NEAR EAST UNIVERSITY
GRADUATE SCHOOL OF SOCIAL SCIENCES
BUSINESS ADMINISTRATION PROGRAM

COMPARISON OF TWO DIFFERENT APPROACHES FOR FORECASTING ACCOMMODATION ESTIMATIONS

MEHMET EMİN AKKAYA

PHD THESIS

NICOSIA
2020

COMPARISON OF TWO DIFFERENT APPROACHES FOR FORECASTING ACCOMMODATION ESTIMATIONS

MEHMET EMİN AKKAYA

NEAR EAST UNIVERSITY
GRADUATE SCHOOL OF SOCIAL SCIENCES
BUSINESS ADMINISTRATION PROGRAM

PHD THESIS

THESIS ADVISOR

ASSOC. PROF. DR. İHSAN TOLGA MEDENİ

NICOSIA

2020

ACCEPTANCE/APPROVAL

We as the jury members certify the ‘Comparison of Two Different Approaches for Forecasting Accommodation Estimations’ prepared by the Mehmet Emin Akkaya defended on 24/01/2020 has been found satisfactory for the award of degree of Phd

JURY MEMBERS

.....

Assoc. Prof. İhsan Tolga Medeni (Supervisor)
Ankara Yıldırım Beyazıt University
Faculty of Business Administration, Management Information Systems Department

.....

Prof. Dr. Şerife Eyüpoğlu (Head)
Near East University
Faculty of Economics and Administrative Sciences, Department of Business Administration

.....

Prof. Dr. Tülen SANER
Near East University
Faculty of Tourism, Department of Tourism and Hotel Management

.....

Prof. Dr. Hakkı Okan YELOĞLU
University Name and Department

.....

Prof. Dr. Oğuz ÖZYARAL
University Name and Department

.....

Prof. Dr. Mustafa Sağsan
Graduate School of Social Science
Director

DECLARATION

I undertake that the thesis I have prepared is entirely my own work and that I provided reference for each citation. I confirm that I have allowed the paper and electronic copies of my thesis to be kept in the archives of the Near East University Graduate School of Social Sciences under the conditions stated below.

- ☐ My entire thesis can be accessed from anywhere.
- ☐ My thesis can only be accessed at the Near East University.
- ☐ I do not want my thesis to be accessible for two (2) years. If I do not apply for an extension at the end of this period, my entire thesis may be become accessible.

Date

Signature

Mehmet Emin AKKAYA

ACKNOWLEDGMENTS

Sales estimates are an indispensable input for enterprises and are important for the planning of enterprises. Data mining is needed for sales forecasts. In a prospective study with time series analysis, it is aimed to find out what kind of analyzes can be made by using the data mining methods of the accommodation facilities and the important points that should be taken into consideration in these analyzes and how they can interpret the results. In this study; When tourism planners and managers face uncertainty about future short and long-term demands for accommodation facilities that make room sales, the development of correct estimation tools will be very helpful in planning and management, It is expected and expected that it will be able to provide infrastructure to the studies, to make analysis with the recorded data of the accommodation facilities, to reveal the important points to be considered in these analyzes and how to interpret the results. I would like to thank my advisor, Assoc. Prof. İhsan Tolga MEDENİ for guiding and supporting me over the years. You have set an example of excellence as a researcher, mentor, instructor, and role model. I would like to thank my thesis committee members for all of their guidance through this process; your discussion, ideas, and feedback have been absolutely invaluable. I would like to thank Director of the Graduate School of Social Sciences Prof. Mustafa SAĞSAN and I would like to thank the university administration for providing scientific studies with modern methods.

Mehmet Emin AKKAYA

ABSTRACT

COMPARISON OF TWO DIFFERENT APPROACHES FOR FORECASTING ACCOMMODATION ESTIMATIONS

This is a comparative study for the provinces of Antalya, Istanbul and Mugla which has the highest tourist overnight stay numbers of Turkey by time series analyses performed by Multilayer Perceptron and Support Vector Regression analyses methods. Annual data range has been used to estimate the tourism demand. Multivariate data have been used. WEKA 3.8 data mining software has been used in this study in which estimation methods have been implemented; estimations have been made according to the total overnight stay numbers by provinces and these numbers have been compared to the domestic and foreign overnight stay number estimation studies by two different regression methods.

In conclusion, it has been found out in this study performed by multivariate data and different tourism destinations that different regression methods have been used for estimated values to yield the results closest to the actual values and regression analyses have shown variance by tourism destinations and determining the regression method according to the tourism destination planned to be estimated has yielded the estimated values closest to the actual values. It is predicted that the results obtained from this study shall be useful for tourism personnel, researches, investors, tourism executives and tourism planning institutions that applies data mining techniques for tourism demand estimation applications.

Keywords: Accommodation Estimates, Tourism Forecasting, Time Series Regression, Demand Forecasting

ÖZ

ÖNGÖRÜSEL KONAKLAMA TAHMİNLERİ İÇİN İKİ FARKLI YAