracy on the data.

2.1.3 Unsupervised Learning

Unsupervised Learning is another machine learning technique in which input data is unlabeled, which means input data where the target output data is unknown. It is based on discovering structure for finding the output; on the other hand it does not use to generalize a mapping from inputs to output for finding the solution.

2.1.4 Semi-supervised Learning

Another learning technique is Semi-supervised Learning. This technique uses the combination of supervised and unsupervised learning techniques. It uses labeled and unlabeled input data together to generate an appropriate solution or value.

2.1.5 Reinforcement Learning

Reinforcement learning is an area of machine learning afflatus from behaviorist psychology. It is about how their output has been more efficient and gains more reward.

2.2 Data Mining

Data mining is firstly introduced in 1990s, and then significant developments were made. Now, it has been applied for many different applications in different areas. Below, basic concepts and terms related with data mining are given.

2.2.1 Basic Definition of Data Mining

Task of data mining is usually confused with machine learning concept. In spite of machine learning, data mining means, reaching the information in large scale data. Another explanation of data mining is, searching on large scale data to discover hidden information, common patterns and correlations between data using algorithms. Data mining studies can provide predictions about future. Before data mining process, data preprocessing is necessary, which includes removing noisy data, cleaning inconsistent data, integration data from various sources, selecting

relevant data and transforming data to another format (normalization and discretization)(Listiani, 2009).

2.2.2 Data Mining Tasks

Data mining involves six common tasks. These are anomaly detection, association rule mining, clustering, classification, regression, and summarization.

2.2.2.1 Anomaly Detection

Anomaly detection, in other words outlier or deviation detection, is the identification of unusual and extreme data values that might cause errors at the results in the research. For example, the records A, B and C in Figure 2.1 are outlier values in data and they can be detected with anomaly detection method in data mining (Listiani, 2009).

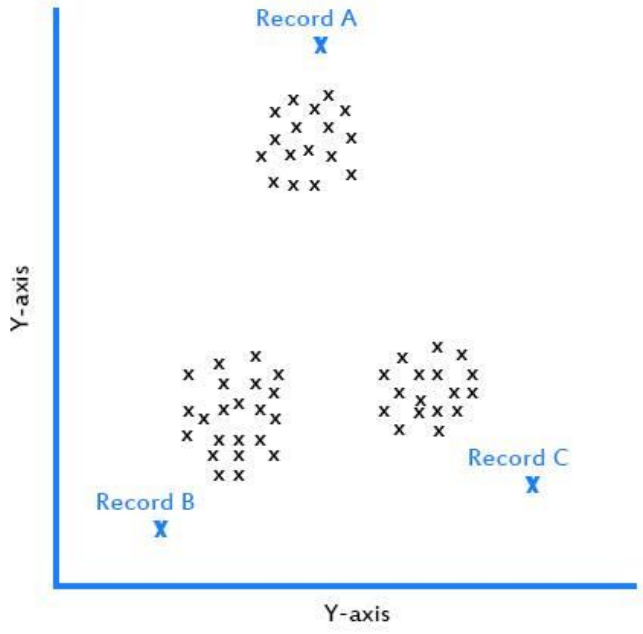


Figure 2.1 Outlier examples

2.2.2.2 Association Rule Mining

The aim of association rule mining is to discover relationships between data values. This technique is very commonly used. For example, it is used in supermarkets which is collecting data on customers what they are buying and the supermarket owners can change their sell behaviors according to the results from their searches.

2.2.2.3 Clustering

Clustering is another data mining task which is discovering structures in data and grouping them that are in the same way or similar. This method do not use known structure in the data. Figure 2.2 shows a clustering example. In that example, data values are separated as three groups and six outliers don't belong to any group.

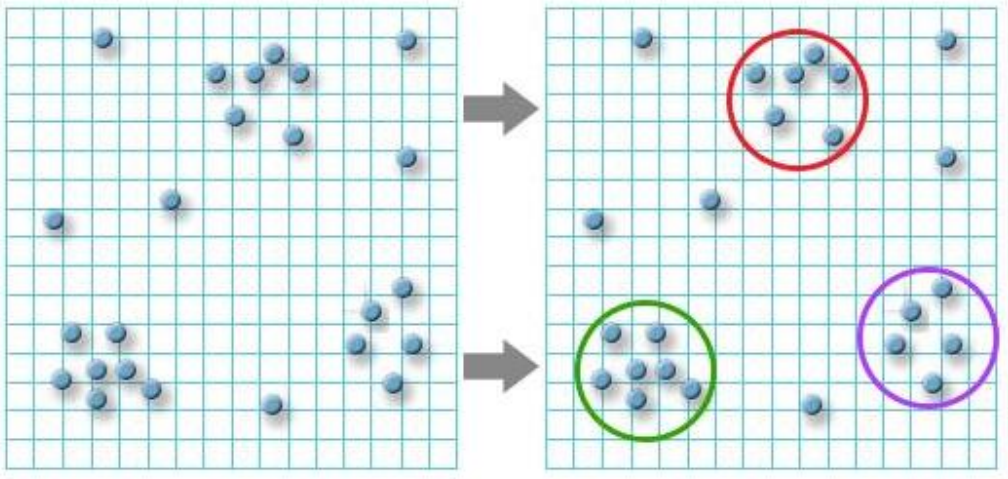


Figure 2.2 Clustering example

2.2.2.4 Classification

Classification is another commonly used task of data mining. Classification method is generalizing the known structure and applies this structure to new incoming data value. As an example, in this study, a system was developed to find the value for price attribute. When the system doing this task, it constructs a model for finding the value for price and the system knows all other attributes except the price attribute.

2.2.2.5 Regression

Regression is also a data mining task that is used to fit an equation to a dataset to predict a number. The aim of this method is to find a function which models the data with the least error in it (Huang et. al. 2012). In Figure 2.3, a regression example is given. After regression is applied, the regression function is called as regression line.

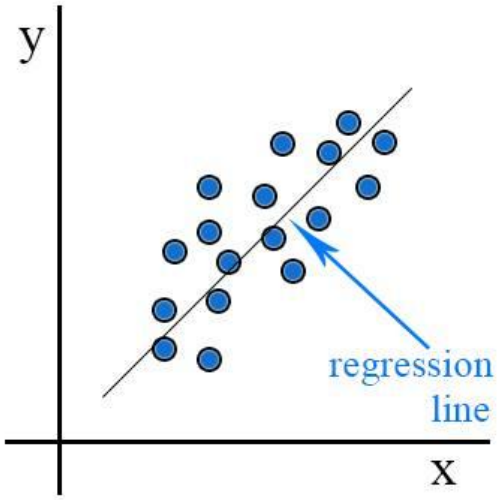


Figure 2.3 An example for regression method and regression line

2.2.2.6 Summarization

Summarization is another task of data mining, which is providing and creating a small or compact representation of the data set, and finds the result from there.

CHAPTER THREE
CLUSTERING

3.1 Basic Definition of Clustering

Clustering is a task of grouping data in a way or using certain methods for separating those data to clusters. In clustering, the similar objects should be in the same cluster. Different clusters are dissimilar to each other if the clusters are well separated. Figure 3.1 shows a clustering example.

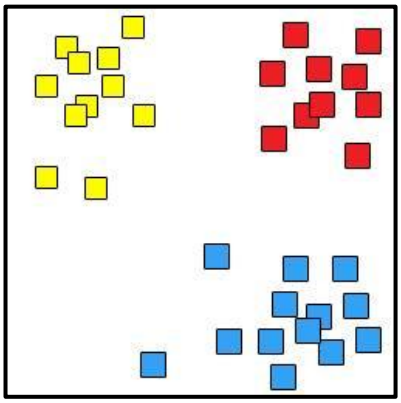
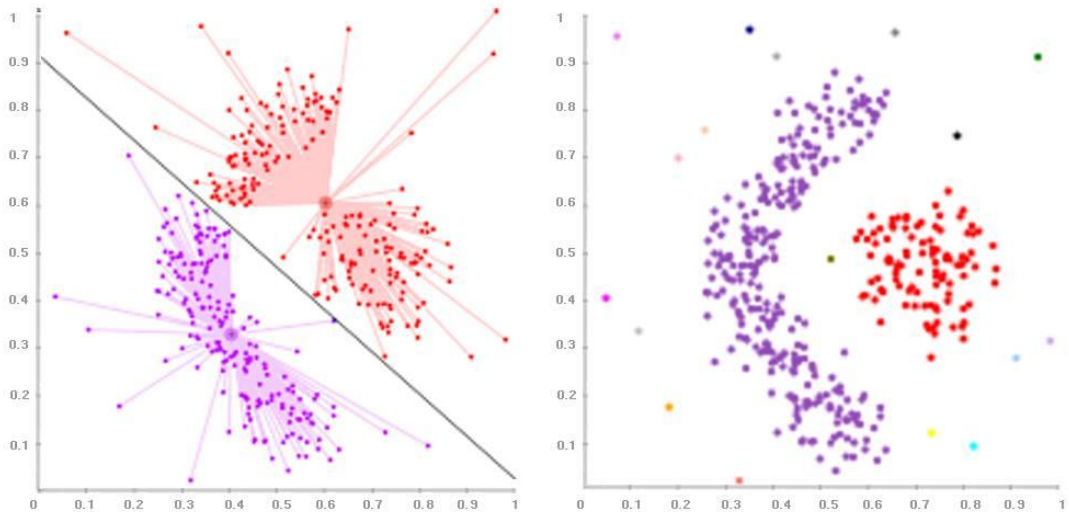


Figure 3.1 A clustering example

There are several models to find clusters in a clustering process. These are; Connectivity models, Centroid models, Distribution models, Density models, Subspace models, Group models, and Graph-based models (Gu, 2012). *Connectivity model* uses distances for relations. *Centroid model* uses a single mean vectors for each cluster, and K-Means algorithm is the best example of this model. *Distribution model* is kind of statistics based model and it distributes most likely data in the same cluster (Gu, 2012). In *density based model*, clusters are defined as areas when higher density is determined. Outliers can also be identified in this model.

Currently, clustering method is used in many areas such as bioinformatics, medicine, business, marketing, manufacturing, telecommunication, financial and banking (Guo et. al. 2012).

In clustering, determining the proper algorithm is a hard and important problem for solving the clustering problem actually. We should carefully decide the correct algorithm to apply on data. If it is chosen as wrong, it might cause very poor and wrong clustering results (Gu, 2012). Figure 3.2 (a) shows an example for a poor clustering result which was constructed by using an algorithm in centroid model (Liang et. al. 2012). Figure 3.2 (b) shows an example for a good clustering result which was constructed by using an algorithm in density based model. When a density based model is applied on the data like in this example, it separates the clusters significantly and clearly. So that, it gives more correct and better results to researchers (Segaran, 2007).



(a) (b)

Figure 3.2 Centroid model vs. Density based model

3.2 K-Means Algorithm

K-means is one of the clustering methods used in data mining. It is a popular algorithm used in many applications. The aim of this algorithm is clustering all data into specific cluster number which is decided by the user. Each data in clusters belongs to the cluster with the nearest mean method. Figure 3.3 shows an example to describe how k-means algorithm works.

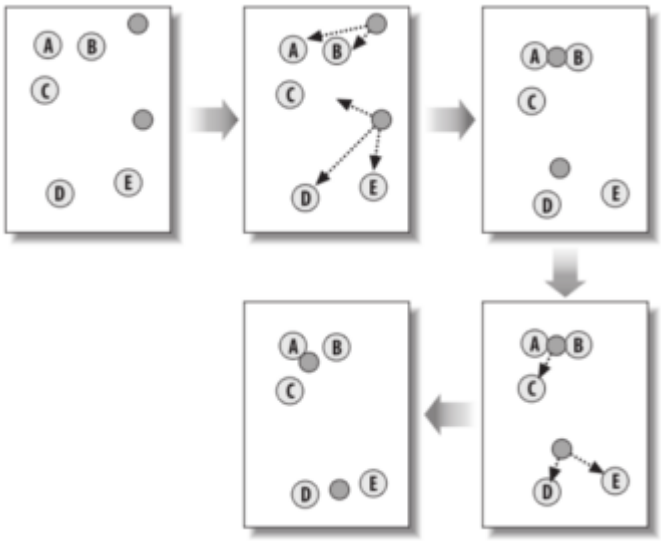


Figure 3.3 K-means clustering example (Segaran, 2007)

When K-Means algorithm begins working, it decides initial starting points, which number of starting points equals number of clusters (Gu, 2012). In Figure 3.4, there is an example of selecting starting points. This figure also shows the iterations of the algorithm step by step. In K-Means algorithm, it is necessary to use a distance metric. The distance measure can be change problem to problem. In this study, Euclidean distance equation is used for calculating the distance between the data values.

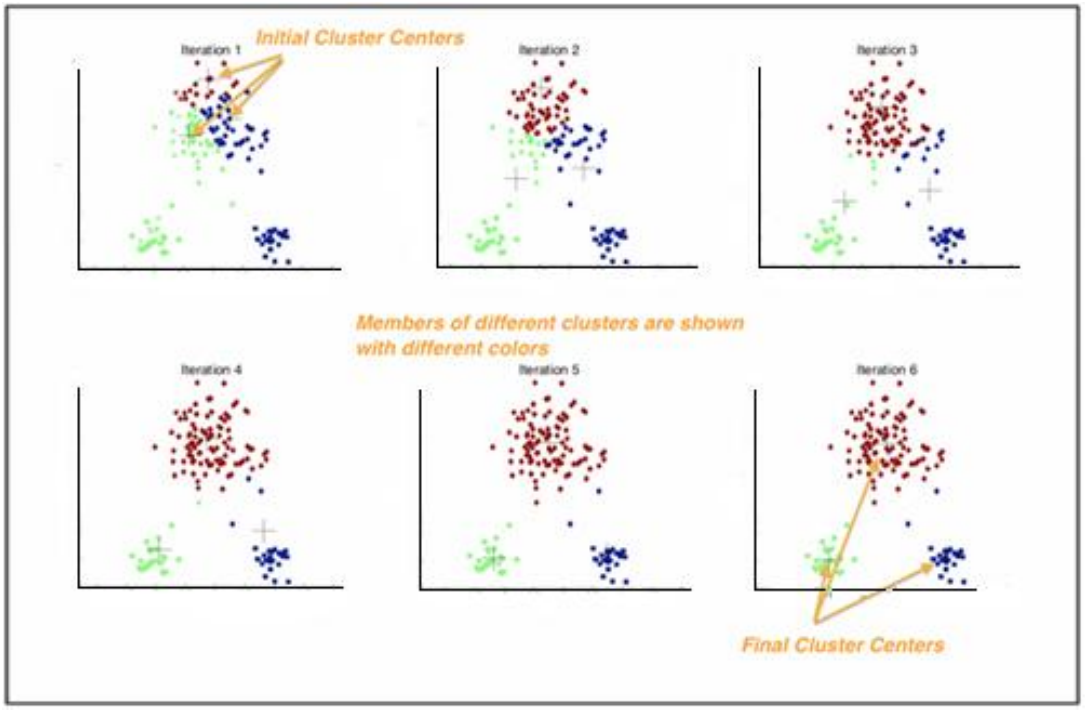


Figure 3.4 Step by step K-Means process

Figure 3.5 shows the pseudo code of K-Means algorithm.

```
Function kmeans(dataSet, k)
{
    // generate centroids randomly
    numFeatures = dataSet.getNumFeatures()
    centroids = getRandomCentroids(numFeatures, k)

    iterations = 0

    // start k-means algorithm
    while not stopCriteria(oldCentroids, centroids) {
        // save old centroids for stability test
        oldCentroids = centroids
        iterations += 1

        // assign labels to each datapoint based on centroids
        labels = getLabels(dataSet, centroids)

        // find new centroids
        centroids = getCentroids(dataSet, labels, k)
    }

    return centroids
}
```

Figure 3.5 Pseudo code of K-Means algorithm

3.2.1 Euclidean Distance

Euclidean distance is a distance measure used to measure similarity based on the characteristics of objects. It is computed as the square root of the sum of the squares of the differences between corresponding values. The formula of Euclidean distance is given in (3.1).

$$d(p,q) = \sqrt{(q1-p1)^2 + (q2-p2)^2 + ... + (qn-pn)^2} = \sqrt{\sum_{i=1}^{n}(qi-pi)^2} \quad (3.1)$$

where $p = (p1, p2,..., pn)$ and $q = (q1, q2,..., qn)$ are two points in n-space.

3.2.2 Sum of Squared Errors

Sum of Squared Errors (SSE) is the sum of the squares differences between values of data and its group's mean. It can be used for determine the number of cluster. Ideal SSE value is zero because it means that all data values are within a different cluster. SSE formula is given in (3.2).

$$SSE = \sum_{i=1}^{n} (X_i - \bar{X})^2 \quad (3.2)$$

where n is the number of observations X_i is the value of the ith observation and $\bar{X}$ is the mean of all the observations.

3.2.3 Elbow Method

Elbow is a definition for big change in SSE values according to number of clusters. Figure 3.6 shows an example SSE values calculated according to the different number of clusters. In this example, the elbow point, so the number of clusters, is determined at 'ten' (k = 10).

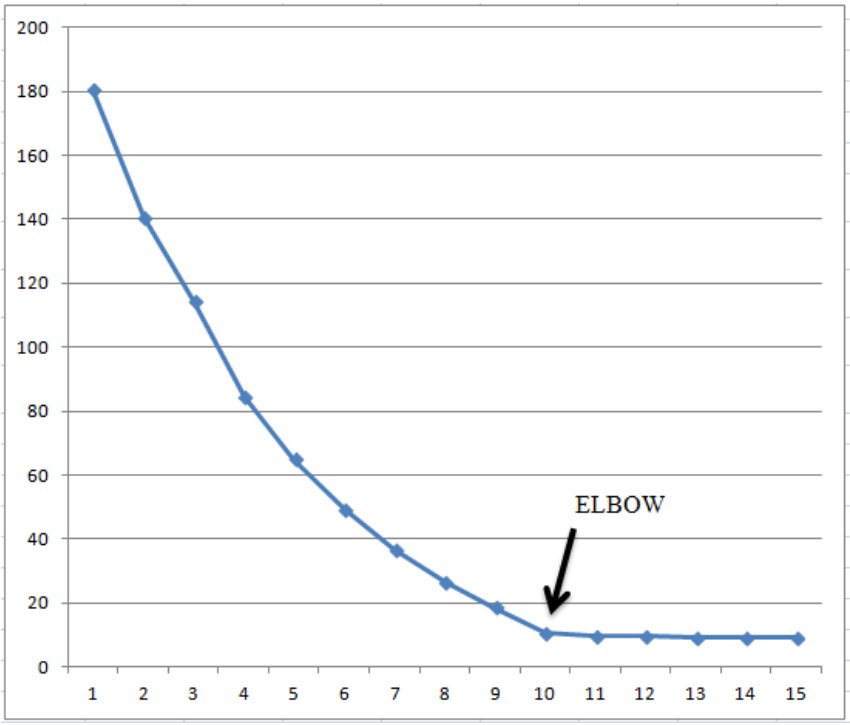


Figure 3.6 An example for Elbow method

CHAPTER FOUR
CLASSIFICATION

4.1 Basic Definition of Classification

Classification is the problem of identifying to which of a set of categories a new observation belongs to, on the basis of a training set. In classification method, data is separated to classes based on their target value. The classes in the classification methods are not fixed. They can change what the target is because the probabilities are always changing what we are searching (Lima & Machado, 2012).

In machine learning terminology, classification belongs to under supervised learning. In addition, classifier is known as an algorithm that implements classification (Vineyard et. al. 2012).

Several classification techniques are used in computer science. They are: support vector machines, decision trees, neural network, kernel estimation and linear classifiers. Many common classification algorithms used in machine learning area are C4.5, Naïve Bayes, Back Propagation and K-Nearest Neighbor (Thuy et. al. 2008). Among them, Naïve Bayes belongs to linear classifiers category, while K-nearest belongs to kernel estimation category (Pang & Vaithyanathan, 2002).

In classification applications, dataset is partitioned into two parts: training set and test set, as shown in Figure 4.1. While training set is used to construct a model, test set is then used to determine the accuracy of the model (Zhang & Ji, 2013).

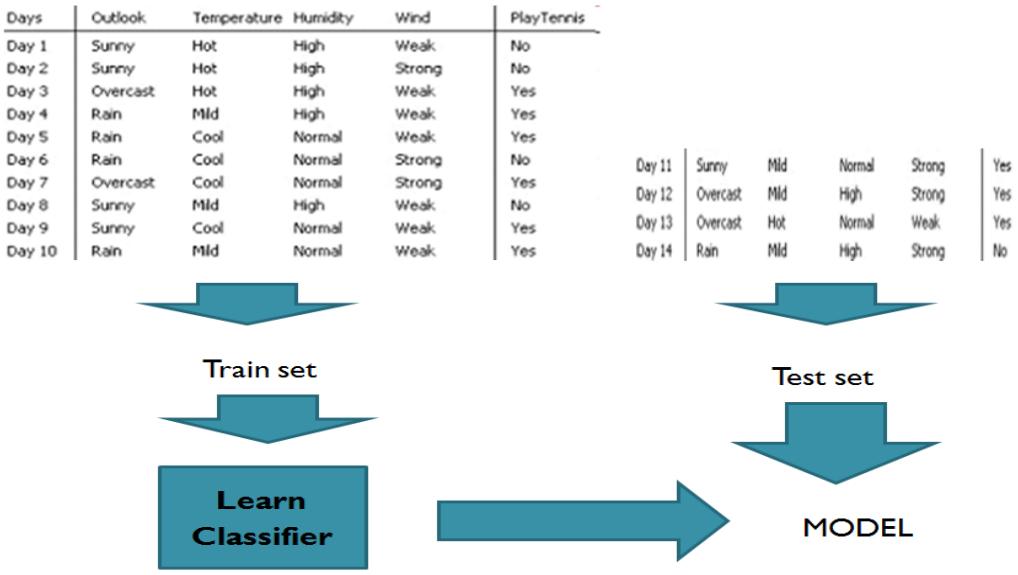


Figure 4.1 Steps of creating model

When researchers are creating train and test sets, they also chose a classification validation technique for their research. They chose it according to the problem and their purpose. If researcher wants a good prediction in their research, they chose one of the validation techniques which have performing better accurate in the test data (Yusuf, Othman & Salim 2010).

Among validation techniques, two validation techniques are commonly used. Those are *simple validation* and *cross validation*. In Figure 4.2, an example is given to show how simple and cross validation techniques separates the data as training and test set.

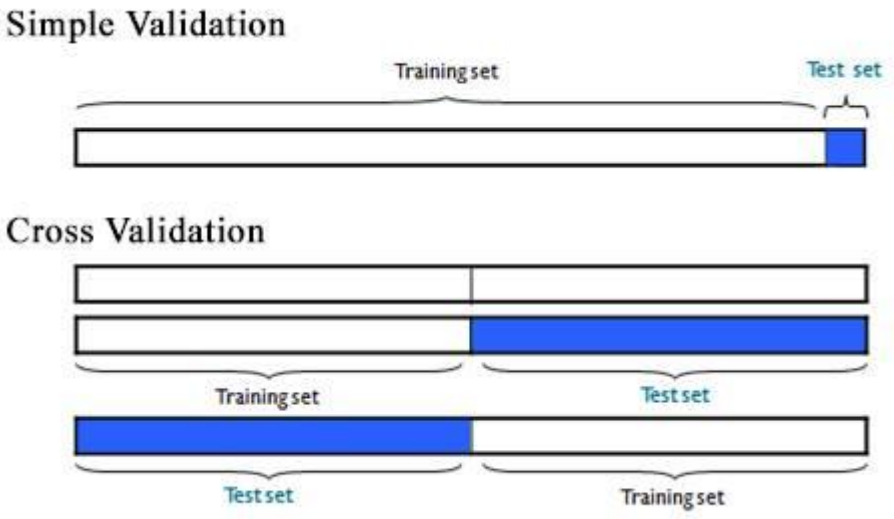


Figure 4.2 Simple and cross validation

K-Nearest Neighbor is also a well-known classification algorithm. K-Nearest Neighbor calculates the distances between instances by using a distance metric such as Manhattan or Euclidean distances (Sahoo & Makur, 2013).

In K-nearest Neighbor algorithm, the distance between the new data and the classes around it is calculated and it is decided to assign the new data to k-closest classes. These distances can be weighted or non-weighted. So that, the deciding mechanism of algorithm uses different formulas for finding the belonging class for new data.

Another well-known classification algorithm is support vector machine. This classification method tries to separate the data as far as possible, like in Figure 4.3. In this example, H3 line is better than H1 and H2 lines because in the first and second tries the data cannot be separated as good as the third one (Rahman et. al. 2011).

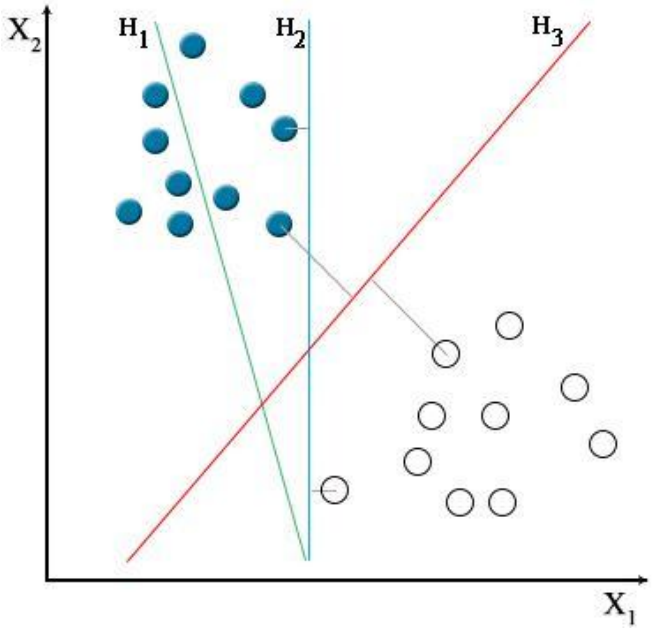


Figure 4.3 Support vector machine

The support vector machine technique finds the ideal line for separating the classes appears clearly (Crawford, 2013). This technique calculates the distances for every line and compares them with each other (Figure 4.4)(Wang et. al. 2014).

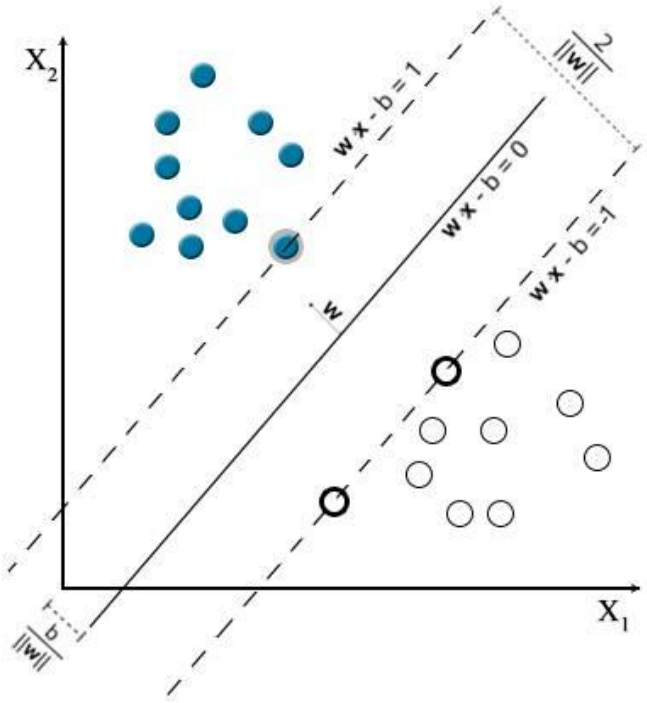


Figure 4.4 Result of support vector machine technique

After construction a classification model, researcher should calculate the accuracy which means the correctness of the research. Accuracy step is realized according to the four conditions: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) (Figure 4.5). They are calculated as looking the actual data and what the test says. If test say the value what the researcher is looking for is actually in the class what the research estimates, it means *True Positive*. On the other hand, if the value is not in the class which is not estimated, it means it is *False Positive*.

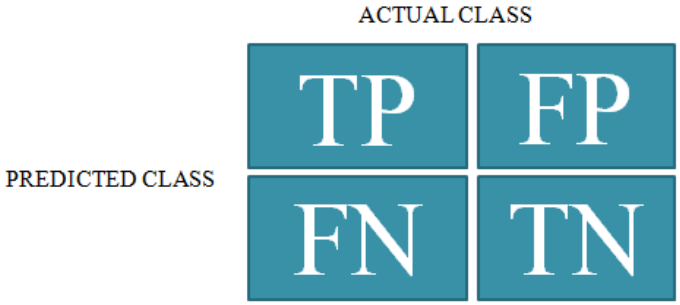


Figure 4.5 Actual class and predicted class

As a final step, the researcher should calculate the success rate with the accuracy formula. Accuracy is calculated as given in (4.1). True cases are added and divided by all cases. The error calculated as given in (4.2).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = Trues / All \tag{4.1}$$

$$Error = \frac{FN + FP}{TP + TN + FP + FN} = Falses / All \tag{4.2}$$

4.2 Naïve Bayes Algorithm

Naïve Bayes algorithm is the one of the classification algorithms in data mining, which is based on calculating every criterion's effect to result value with using possibility calculations. It is the one of the most commonly used and well known classification algorithm. In Formula (4.3), Naïve Bayes algorithm's probability calculation is shown.

$$P(A/B) = P(B/A)P(A)/P(B) \tag{4.3}$$

where P(A) is the prior probability, P(A|B) is the conditional probability and P(A|B) is the posterior probability.

In Naïve Bayes algorithm, firstly, all possibilities of all criterions are calculating one by one. After these calculations, results from every cluster's possibilities are multiplied with each other.

Naïve Bayes algorithm can be explained with an example easily (Ziemniak, 2011). Assume that we have the dataset given in Table 4.1. In this data, the researcher is looking for can player play tennis in day fifteen. So that, researchers enter the data about 'Outlook', 'Temperature', 'Humidity', and 'Wind' attributes in day fifteen. As an example, in day fifteen, weather values like this, 'Outlook' is sunny, 'Temperature' is cool, 'Humidity' is high and, 'Wind' is strong. So that, researcher will look for those data values for their calculations for yes and no result.

Table 4.1 Example dataset

Days	Outlook	Temperature	Humidity	Wind	PlayTennis
Day 1	Sunny	Hot	High	Weak	No
Day 2	Sunny	Hot	High	Strong	No
Day 3	Overcast	Hot	High	Weak	Yes
Day 4	Rain	Mild	High	Weak	Yes
Day 5	Rain	Cool	Normal	Weak	Yes
Day 6	Rain	Cool	Normal	Strong	No
Day 7	Overcast	Cool	Normal	Strong	Yes
Day 8	Sunny	Mild	High	Weak	No
Day 9	Sunny	Cool	Normal	Weak	Yes
Day 10	Rain	Mild	Normal	Weak	Yes
Day 11	Sunny	Mild	Normal	Strong	Yes
Day 12	Overcast	Mild	High	Strong	Yes
Day 13	Overcast	Hot	Normal	Weak	Yes
Day 14	Rain	Mild	High	Strong	No

Firstly, when ‘PlayTennis’ attribute is calculating, researcher must count how many yes value the data got in ‘PlayTennis’ attribute.

The attribute which is trying to find the result about it is called target attribute. In addition, for this example ‘PlayTennis’ attribute is target attribute. The data which is yes in target attribute is shown in Table 4.2.

In this example, the ‘yes’ result for target is 9/14. So, after this result the researcher will continue to calculate positive results over those nine data. After this step, researcher looks for all possibilities for all attributes like when ‘Outlook’ is sunny, yes for target is 2/9. And, when ‘temperature’ is cool, yes value for target is 3/9. For ‘Humidity’ attribute, when it is high, its possibility is 3/9 and for ‘Wind’, when it is strong, its probability is 3/9. As we can see, all calculations done for yes and over nine value.

Table 4.2 ‘Yes’ values in target attribute

Days	Outlook	Temperature	Humidity	Wind	PlayTennis
Day 1	Sunny	Hot	High	Weak	No
Day 2	Sunny	Hot	High	Strong	No
Day 3	Overcast	Hot	High	Weak	Yes
Day 4	Rain	Mild	High	Weak	Yes
Day 5	Rain	Cool	Normal	Weak	Yes
Day 6	Rain	Cool	Normal	Strong	No
Day 7	Overcast	Cool	Normal	Strong	Yes
Day 8	Sunny	Mild	High	Weak	No
Day 9	Sunny	Cool	Normal	Weak	Yes
Day 10	Rain	Mild	Normal	Weak	Yes
Day 11	Sunny	Mild	Normal	Strong	Yes
Day 12	Overcast	Mild	High	Strong	Yes
Day 13	Overcast	Hot	Normal	Weak	Yes
Day 14	Rain	Mild	High	Strong	No

After that, researcher needs to find the all yes possibility and they do it with multiplying all possibilities which find them before. ‘Yes’ value probability will be found 0.0053 as shown in (4.4).

$$P(y) = 9/14 * 2/9 * 3/9 * 3/9 * 3/9 = 0{,}0053 \quad (4.4)$$

Furthermore, after researcher found values for ‘yes’, time to find the ‘no’ possibilities. In Table 4.3, ‘no’ values is shown.

When researcher calculating the ‘no’ possibilities, they divide the possibilities over five for this example. So that, when ‘Outlook’ is sunny its possibility is 3/5 and when ‘Temperature’ is cool its possibility is 1/5 and when ‘Humidity’ is high its possibility is 4/5 and, when ‘Wind’ is strong its possibility is 3/5. So that, its value is be 0.0205 as shown in (4.5).

$$P(n) = 5/14 * 3/5 * 1/5 * 4/5 * 3/5 = 0{,}0205 \quad (4.5)$$

Table 4.3 ‘No’ values in target attribute

Days	Outlook	Temperature	Humidity	Wind	PlayTennis
Day 1	Sunny	Hot	High	Weak	No
Day 2	Sunny	Hot	High	Strong	No
Day 3	Overcast	Hot	High	Weak	Yes
Day 4	Rain	Mild	High	Weak	Yes
Day 5	Rain	Cool	Normal	Weak	Yes
Day 6	Rain	Cool	Normal	Strong	No
Day 7	Overcast	Cool	Normal	Strong	Yes
Day 8	Sunny	Mild	High	Weak	No
Day 9	Sunny	Cool	Normal	Weak	Yes
Day 10	Rain	Mild	Normal	Weak	Yes
Day 11	Sunny	Mild	Normal	Strong	Yes
Day 12	Overcast	Mild	High	Strong	Yes
Day 13	Overcast	Hot	Normal	Weak	Yes
Day 14	Rain	Mild	High	Strong	No

When researcher found class *(yes* and *no)* probabilities separately, it’s time to look which value is bigger than the other. For this example, probability of cannot play tennis which means ‘no’ value is bigger than the can playable which means ‘yes’ value, so that, in day fifteen when the weather values been like this, ‘Outlook’ is sunny, ‘Temperature’ is cool, ‘Humidity’ is high and, ‘Wind’ is strong, the tennis playability is been ‘no’.

CHAPTER FIVE
DATA PREPROCESSING

Data preprocessing is an important step in Knowledge Discovery in Databases (KDD) and should be done before data mining process to get accurate results. In this thesis, four major tasks of data preprocessing have been done: Data Collection, Data Reduction, Data Transformation, and Data Integration. In this chapter, these stages are explained step by step.

5.1 Data Collection

Data collection is the first step in data preparation process. It includes identifying the relevant data, collecting it efficiently and accurately as well as storing in a database. After identifying the goal of data mining, data needs to be gathered from various sources which will be useful for the analysis. Based on the problem definition, the data attributes necessary to be collected can be defined.

Data may collect from various sources like web sites, files, databases, data marts and so on. Large datasets are required to build accurate predictive models.

The aim of this thesis is to design and implement a new system for price prediction by using machine learning algorithms. As an application, car price prediction system was implemented to demonstrate the success of the proposed approach. For prediction of used car prices, we will require the past car data across years which capture the information on the characteristics of the cars and their prices. Additional data can be captured on the features of the car such as first-owner, from-owner, guarantee, smoke-free, accessories, from-lady, ABS, alarm, air-bag, park-distance-check and turbo.

At the beginning stages of this thesis, a program was coded to collect data from a website. This program was roamed inside the website (http://www.arabam.com/) and

collected necessary data for this thesis. The collected dataset contains 76,763 car records. Data was collected within the period from May 2012 to August 2013.

The program developed to collect data from the web site uses regular expression codes for choosing necessary values from the web site. Figure 5.1 shows some of regular expression codes used to collect data from the web site.

```
string result = wc.DownloadString("http://www.arabam.com/ikinciel/Arama/?arctp=1&krlk=-1&hsr=-1&PageNo=" + x + "&ItemPerPage=50");

Regex reg1 = new Regex("(?<=contentSMiddle\">).*\\r\\n(.*\\r\\n)*\\s*(?=<div class=\"sBottomBox1\">)");  //for main page
Regex reg2 = new Regex("(?<=sMiddleBox\">).*\\r\\n(.*\\r\\n)*?\\s*(?=<div class=\"sMiddleBoxIn10 fLeft\">)");   //for car ad part
Regex reg3 = new Regex("(?<=<strong>).*?\\s*(?=</strong>)");    //for page part
Regex reg4 = new Regex("(?<=<a href=\").*?\\s*(?=\")");    //for links
Regex reg5;   //
Regex reg6;   //  status
Regex reg7;  //  chassis
Regex reg8;   // color
Regex reg9;  //  city town
Regex reg10 = new Regex("(?<=<strong>).*?(?= TL</strong></a>)");    //    price
#region contentSMiddle
foreach (Match match in reg1.Matches(result))
{
    foreach (Match match2 in reg2.Matches(match.Value))
    {
        listBox3.Items.Add(reg4.Match(match2.Value)); // "details" link of each ad
        listBox1.Items.Add(match2.Value); // DIV of each ad
        listBox2.Items.Add(reg3.Match(match2.Value.Replace(" ", " "))); // title  of each ad
        Match prc = reg10.Match(match2.Value);
        price.Add(Convert.ToInt32(prc.Value.Replace(".","")));
        textBox2.Text = "";
    }

}
```

Figure 5.1 Program code written to collect main data

After collecting main data about cars from a web page, as named as specialties for cars, additional information about cars was also collected from another web page. Figure 5.2 shows a part of the program code which is coded for collecting additional data, named as features for cars, from another web page.

```
    #region detailsPage
    foreach (var arabalink in listBox3.Items)   //taking every cars' ad
    {
        conn.Open();
        result = wc.DownloadString("http://www.arabam.com" + arabalink);
        //textBox1.Text = result;
        reg1 = new Regex("(?<=<div class=\"facePop\"[^>]*>).*\\r\\n(.*\\r\\n)*\\s*(?=<div class=\"contentRight\">)");   //inside page
        reg2 = new Regex("(?<=<dd>)(.|\n)*?(?=</dd>)");   //  all features  part
        reg3 = new Regex("(?<=<a[^>]*>)(.|\n)*?(?=</a>)");    //links part
        reg4 = new Regex("(?<=stOn[^>]*>)(.|\n)*?(?=</li>)");    //liston part which is chosen ones
        reg5 = new Regex("(?<=0km_ikinciel_model.arabam[^>]*>)(.|\n)*?(?=</a>)");   //
        reg6 = new Regex("(?<=hsr=[^>]*>)(.|\n)*?(?=</a>)");   //damaged or not part
        reg7 = new Regex("(?<=gvd=[^>]*>)(.|\n)*?(?=</a>)");   //chassis part
        reg8 = new Regex("(?<=rnk=[^>]*>)(.|\n)*?(?=</a>)");   // color part
        reg9 = new Regex("(?<=marka=[^>]*>)(.|\n)*?(?=</a>)");   //model part

        foreach (Match match in reg1.Matches(result))     //find and replacing part of ads
        {
            Match mdl = reg9.Match(result);
            model = mdl.Value.Replace("        ", "").Replace("\r", "").Replace("\n", "");

            Match brand = reg5.Match(result);
            string brandname = brand.Value.Replace("\n", "").Replace("\r", "").Replace("                ", "");
            SqlCommand findModel = new SqlCommand("SELECT brand_id FROM BRAND WHERE Brand_name='" + brandname + "'", conn);
            SqlDataReader readModel = findModel.ExecuteReader();
            while (readModel.Read())
            {
                brand_id = Convert.ToInt32(readModel.GetValue(0));
            }
            readModel.Dispose();
            MatchCollection mc = reg2.Matches(match.Value); // Specialities
            MatchCollection mc2 = reg4.Matches(match.Value); // Features
```

Figure 5.2 Program code written to collect additional information

After getting data automatically from web pages via developed programs, it was stored in the database. Figure 5.3 shows the code written to insert data into the database.

```
SqlCommand command2 = new SqlCommand("INSERT INTO SPECIALITIES(Brand_id,Km,Year,Fueltype,Geartype,Cc,Hp,Model,City,Town,
command2.Parameters.AddWithValue("@brand_id", brand_id);
command2.Parameters.AddWithValue("@km", km);
command2.Parameters.AddWithValue("@year", year);
command2.Parameters.AddWithValue("@fueltype", fueltype);
command2.Parameters.AddWithValue("@geartype", geartype);
command2.Parameters.AddWithValue("@cc", cc);
command2.Parameters.AddWithValue("@hp", hp);
command2.Parameters.AddWithValue("@model", model);
command2.Parameters.AddWithValue("@city", city);
command2.Parameters.AddWithValue("@town", town);
command2.Parameters.AddWithValue("@status", status);
command2.Parameters.AddWithValue("@chassis", chassis);
command2.Parameters.AddWithValue("@color", color);
command2.Parameters.AddWithValue("@price", price[counter]);
command2.Parameters.Add("@ID", SqlDbType.Int).Direction = ParameterDirection.Output;

int specid = command2.ExecuteNonQuery();
specid = Convert.ToInt32(command2.Parameters["@ID"].Value.ToString());
command2.Dispose();
```

Figure 5.3 Program code written to insert collected data into database

The original data contains several attributes with continuous values, such as kilometer, engine size, year etc., as well as nominal values such as model, color and other features like them.

5.2 Data Reduction

Data reduction is another possible objective for data mining. It is the selection of the subset from the whole dataset. Due to huge size of data and because some of them is irrelevant to objective, some of the data is redundant to increase performance of the algorithm and to get more processing power.

In this thesis, instead of using collected data as a whole, some attributes were eliminated based on their effects to the outcome. That means, before using raw data, the elimination process has been applied and processed data was written in view tables in the database.

5.3 Data Transformation

Data transformation is a key aspect of data mining. Data transformation is to use simple mathematical formulations or methods to convert data into different format for the purpose of data analysis. This process typically involves normalization and discretization of data features/attributes.

In this thesis, before applying machine learning algorithms on data, both data normalization and discretization processes have been done.

5.3.1 Data Discretization

After all data collected, for applying algorithms, that data must be converted to the information. So that, data warehousing applied to the database and unnecessary data was eliminated but not deleted. They just passed to the view tables. After these steps, thesis got necessary information for executing algorithms on it. For applying algorithms on this information, other software was coded.

Table 5.1 shows the intervals of price attribute and their corresponding discrete values obtained after discretization.

Table 5.1 Discretization of *price* attribute

Interval	Discrete Value
0 - 2750	Very Low
2750 - 9000	Low
9001 - 20000	Mid
20001- 41000	High
41001 or higher	Very High

Table 5.2 shows the intervals of price attribute and their corresponding discrete values obtained after discretization.

Table 5.2 Discretization of *cc* attribute

Interval	Discrete Value
< 1600	cc-low
>= 1600	cc-high

5.3.2 Data Normalization

Data normalization is one of the most used data transformation methods. It is commonly made by updating numeric values to keep them in a value range. Such as, taking all data values and getting new values them in between zero and one. If this normalization progress is not applied before applying a clustering algorithm, the result is not taken properly because the technique will find a result which is based on the biggest weighted values. As an example, if clustering technique is applying to a second hand car data, like in this study, normalization must be apply before study because if normalization is not applied before study, having a large number values in data field like kilometer value within all other data is having weight on it and it cause obtaining false results on the technique. Thus, it is essential to normalize the inputs so that their variability reduces their importance.

Mostly used normalization methods are Decimal Scaling, Min-Max Normalization and Z-Score Normalization. In decimal scaling method, each

numerical value is divided by the same power of 10. Min-max normalization technique is computed based on the minimum and maximum values of the data for transformation of data to a desired range. In z-score normalization, the data is rescaled according to the mean and standard deviation that are computed from the data.

In this thesis, Min-Max Normalization technique was used. This technique changes values into a range to provide equal weight distribution for all attributes. Min-Max Normalization formula is given in Formula 5.1.

$$newvalue = \frac{currentvalue - min}{max - min}(newmax - newmin) + newmin \quad (5.1)$$

where *newvalue* is the new normalized value, *currentvalue* is the original variable value, *min* and *max* are the minimum and maximum values of the attribute, *newmin* and *newmax* are the minimum and maximum value for the normalized range like [newmin..newmax]. Note that if the new range is [0..1], then the Formula (5.1) simplifies to Formula (5.2).

$$newvalue = \frac{value - min}{max - min} \quad (5.2)$$

Table 5.3, Table 5.4 and Table 5.5 shows example normalizations worked on *kilometer*, *cc* and *hp* attributes.

Table 5.3 Normalization example for *kilometer* attribute

Original Value	**Normalized Value**
750	0.00
41.400	0.30
69.000	0.50
83.800	0.60
138.750	1.00

Table 5.4 Normalization example for *cc* attribute

Original Value	Normalized Value
0	0.00
1242	0.39
1560	0.50
1998	0.63
3148	1.00

Table 5.5 Normalization example for *hp* attribute

Original Value	Normalized Value
16	0.00
163	0.25
332	0.51
487	0.76
661	1.00

5.4 Data Integration

After collecting data in the database, data integration process has been done by joining nine tables. At the end of data integration, five *view tables* have been constructed for using in research.

'Specialties' table in the database may be called the main table because it holds the main characteristic of a car such as model, kilometer, year, fuel-type, gear-type, color, price etc. 'Features' table is another supportive table which holds additional car features inside it such as first-owner, from-owner, guarantee, smoke-free, accessories, from-lady, tradable, AAS, ABS, alarm, back-view-cam', air-bag-driver, park-distance-check, turbo etc.

One of the view tables given in Figure 5.6 was constructed from ‘Specialties’ given in Figure 5.4 and ‘Features’ table given in Figure 5.5. In the system, machine learning techniques were applied on view tables. The query written for data integration is given in Figure 5.7.

The data values in view tables were changed as numerical values. The reason of that is, K-Means clustering algorithm cannot be applied on the verbal data, it needs scalar data for apply. So that, when data was transferring to the view tables, it was changed into the scalar form.

Specia...	Brand_id	Km	Year	Fueltype	Geartype	Cc	Hp	Model	City	Town	Status	Chassis	Color	Price
239	73	102000	2011	2	1	1493	110	Accent Era 1.5 CRDi-...	İstanbul	Sultan...	2	1	6	25750
240	144	216000	2003	2	3	2148	143	C 220 CDI Elegance ...	İstanbul	Sultan...	2	1	10	58500
241	144	37000	2012	1	3	1595	156	C 180 BlueEFFICIENC...	İstanbul	Sultan...	2	1	6	93500
242	144	266000	2000	3	3	3148	200	S 320 Long	İstanbul	Sultan...	2	1	9	45750
243	162	96000	2008	1	1	1498	90	Astra 1.4 NB Twinport	İstanbul	Sultan...	2	1	42	23750
244	162	262000	1991	3	1	1791	100	Vectra 1.8 Comfort	İstanbul	Sultan...	2	1	6	9750
245	185	34000	2012	2	2	1598	105	Leon 1.6 CR TDI Styl...	İstanbul	Sultan...	2	2	6	48500
246	185	11200	2012	1	1	1399	97	Accent Era 1.4 Start A...	Osma...	Merkez	2	1	12	27250
247	216	81000	2011	2	1	1600	105	Golf 1.6 TDi Highline	İstanbul	Bahçel...	2	2	12	42950
248	73	198000	2004	2	1	1560	110	Focus C-Max 1.6 TDC...	Konya	Karatay	2	2	42	26500
249	216	110000	2010	2	2	1598	105	Jetta 1.6 TDI Exclusiv...	Sakarya	Adapa...	2	1	12	44750
250	37	92000	2009	3	2	1598	113	Cruze 1.6 LS Otm.	Zongu...	Merkez	2	1	9	28000
251	37	82000	2004	1	1	1242	80	Albea 1.2 EL	Zongu...	Merkez	2	1	12	16500
252	37	94000	2011	2	1	1248	90	Linea 1.3 Multijet Acti...	Çanak...	Merkez	2	1	12	25500
253	37	37000	2012	2	1	1598	105	Linea 1.6 Multijet Emo...	Zongu...	Merkez	2	1	14	34250
254	144	130000	2001	1	3	1998	163	C 200 Kompressor Sp...	Çanak...	Merkez	2	5	40	35500
255	175	100000	2010	2	1	1461	85	Fluence 1.5 dCi Expre...	Çanak...	Merkez	2	1	10	36000
256	175	74000	2006	1	1	1149	75	Clio 1.2 16V Authentiq...	Çanak...	Merkez	2	2	5	18500
257	208	56000	2008	1	1	1598	124	Corolla Sedan 1.6 Ele...	Çanak...	Merkez	2	1	14	35000

Figure 5.4 Sample data in Specialties table

	Features_id	Firstow...	Notusedcar	Guarantiee	Service	Smokefree	Aksesories	Fromlady	Emergency	Tradable	Unnegotiable	Maturitiable	Scotfree
251	251	0	0	0	0	0	0	0	0	1	0	0	1
252	252	1	1	0	1	1	0	0	0	1	0	0	1
253	253	0	0	0	0	0	0	0	0	1	0	0	1
254	254	0	1	0	0	0	1	0	0	0	0	0	0
255	255	1	1	0	1	1	0	0	0	1	0	0	1
256	256	1	0	0	0	0	0	0	0	0	0	0	0
257	257	0	0	0	1	0	0	0	0	0	0	0	1
258	258	0	0	0	0	0	0	0	0	1	0	0	1
259	259	0	1	0	0	1	0	0	0	1	0	0	1
260	260	0	0	0	0	0	0	0	0	0	0	0	1
261	261	0	0	0	1	1	0	0	0	0	0	0	1
262	262	0	1	0	1	1	0	0	0	0	0	0	1
263	263	1	1	0	1	1	0	0	0	1	0	0	1
264	264	1	0	1	1	0	0	0	0	1	0	0	1
265	265	0	0	0	1	0	0	0	0	1	0	0	1
266	266	0	1	0	0	1	0	0	0	0	0	0	1
267	267	1	1	0	1	1	0	0	0	1	0	0	1
268	268	0	0	0	0	0	0	0	0	0	0	0	1
269	269	0	0	0	0	0	0	0	0	0	0	0	1
270	270	1	0	0	1	1	1	0	1	1	0	0	1
271	271	1	0	0	1	1	0	0	0	0	0	0	0
272	272	0	0	0	1	0	0	0	1	0	0	0	1

Figure 5.5 Sample data in Features table

	Specialities_id	Brand_id	Km	Year	Fueltype	Geartype	Cc	Hp	Status	Price	Chassis	Color	Firstowner	Notusedcar	Guarantiee	Ser
1	65	72	25000	1999	3	1	1581	83	2	1000	1	3	0	0	0	0
2	66	72	100000	1985	3	1	1600	110	2	3000	1	16	1	1	0	1
3	67	72	164000	1984	3	1	1600	150	2	50000	1	28	0	0	0	0
4	115	29	3000	2013	2	3	1995	184	2	137900	1	9	1	1	1	0
5	116	29	2000	2013	2	3	1995	184	2	138750	1	6	1	1	1	0
6	117	29	5000	2013	2	3	1995	184	2	137750	1	6	1	1	1	0
7	118	37	190000	2004	1	1	1400	80	2	15500	1	6	0	0	0	0
8	119	37	174000	2001	1	1	1600	75	2	10500	1	6	0	0	0	0
9	120	37	25318	2012	2	1	1248	75	2	31300	2	42	0	0	0	0
10	121	73	45000	2011	1	1	1600	180	2	43000	2	9	0	0	1	0
11	122	88	112000	2009	1	2	1600	100	1	33000	1	9	0	0	0	0
12	123	88	112000	2009	1	2	1600	100	1	33000	1	9	0	0	0	0
13	124	162	26000	2011	2	1	1300	100	2	33500	2	42	0	0	0	0
14	125	175	90000	2010	1	1	1400	77	2	20500	1	6	0	0	0	0
15	126	175	110000	2010	1	1	1400	77	2	20500	1	42	0	0	0	0
16	127	175	25000	2012	2	1	1500	85	2	43500	1	42	0	0	1	0
17	128	175	160000	2010	2	1	1500	70	2	21500	1	42	0	0	0	0
18	129	208	88000	2008	2	3	1400	100	2	34500	1	9	0	0	0	0

Figure 5.6 Sample data in view table

```
SELECT    dbo.SPECIALITIES.Km,       dbo.SPECIALITIES.Year,
      dbo.SPECIALITIES.Fueltype,     dbo.SPECIALITIES.Geartype,
      dbo.SPECIALITIES.Cc,           dbo.SPECIALITIES.Hp,
      dbo.SPECIALITIES.City,         dbo.SPECIALITIES.Status,
      dbo.SPECIALITIES.Chassis,      dbo.SPECIALITIES.Color,
      dbo.SPECIALITIES.Price,        dbo.FEATURES.Firstowner,
      dbo.FEATURES.Notusedcar,       dbo.FEATURES.Guarantiee,
      dbo.FEATURES.Service,          dbo.FEATURES.Smokefree,
      dbo.FEATURES.Aksesories,       dbo.FEATURES.Fromlady,
      dbo.FEATURES.Emergency,        dbo.FEATURES.Tradable,
      dbo.FEATURES.Unnegotiable,     dbo.FEATURES.Maturitiable,
      dbo.FEATURES.Scotfree,         dbo.FEATURES.Fromowner,
      dbo.FEATURES.AAS,              dbo.FEATURES.ABS,
      dbo.FEATURES.Alarm,            dbo.FEATURES.Backwievcam,
      dbo.FEATURES.ASR,              dbo.FEATURES.EBD,
      dbo.FEATURES.EDL,              dbo.FEATURES.ESP,
      dbo.FEATURES.Airbagcurtain,    dbo.FEATURES.Airbagdriver,
      dbo.FEATURES.Airbagside,       dbo.FEATURES.Airbagpass,
      dbo.FEATURES.Isofix,           dbo.FEATURES.Tiremalfsign,
      dbo.FEATURES.Centrallock,      dbo.FEATURES.Parkdistancecheck,
      dbo.FEATURES.Woodensteer,      dbo.FEATURES.Locklessentry,
      dbo.FEATURES.Leatherwoodensteer,dbo.FEATURES.Leatherfabricseat,
      dbo.FEATURES.Leathersteer,    dbo.FEATURES.Leatherseat,
      dbo.FEATURES.Deepchangeablesteer, dbo.FEATURES.Autowindows,
      dbo.FEATURES.Automemoryseat,  dbo.FEATURES.Crusecontrol,
      dbo.FEATURES.Hydrosteer,       dbo.FEATURES.Heatedsteer,
      dbo.FEATURES.Heatedseat,       dbo.FEATURES.Ac,
      dbo.FEATURES.Fabricseat,       dbo.FEATURES.Startstop,
      dbo.FEATURES.Sunroof,          dbo.FEATURES.Phonebluetooth,
      dbo.FEATURES.Tripcomp,         dbo.FEATURES.Heightchangeblesteer,
      dbo.FEATURES.Alloywheel,       dbo.FEATURES.Windowroof,
      dbo.FEATURES.Autosidemirror,  dbo.FEATURES.Headlightsign,
      dbo.FEATURES.Autowashheadlight,dbo.FEATURES.Modified,
      dbo.FEATURES.Foglights,        dbo.FEATURES.Xenon,
      dbo.FEATURES.Rainsensor,       dbo.FEATURES.Cdplayer,
      dbo.FEATURES.Cdchanger,        dbo.FEATURES.Ipodconnector,
      dbo.FEATURES.Mpthreeplayer,    dbo.FEATURES.Radio,
      dbo.FEATURES.Navi,             dbo.FEATURES.Fourwheel,
      dbo.FEATURES.Turbo
FROM    dbo.SPECIALITIES INNER JOIN  dbo.SPECFEAT ON
        dbo.SPECIALITIES.Specialities_id =
        dbo.SPECFEAT.Specialities_id
        INNER JOIN dbo.FEATURES ON dbo.SPECFEAT.Features_id =
        dbo.FEATURES.Features_id
```

Figure 5.7 Query written for data integration

Data integration should also been done to get names of the features from another table according to their ids. For example, 'Specialties' table includes brand-id attribute, it is possible to get the name of the brands from another table 'Brand' table in the database. Figure 5.8 shows a sample data in 'Brand' table.

	Brand_id	Brand_name
24	24	BEDFORD
25	25	BENELLI
26	26	BENTLEY
27	27	BISAN
28	28	BMC
29	29	BMW
30	30	BOMBARDI...
31	31	BUGATTI
32	32	BUICK
33	33	CADILLAC
34	34	CASE IH
35	35	CEZETA
36	36	CHERY

Figure 5.8 Sample data in Brand table

When integrating two tables, it is possible to append the attributes from one onto the other. It is possible when two tables share a key field, such as an ID code. For example, Specialties given in Figure 5.4 and Chassis tables given in Figure 5.9 have the same key field chassis-id, so it can be joined during data integration process.

Results | Messages

	Chassis_id	Chassis
1	1	Sedan
2	2	Hatchback/5
3	3	Hatchback/3
4	4	Station Wagon
5	5	Coupe
6	6	Cabrio
7	7	MPV
8	8	Minivan
9	9	Van
10	10	Camli Van

Figure 5.9 Data in Chassis table

CHAPTER SIX
DESIGN

This chapter explains the design of the system in detail. It presents UML diagram such as class diagram and sequence diagram. It explains database in detail by giving Entity Relation (E/R) diagram, database tables and view lists.

6.1 E/R Diagram

Figure 6.1 shows the design of the database. It presents Entity / Relation (E/R) Diagram for explaining database in detail.

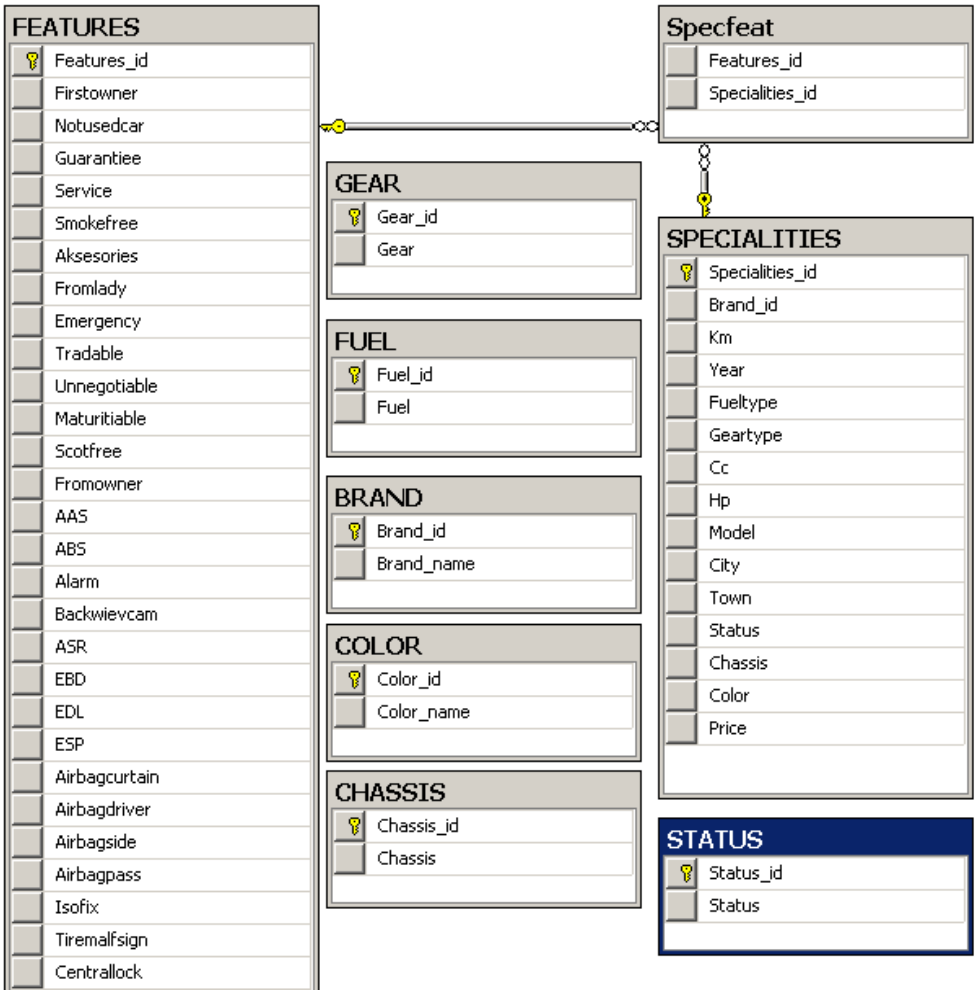


Figure 6.1 Entity / Relation diagram

Database used in this thesis mainly consists of nine tables: Brand, Chassis, Color, Status, Gear, Fuel, Specialties-Features, Specialties and Features Tables which are shown in Table 6.1, Table 6.2, Table 6.3, Table 6.4, Table 6.5, Table 6.6, Table 6.7, Table 6.8 and Table 6.9 respectively.

Table 6.1 Brand table

Name	Type	Explanation
Brand_Id	Int	Primary key for Brand table
Brand_Name	Nchar(20)	The name of the Brands

Table 6.2 Chassis table

Name	Type	Explanation
Chassis_Id	Int	Primary key for Chassis table
Chassis	Nchar(20)	The name of Chassis types

Table 6.3 Color table

Name	Type	Explanation
Color_Id	Int	Primary key for Color table
Color_Name	Nchar(20)	The name of the Colors

Table 6.4 Status table

Name	Type	Explanation
Status _Id	Int	Primary key for Status table
Status	Nchar(20)	The name of status situations for cars

Table 6.5 Gear table

Name	Type	Explanation
Gear_Id	Int	Primary key for Gear table
Gear_Name	Nchar(20)	Gear types

Table 6.6 Fuel table

Name	Type	Explanation
Fuel _Id	Int	Primary key for Fuel table
Fuel	Nchar(20)	The type of the fuel for cars

Table 6.7 Specialties-features table

Name	Type	Explanation
Features _Id	Int	A connective table between Specialities and Features, so that, it holds features ids
Specialities_Id	Int	A connective table between Specialties and Features, so that, it holds specialties ids

Table 6.8 Specialties table

Name	Type	Explanation
Specialities_Id	Int	Primary key for specialties table
Brand_Id	Int	Connects with Brand table
Km	Int	Km information for cars
Year	Int	Year information for cars
Fueltype	Int	Connects with Fuel table
Geartype	Int	Connects with Gear table
Cc	Int	Cc information for cars
Hp	Int	Hp information for cars
Model	Nchar(60)	Model information for cars
City	Nchar(20)	City information for cars
Town	Nchar(20)	Town information for cars
Status	Int	Connects with Status table
Chassis	Int	Connects with Chassis table
Color	Int	Connects with Color table
Price	Int	Price information for cars

Table 6.9 Features table

Name	**Type**	**Explanation**
Features_Id	Int	Primary key for features table
Firstowner	Bit	This car is sold from first owner or not
Notusedcar	Bit	It specified the car is not used before or not
Guarantiee	Bit	It specified car still in warranty or not
Service	Bit	The seller of car send it to service as in time
Smokefree	Bit	The seller never smokes in their car
Aksesories	Bit	Accessories was added to the car
Fromlady	Bit	The seller is woman or not
Emergency	Bit	The seller wants to sell it as soon as possible
Tradable	Bit	The car is tradable with another car with buyer
Unnegotiable	Bit	The car price is not changeable
Scotfree	Bit	The car is scotfree feature with it or not
Fromowner	Bit	The car is selling from owner or not
AAS	Bit	The car has AAS features or not
ABS	Bit	The car has ABS feature with it
Alarm	Bit	The car which is sold alarm feature in it
Backviewcam	Bit	The car has camera to show the back side
ASR	Bit	The car has ASR features or not
EBD	Bit	The car has EBD features or not
EDL	Bit	The car has EDL features or not
ESP	Bit	The car has ESP features or not
Airbagcurtain	Bit	The car has airbags on side of the car
Airbagdriver	Bit	The car has airbag on just in driver seat
Airbagside	Bit	The car has airbags on side of the seats
Airbagpass	Bit	The car has airbag on passenger side
Isofix	Bit	The car has Isofix holders at seats
Tiremalfsign	Bit	The car has sign in its dashboard which is shows tire has got malfunction or not

Table 6.9 Features table(cont.)

Centrallock	Bit	The car has central lock feature
Parkdistancecheck	Bit	The car has a park sensor or not
Woodensteer	Bit	The car has steer which is made from wood
Locklessentry	Bit	Car opens doors without using key
Leatherwoodensteer	Bit	Car has steer which is made of leather and wood
Leatherfabricseat	Bit	Car's seat are made of leather and fabric
Leathersteer	Bit	Car has steer which made of leather
Leatherseat	Bit	Car has seats which is leather
Deepchangeblesteer	Bit	Car's steer is changeable to deep formation
Autowindows	Bit	Car's windows are electrical and it can open with one button
Automemoryseat	Bit	information about seat of car and car changes seat deepness with its memory
Crusecontrol	Bit	Car has cruise-control feature in it or not
Hydosteer	Bit	The car has got hydraulic steer or not
Heatedsteer	Bit	Car has got heated steer features in it or not
Heatedseat	Bit	Seat of car is heated features in it
Ac	Bit	Air conditioner is exist inside car
Fabricseat	Bit	Seats of car is made of fabric
Startstop	Bit	Start Stop technology exist in car
Sunroof	Bit	Car has sunroof or not
Phonebluetooth	Bit	Driver can connect their phone to car's music system
Tripcomp	Bit	Car has trip computer system in it
Heightchangeblesteer	Bit	Car's steer is changeable with height information
Alloywheel	Bit	Car's wheels are changed to the alloy
Windowroof	Bit	Car's roof is made of glass
Autosidemirror	Bit	When the car is parking its side mirrors are closing

Table 6.9 Features table(cont.)

Headlighsign	Bit	Car's dashboard is showing the information about lights on or not
Autowashheadlight	Bit	When the dirt is come to the lights it is automatically cleans itself
Modified	Bit	Seller of this car has done modifies to the car
Foglights	Bit	Car got fog lights in it or not
Xenon	Bit	Car got Xenon lights in it or not
Rainsensor	Bit	Car got rain sensors on its lights or not
Cdplayer	Bit	Car got CD player in it or not
Cdchanger	Bit	Car got CD changer in it or not
Ipodconnector	Bit	Ipods can connect to car's music system
Mpthreeplayer	Bit	Car can play mp3 format music in it
Radio	Bit	Car has radio in it or not
Navi	Bit	Navigation system exists in the car or not
Fourwheel	Bit	Car has four wheel drive system in it
Turbo	Bit	The car has turbo power system in its engine

6.2 Class Diagram

This section shows the design of the system with class diagram. It describes the classes by giving the diagram in Figure 6.2.

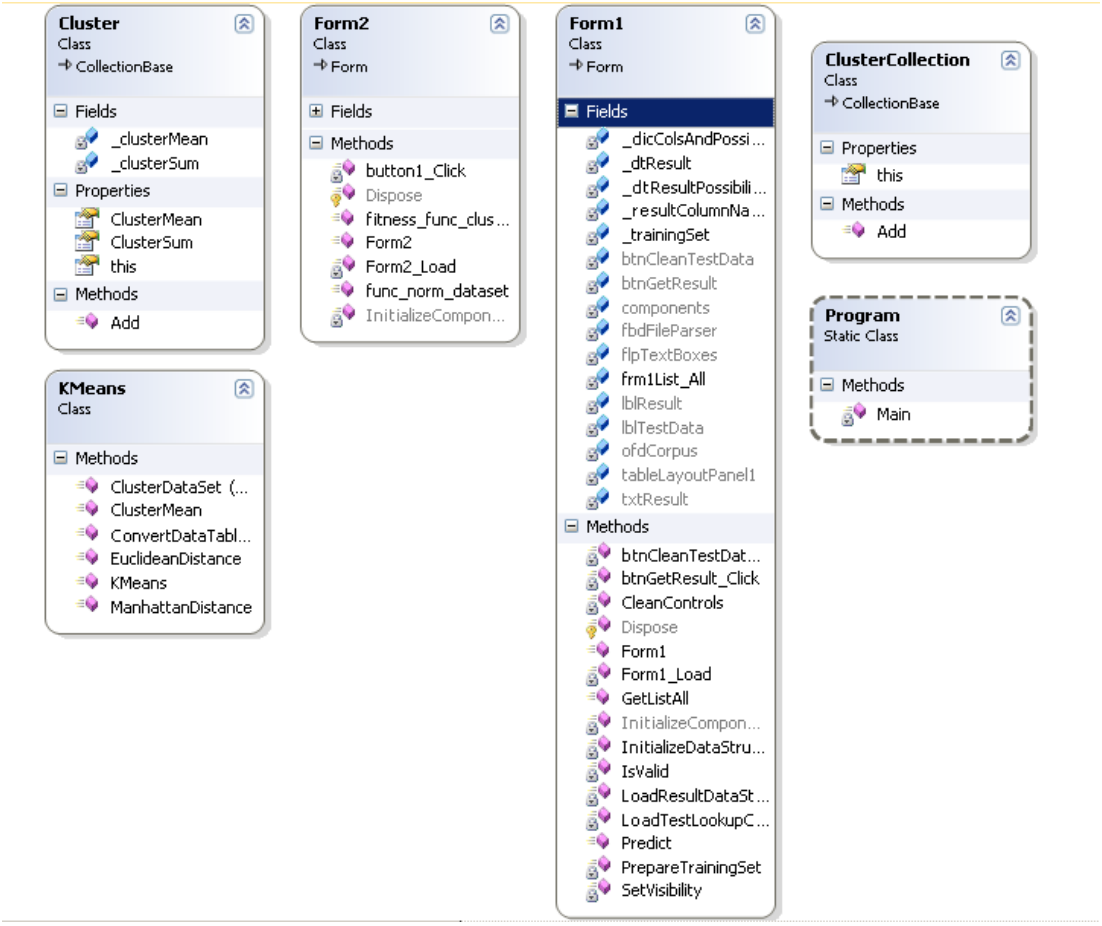


Figure 6.2 Class diagram

6.3 Sequence Diagram

This section presents the sequence diagram drawn in this study. It shows the sequence of system with the sequence diagram. In Figure 6.3, the sequence diagram of the system is shown until the collection of data. In Figure 6.4, the sequence diagram of the system is shown which is finding the ideal value for user.

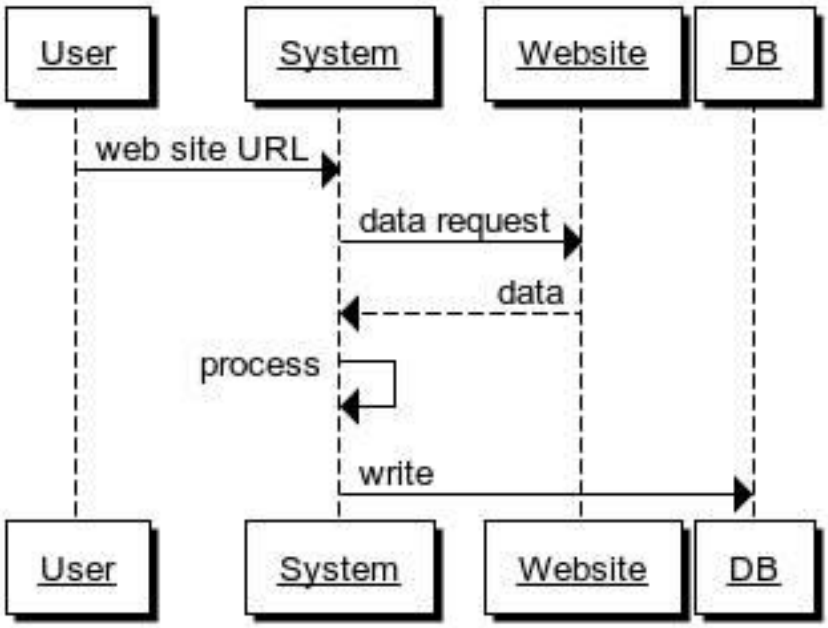


Figure 6.3 Sequence diagram for data collection

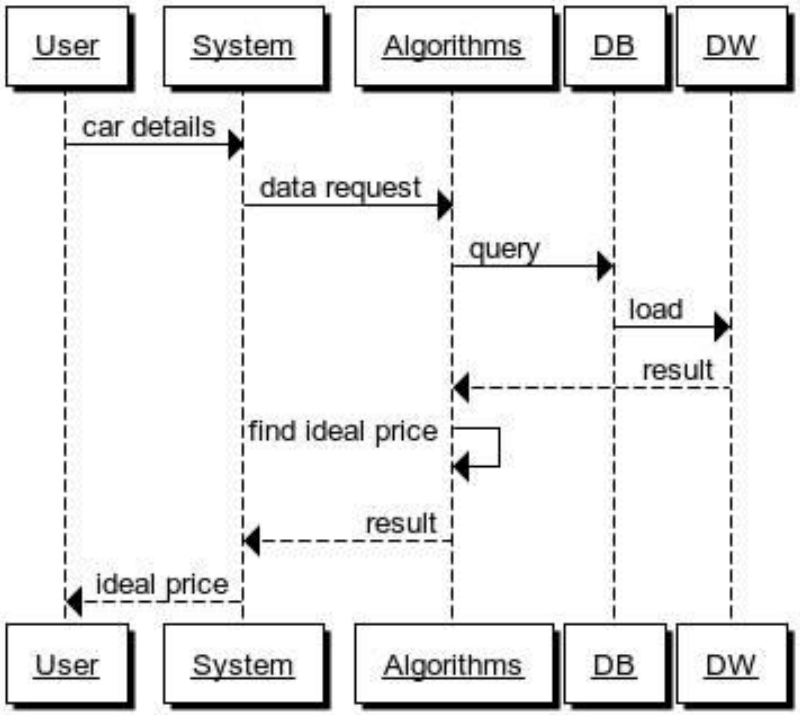


Figure 6.4 Sequence diagram for finding the ideal value

CHAPTER SEVEN
IMPLEMENTATION

7.1 Car Price Prediction System

Accurate prediction of car sales price is important in the operation of the automotive market. Car sellers and buyers wish to know a fair value for their cars. A precise estimate of the sales price of a car is of real importance to buyers who face choices among several cars.

Some experts try to obtain an accurate estimate of the market value, but they may require a cost for this service, they may consider only several factors and they may decide a price according to their profits. In addition, inaccurate appraisal of car values results in low earnings for sellers or money loss for buyers. For this reason, in this thesis, a car price prediction system was developed to predict a fair value for a car according to its features, using the real data collected from past.

However, the accurate prediction of the car price is difficult because cars are typically sold as a package of various factors, such as location, mark, model, year, kilometer, color etc. It is not obvious how to account for these factors in predicting the selling price of a car. The solution of this problem is data mining. Because data mining automates the process of discovering complex relationships in data and finding predictive information.

Figure 7.1 shows decision support processes followed in our system. After data collection, data preprocessing and data mining steps, prices suggestion is given to user according to his/her inputs.

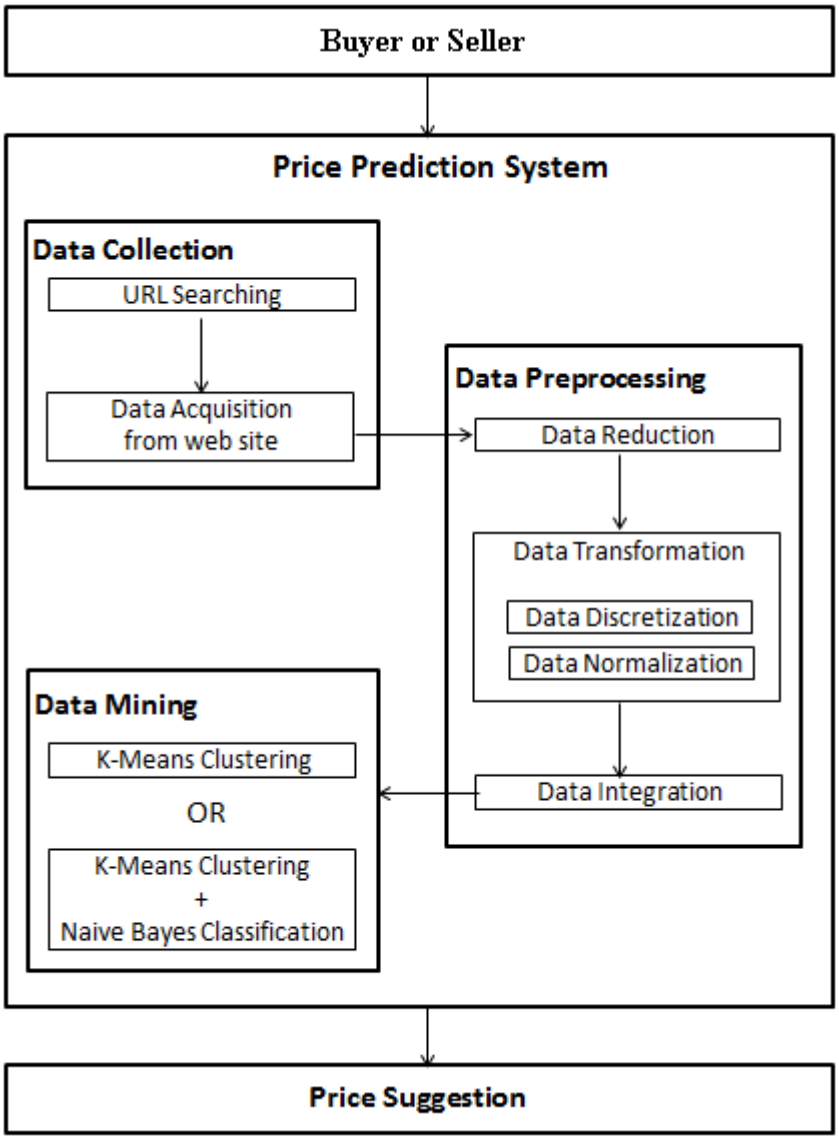


Figure 7.1 Screenshot of the database

7.2 Tools

In this section, development tools and database system used to implement the system is explained. These are MS SQL Server to manage database, MS Visual Studio to write code, MS Visio to draw diagrams and MS Expression Design to draw design diagrams.

7.2.1 MS SQL Server Database

In this thesis, MS SQL Server was used to keep data and to manage the database. It is one of the best database systems to develop critical applications in business and technical areas. It has a strong management tool, named Management Studio, to create tables, procedures, triggers and functions. Figure 7.2 shows the screenshot of the databases created in this thesis.

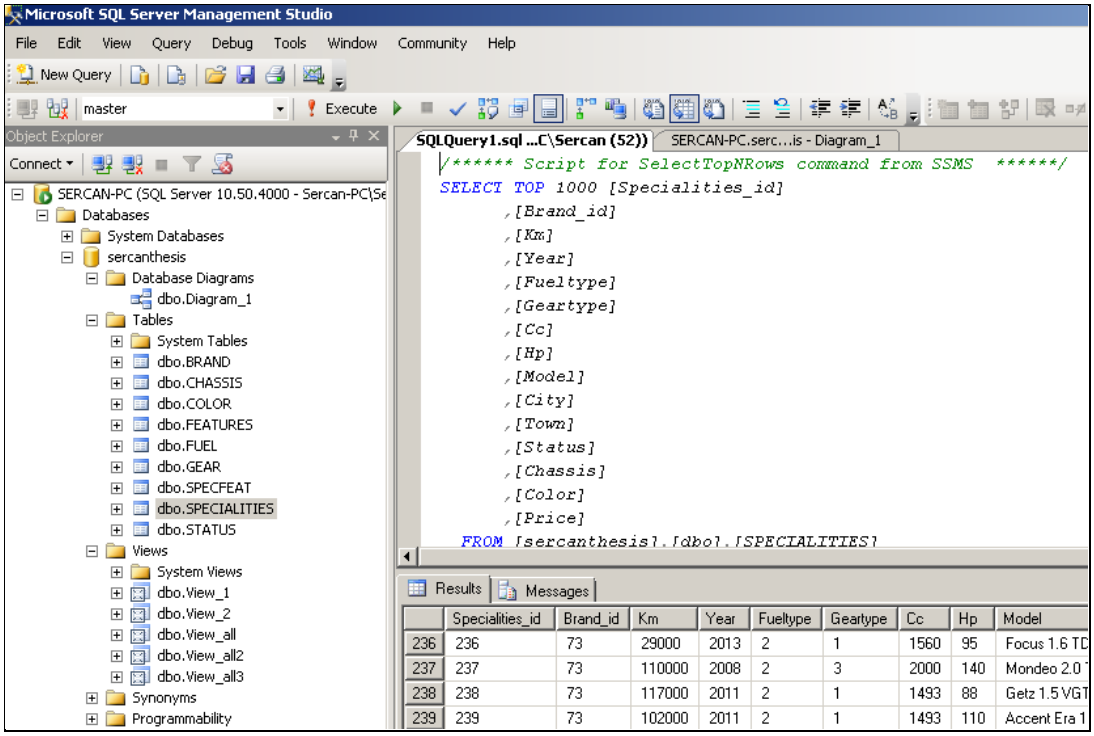


Figure 7.2 Screenshot of the database

7.2.2 MS Visual Studio

MS Visual Studio is an IDE (Integrated Development Environment) tool which is used to develop web, mobile and desktop applications. Machine learning algorithms were coded in this platform using C# programming language. Figure 7.3 shows the screenshot from our application in Visual Studio platform.

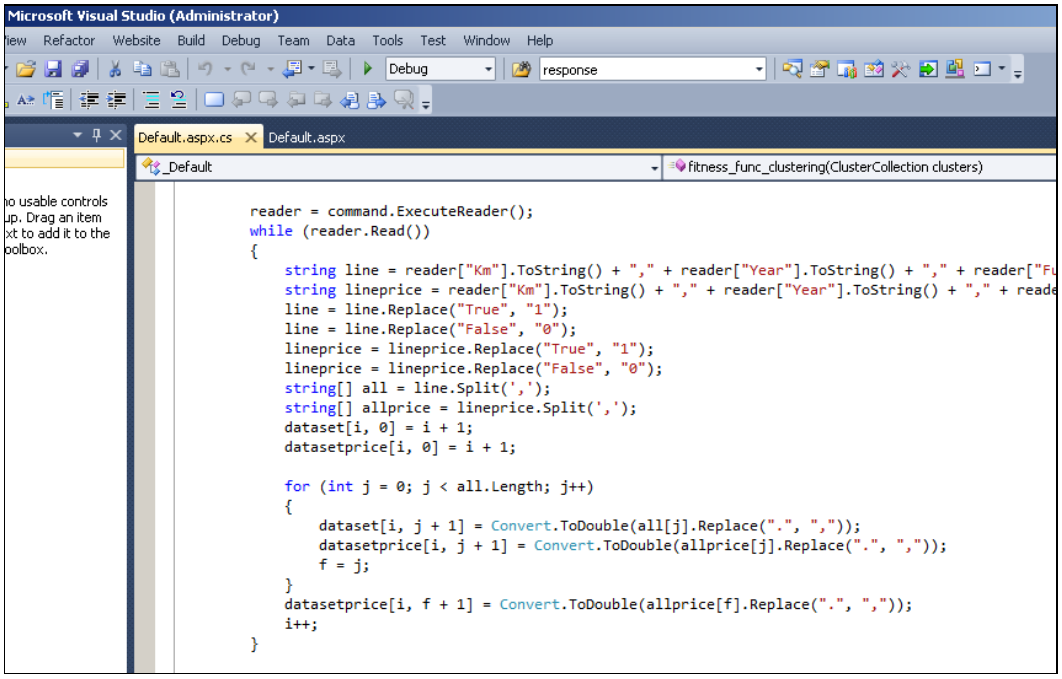


Figure 7.3 Screenshot from Visual Studio

7.2.3 MS Visio

MS Office Visio provides a platform to draw diagrams, flowcharts, maps, scheduling diagrams, detailed network diagram, industrial control systems and others systems diagrams. At this project, diagrams and charts are drawn by this software.

7.2.4 MS Expression Design

In this study, design process is done by using Microsoft Expression Design program. This software usually uses for drawing designs and diagrams in software development studies.

7.3 Proposed Model

In this thesis, we propose an approach to estimate the market value of a used-car depending on real data collected from past which include the sales price and information on the car such as its age, mileage, fuel type, gear type, engine size, etc. Our approach automatically discovers the association between the sales price and car

attributes using machine learning techniques. Proposed solution, in this thesis, does not depend on the input space dimension; therefore it is suitable to solve a high dimensional problem.

Two applications have been developed to compare our proposed model with the classic model for price prediction. In the applications, different approaches was applied of data collected form a web site and then accuracy rates were calculated using a validation method.

In the first approach as shown in Figure 7.4a, only K-Means clustering algorithm is applied on data and then the closest group to the current car, whose price is to be forecasted, is used for prediction. The recurrent problem of finding the value of k is solved by employing Elbow method.

In the second approach, which is proposed in this thesis, (i) firstly data is partitioned into groups of similar instances by K-Means clustering algorithm and (ii) after clustering, Naïve Bayes classification algorithm is applied on clustered data for the prediction, as shown in Figure 7.4b.

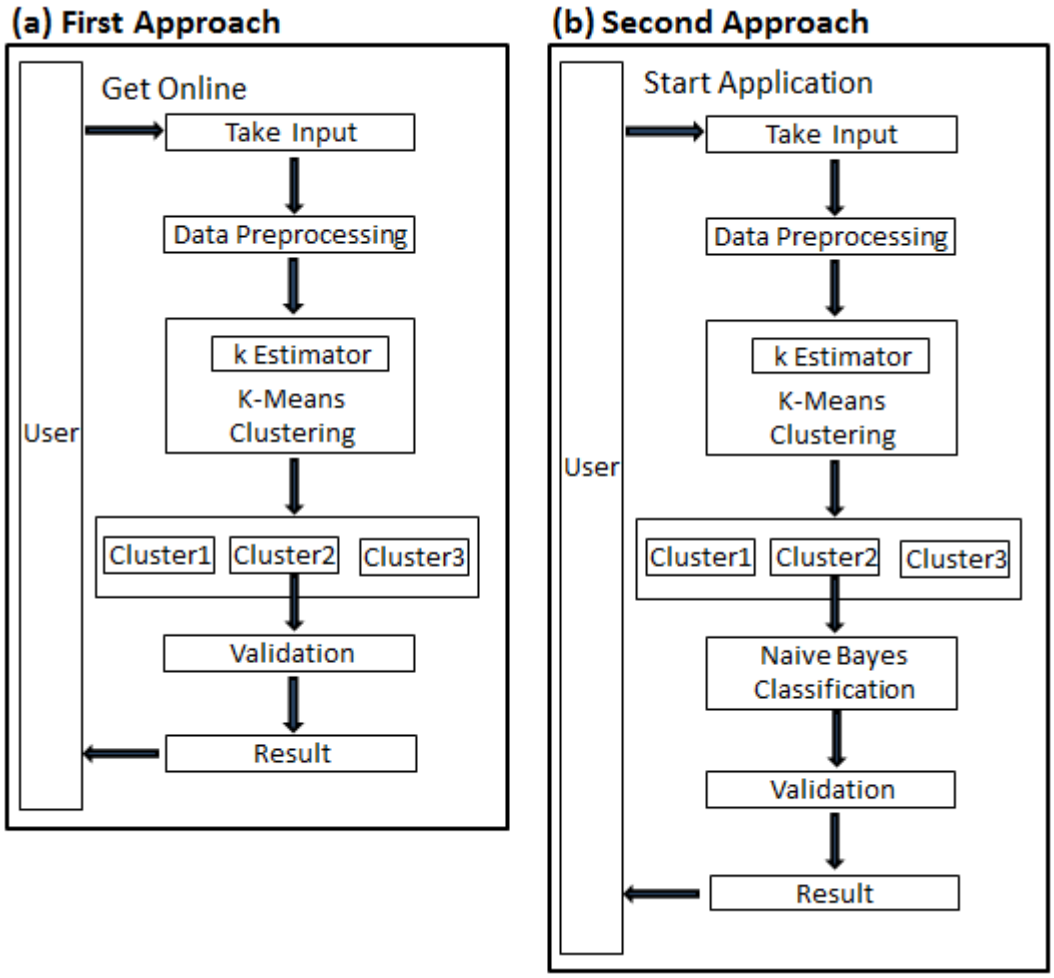


Figure 7.4 Block diagrams (a) for the first approach and (b) for the second approach proposed in this thesis

7.3.1 First Approach: K-Means Clustering

When K-means algorithm is applied, user has to decide the value of *k*, which is the number of clusters. It is an important issue for actually solving the clustering problem. The optimal choice of *k* is often ambiguous, because increasing the value of *k* continuously decreases the error and increases the computation time.

In this thesis, we use elbow method to estimate the value of *k*. The idea of this method is to choose the *k* at which the decrease in Sum of Squared Error (SSE) slows down. Elbow point can also be calculated by checking the percentage of variance measure against the number of clusters.

Table 7.1 shows SSE values measured according to the number of clusters in our experiment. Elbow point occurs after ten numbers of clusters as shown in Figure 7.5. Because SSE value at point nine is 18.58, it suddenly drops at the point ten (10.68), but after that the decrease slow down as 9.85 at point eleven, 9.67 at point twelve, 9.14 at point thirteen etc. So we divided the input dataset in ten clusters considering a set of attributes i.e. kilometer, year, fuel-type, gear-type etc.

Table 7.1 SSE values calculated according to the number of clusters

Number of Clusters	SSE Values
1	180.56
2	139.96
3	117.84
4	84.73
5	62.19
6	50.75
7	37.98
8	26.59
9	18.58
10	10.68
11	9.85
12	9.67
13	9.14
14	8.93
15	8.83

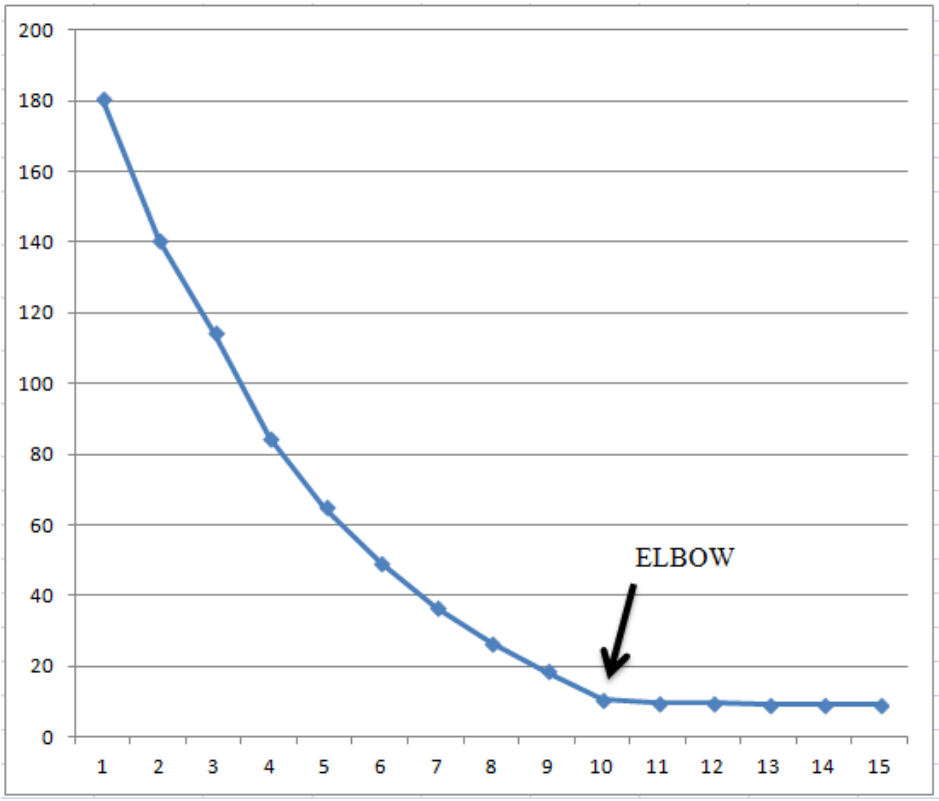


Figure 7.5 Elbow point for this study

After applying k-means algorithm, the percentages of data on each cluster are given in Table 7.2.

Table 7.2 Percentage of data on each cluster

Cluster ID	Percentage of Data
1	13%
2	6%
3	19%
4	27%
5	3%
6	8%
7	4%
8	13%
9	5%
10	2%

A screenshot from the web application developed to apply only K-Means algorithm on the data is shown in Figure 7.6. This application requires information about a car as shown in Figure 7.7 and then gives a fair price interval as shown as in Figure 7.8. After clustering, the closest cluster to the current instance is found, and then price interval is predicted according to the prices in this cluster.

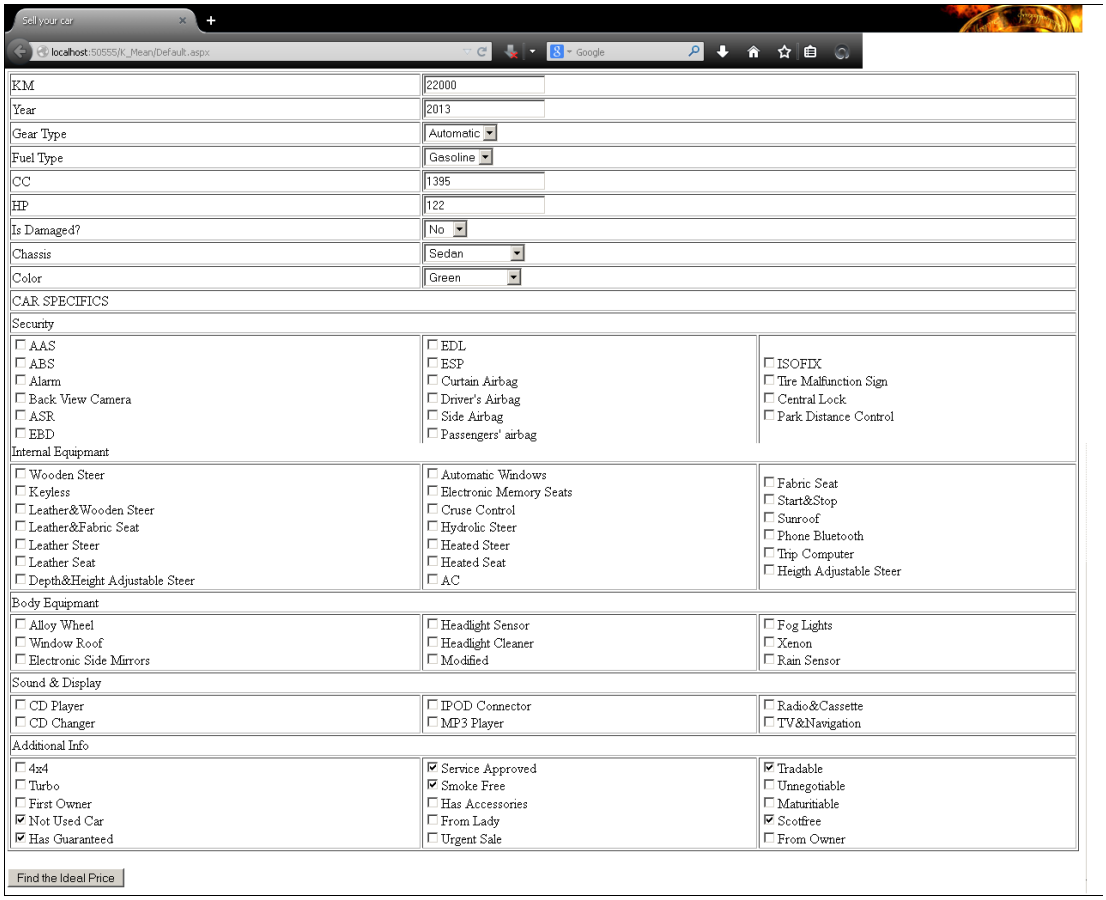


Figure 7.6 A screenshot from the web application developed in this thesis

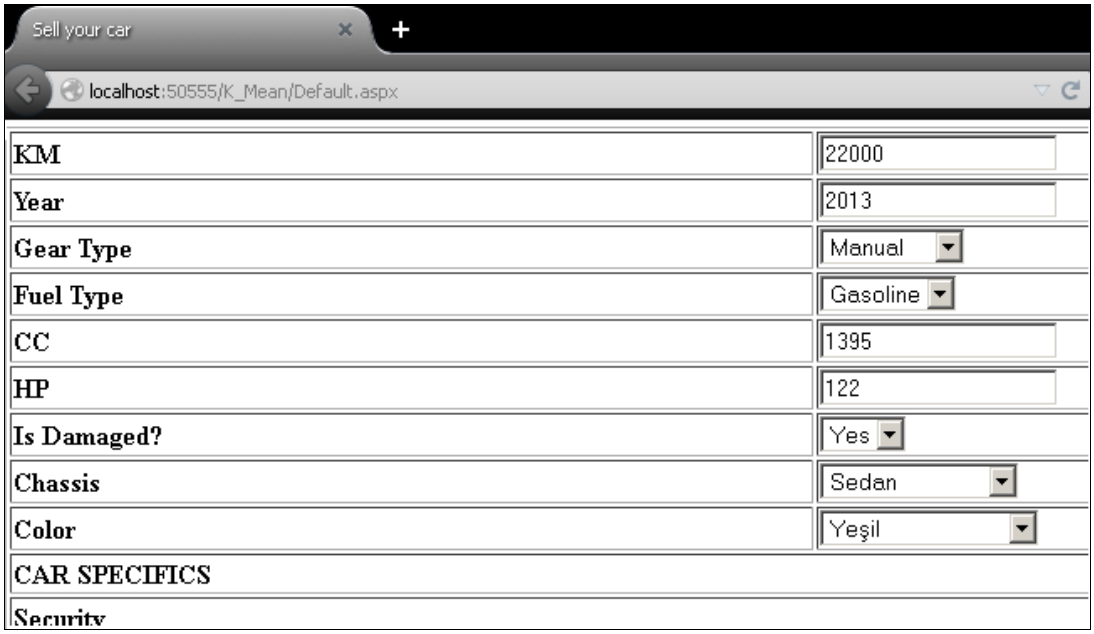


Figure 7.7 Application developed for the first approach

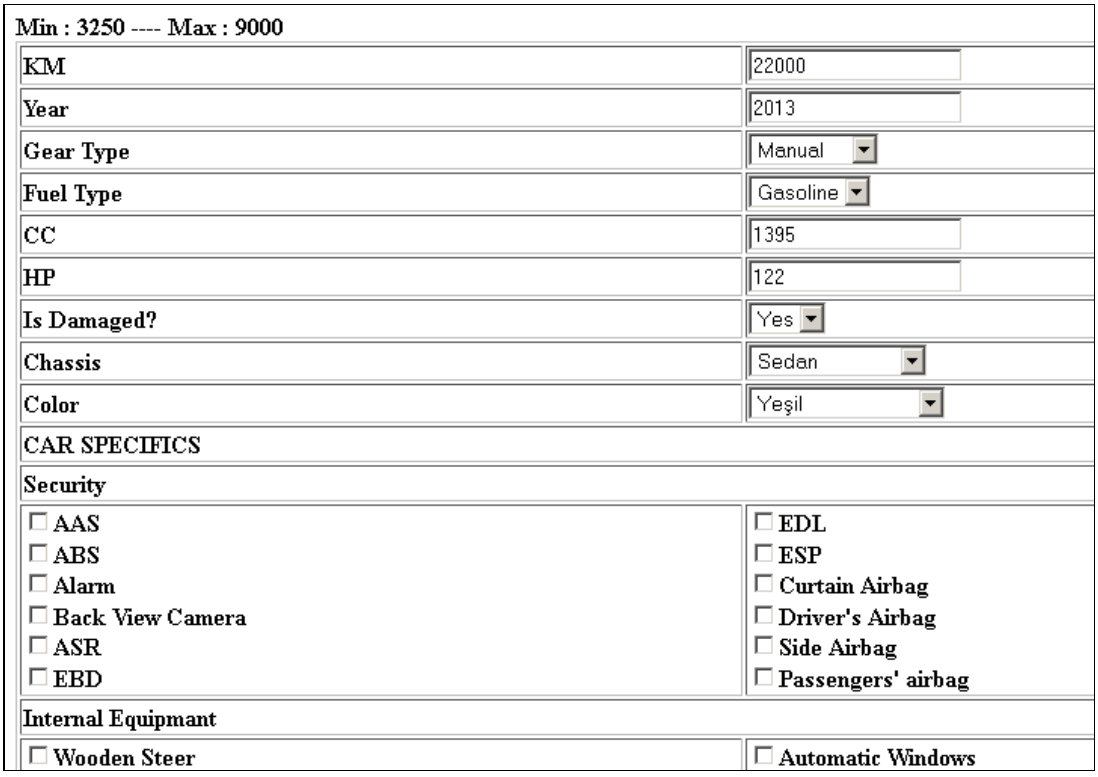


Figure 7.8 The result obtained from the first application

In the application, a web form was created that allows anyone to enter desired car parameters which are evaluated using past data stored in the database, resulting in a

price estimate that is returned to the user. Using this application, users can be able to predict the price of a car correctly based on past data.

When K-Means algorithm applied on data, it is necessary to use a distance measure. In this experiment, Euclidean distance was used. Written code to calculate Euclidean distance during K-Means algorithm is shown in Figure 7.9.

```
public static double EuclideanDistance(double [] X, double []Y, double[] coefficients)
{
    int count = 0;

    double distance = 0.0;

    double sum = 0.0;


    if(X.GetUpperBound(0) != Y.GetUpperBound(0))
    {
        throw new System.ArgumentException("the number of elements in X must match the number
    }
    else
    {
        count = X.Length;
    }

    for (int i = 1; i < count; i++)
    {

        sum = sum + Math.Pow(Math.Abs(X[i] - Y[i]), 2.0) * coefficients[i-1];
    }

    distance = Math.Sqrt(sum);

    return distance;
}
```

Figure 7.9 Finding the Euclidean distance

7.3.2 Second Approach: K-Means Clustering + Naïve Bayes Classification

In the second approach, price is predicted by applying Naïve Bayes classification algorithm on the target cluster which is generated by applying k-means algorithm on the dataset.

An application was developed to apply K-means and Naïve Bayes techniques consecutively. A screenshot from the application is given in Figure 7.10. This application also takes information about a car from the user and gives an interval for a fair price value for the car. But, differently from the previous application, first of

all, K-Means technique is applied, after K-means result are taken as a cluster, Naive Bayes technique is applying to just in this cluster data. So that, results are getting better, the reason of that Naïve Bayes technique is applied on data which is smaller and have similar values.

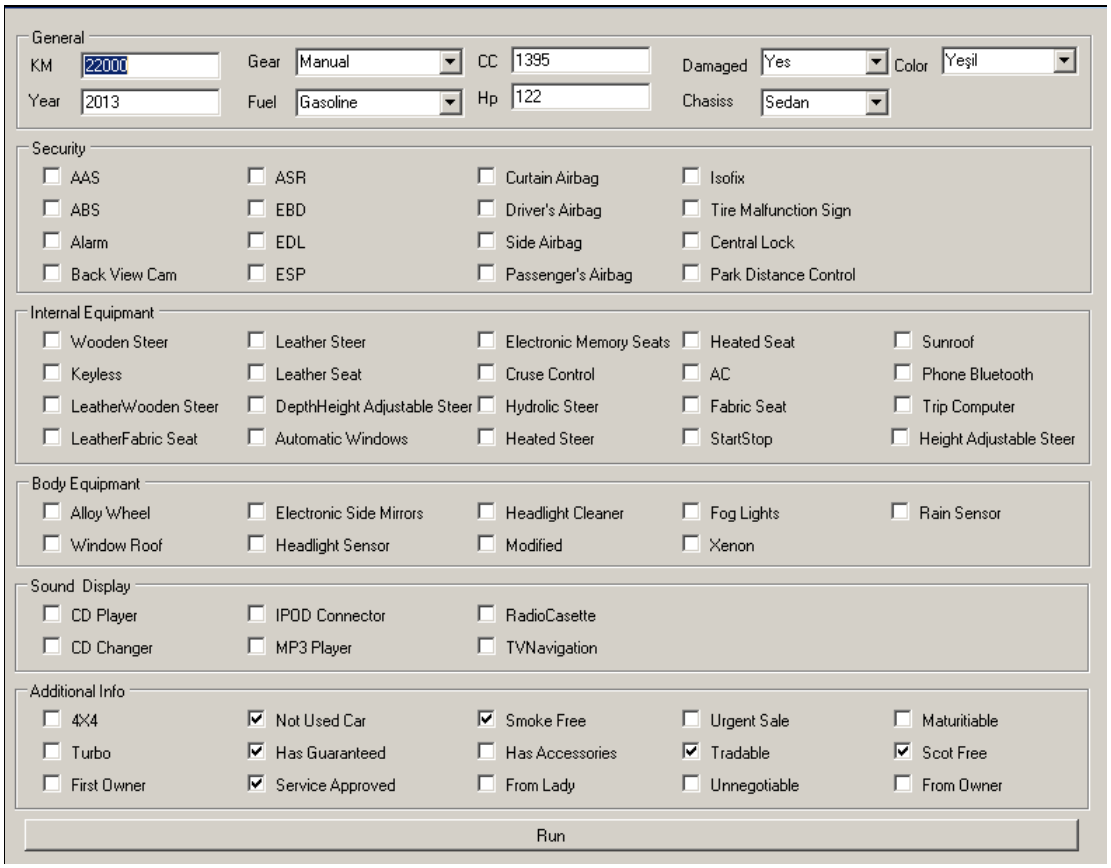


Figure 7.10 Application developed for the second approach

After taking information about a car from the user, the application automatically gives an interval for a fair price value for the car. Figure 7.11 shows the result screen of the second application which includes the usage of two machine learning techniques one after another. It gives the results to users as an interval like the first application. But it provides better predictions then the first application.

Figure 7.11 The result obtained from the second application

The second application also calculates SSE values before separating to clusters for deciding how many clusters is ideal for applying K-Means technique. A function given in 7.12 was written in the code to calculate SSE values to determine the optimal value of *k*. The inner *for* structure in the code calculates the sum of the squared differences between group mean and each observation in a single group. The outer *for* structure loops as the number of clusters and sums the local SSE values of all clusters to find the global SSE value.

```
public double SSE_Calculation(ClusterCollection clusters)
{
    double SSE = 0;
    for (int i = 0; i < clusters.Count; i++)
    {
        double sum = 0;
        for (int j = 0; j < clusters[i].Count; j++)
        {
            for (int k = 1; k < clusters[i].ClusterMean.Length;k++)
                sum += Math.Pow(Math.Abs(clusters[i].ClusterMean[k]
                                        - clusters[i][j][k]), 2);
        }

        SSE += Math.Sqrt(sum);
    }

    return SSE;
}
```

Figure 7.12 Finding the SSE value

CHAPTER EIGHT
EXPERIMENTAL WORK

8.1 Application of Different Approaches

In the experimental works, the proposed approach was successfully applied for used-car price prediction as a case study. This study deals with the prediction of price variable according the car features such as model, year, kilometer etc. Because, without any knowledge, the seller may sell his/her used-car with a small profit. In order to ensure a reasonable profit, the seller needs to be able to predict the price fairly. The aim of this study is to develop an application to help sellers in this process.

Selecting the right machine learning technique for any given problem is critical. In analytics projects, often we use several techniques and algorithms to test on various data problems. Since the problem involved in this experimental work is predicting of price values, clustering and classification techniques can be used. Many algorithms can implement clustering technique such as K-Means, DBSCAN, Self-Organized Maps, Single-Link Hierarchical Clustering which vary in terms of complexity of implementation and scope of the problem. At this point, we selected K-Means algorithm, because the general idea of this algorithm is more relevance here. Many algorithms can implement classification technique such as Neural Networks, C4.5 Decision Tree, Naïve Bayes and Support Vector Machine. It is necessary to know the underlying concepts while selecting and using these techniques in real-time projects. At this point, we selected Naïve Bayes algorithm, because it has successfully proved its capability in generating a good prediction model.

8.2 Comparison of Results and Discussion

In this thesis, two different approaches were demonstrated on the same data to determine the better one. The outcome of the proposed model with classification is compared with the classic (without classification) model for price prediction. The

model is evaluated by a validation technique. In this section, the comparison results are given related with two approaches.

Figure 8.1 shows how Naïve Bayes algorithm is applied after K-Means algorithm. Classification algorithm is not applied on whole data; instead, it is applied to the data in the closest cluster to the current instance.

Figure 8.1 Target area of classification method applied

Table 8.1 shows the comparison between two approaches on the test set in this experiment. Firstly, only K-Means clustering algorithm was used for finding the fair values of cars and its success rate was measured as 62%. Secondly, K-Means clustering and Naïve Bayes classification algorithm were used consecutively for observing the differences between the separately used techniques and using techniques together. In this case, the success rate is measured as 86%. That means: applying more than one machine learning technique or algorithm can provides better success rates than applying only just one technique or algorithm.

Table 8.1 Comparison of two approaches according to accuracy rates

Approach	**Accuracy**
K-Means Clustering	62%
K-Means Clustering + Naïve Bayes Classification	86%

According to the experimental works, it is discovered that using hybrid combinatorial method of clustering and classification may give better results than

using one technique. Error can be reduced by applying classification after clustering because it focuses on similar products based on their characteristics, which means clustering minimizes the model complexity. In conclusion, executing more than one machine learning techniques on specific data can provide better prediction than executing just one machine learning technique on that same data. In this thesis, that is proven that higher accuracy can be achieved using different approaches. In other words, this study proves that the prediction accuracy can be improved substantially by applying two machine learning algorithms one after another.

The findings in this thesis would be interesting in future studies to be able to predict car prices better. Because, the car sellers need a professional price calculation, in order to avoid the expensive cost of wrong prediction. The application developed in this thesis can provide this, because it yields a high accuracy.

CHAPTER NINE
CONCLUSION AND FUTURE WORK

9.1 Conclusion

The aim of this study is to design and develop a price prediction system which finds an ideal or closest to ideal value for an instance. In this thesis, predicting price of an instance has been stated as a machine learning problem and has been solved using the combination of clustering and classification techniques.

Clustering and classification techniques were used in this study and the models constructed with these techniques were evaluated in terms of accuracy. A price prediction system was designed to helps buyers and sellers to make effective decision.

In the experimental works, used-car price prediction system was implemented to demonstrate the success of the proposed approach. From a website, data were collected on the previous sale prices of used-cars. Our application automatically gathers data for past car records and stores that data in a database. The data include the sales price and information on the car such as its model, kilometer, fuel type, engine size, gear type etc. Before applying machine learning techniques, gathered data undergone a set of data operations, such as data reduction, data integration and data transformation such as discretization and normalization. All these operations are collectively referred to as data pre-processing. These steps provide data to be transformed in relevant format which is then given as input to various algorithms.

In order to compare two different approaches, two applications are developed and they are compared in terms of accuracy. While, the first application only runs K-Means algorithm on collected data, the second one runs K-Means and Naive Bayes algorithms and builds a better price prediction model. Experimental results show that using more than one machine learning techniques can give better results than using just one machine learning technique.

In this thesis, the price prediction with the combination of two machine learning techniques has been proved as better, instead of using only one technique. The proposed model has successfully proved its capability in generating a good prediction. With the right parameter setting, it achieves a better accuracy than the standard solution. This knowledge may make a significant contribution to computer science, especially to data mining works.

9.2 Future Work

For the future work, this study can applied any kind of data, instead of car data. In the case of used cars price prediction example, determining value of a used car based on a variety of characteristics such as model, engine, color etc. When developing a new application for another problem, other characteristics can be used related to the problem.

Moreover, the model construction takes place periodically, when the model needs to be updated because of inserting new data.

In addition, the application can be developed as a mobile application to make easy way to use in anywhere. Mobile application can be developed for every major platform such as Android, iOS, Mobile Phone or can be developed as a mobile web application independently from the platform.

Moreover, the model proposed in this study can be improved by using more machine learning techniques consecutively or applying different algorithms such as neural network, C4.5 decision tree or support vector machine.

REFERENCES

Alkhatib, K., Najadat, H., Hmeidi, I., & Shatnawi, M.K.A. (2013). Stock price prediction using k-nearest neighbor algorithm. *International Journal of Business, Humanities and Technology*, *3*, 32-44.

Bin, O. (2004). A prediction comparison of housing sales prices by parametric versus semi-parametric regressions. *Journal of Housing Economics*, *13*, 68-84.

Chan, H.C.C. (2008). Intelligent spider for information retrieval to support mining-based price prediction for online auctioning. *Expert Systems with Applications*, *34*, 347-356.

Choudhury, S., Ghosh, S., Bhattacharya, A., Fernandes, K., & Tiwari, M. (2014). A real time clustering and SVM based price-volatility prediction for optimal trading strategy. *Neurocomputing, 131,* 419-426.

Cortez, P., (2012). Data mining with multilayer perceptrons and support vector machines. *Data Mining: Found & Intell Paradigms, 24*, 9-25.

Crawford, M.M., Tuia, D., & Yang, H.L. (2013). Active learning: Any value for classification of remotely sensed data? *Proceedings of the IEEE, 101*, 593-608.

Decherchi, S., Gastaldo, P., Leoncini, A., & Zunino, R. (2012). Efficient digital implementation of extreme learning machines for classification. *IEEE Transactions on Circuits and Systems, 59*, 496-500.

Devi, I.M., Rajaram, R., & Selvakuberan, K. (2007). Machine learning techniques for automated web page classification using URL features. *International Conference on Computational Intelligence and Multimedia Applications, 2*, 116-118.

Gacovski, Z., Kolic, J., Dukova, R. & Markovski, M. (2012). Data mining application for real estate valuation in the city of Skopje. *ICT Innovations*, 537-538.

Gu, L., (2012). Two semi-supervised locality sensitive k-means clustering algorithms by Seeding. *IEEE Fifth International Conference on Advanced Computational Intelligence*, 296-299.

Guo, K., Wang, J., Shi, G., & Cao, X. (2012). Cluster analysis on city real estate market of China: based on a new integrated method for time series clustering. *Procedia Computer Science, 9*, 1299-1305.

Huang, G., Zhou, H., Ding, X., & Zhang, R. (2012). Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics*, *42*, 513-529.

Kisilevich, S., Keim, D., Rokach, L., (2012). A gis-based decision-support system for hotel room rate estimation and temporal price prediction: The hotel brokers context. *University of Konstanz: Data Mining, Database and Visualization, Department of Computer and Information Science.*

Li, W., Guo, Q., & Elkan, C. (2011). A positive and unlabeled learning algorithm for one-class classification of remote-sensing data. *IEEE Transactions on Geosciences and Remote Sensing*, *49*, 717-725.

Liang, J., Bai, L., Dang, C., & Cao, F. (2012). The k-means-type algorithms versus imbalanced data distributions. *IEEE Transactions on Fuzzy Systems, 20*, 748-745.

Lima, B., & Machado, V. (2012). Machine learning algorithms applied in automatic classification of social Network Users. *IEEE Computational Aspects of Social Networks* 58-62.

Lin, H.H., Liu, T. & Chuang, J. (2009). Learning a scene background model via classification. *IEEE Transactions on Signal Processing, 57*, 1641-1654.

Listani, M. (2009). *Support vector regression analysis for price prediction in a car leasing application*. Master Thesis, Hamburg University of Technology, Hamburg.

Nasira, G.M., & Hemageetha, N. (2012) Forecasting model for vegetable price using back propagation neural network. *International Journal of Computational Intelligence and Informatics, 2*, 110-115.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of EMNLP*, 79-86.

Rahman, M., Antani, S.K., & Thoma, G..R. (2011). A learning-based similarity fusion and filtering approach for biomedical image retrieval using SVM classification and relevance feedback. *IEEE Transactions on Information Technology in Biomedicine, 15*, 640-646.

Raykhel, I., & Ventura, D., (2009). Real-time automatic price prediction for eBay online trading. *Proceedings of the Twenty-First Innovative Applications of Artificial Intelligence*, 135-140.

Sahoo, S.K., & Makur, A. (2013). Dictionary training for sparse representation as generalization of k-means clustering. IEEE Signal Processing Letters, *20*, 587-599.

Segaran, T., (2007). *Programming collective intelligence: building smart Web 2.0 applications*. Sebastopol, CA: O'Reilly Media.

Shumeli, G., Patel, N.R., & Bruce, P.C., (2007). *Data mining in excel: Lecture notes and cases*. Resampling Stats, Inc.

Tao, Q., Sun, Z., & Kong, K. (2012). Developing learning algorithms via optimized discretization of continuous dynamical systems. *IEEE Transactions on Systems, Man, and Cybernetics, 42*, 140-149.

Thuy, T.T., Nguyen, A., (2008). A survey of techniques for internet traffic classification using machine learning. *IEEE Communications Surveys and Tutorials, 10*, 56-74.

Vineyard, C.M., Heileman, G.L., Verzi, S.J., & Jordan, R. (2012). Game theoretic mechanism design applied to machine learning classification. *International Workshop on Cognitive Incromation Processing* 1-5.

Wang, X., Wen, J., Zhang, Y., & Wang, Y. (2014). Real estate price forecasting based on SVM optimized PSO. *Optik-International Journal for Light and Electron Optics, 125*, 1439-1443.

Yusuf, L. M., Othman, M. S., & Salim, J. (2010). Web classification using extraction and machine learning techniques. *IEEE Information Technology International Symposium, 2*, 765-770.

Zhang, W.B., & Ji, H.B. (2013). Fuzzy extreme learning machine for classification. *Electronics Letters, 49*, 448-450.

Ziemniak, T. (2011). Use of machine learning classification techniques to detect atypical behavior in medical applications. *Sixth International Conference on IT Security Incident Management and IT Forensics*, 149-162.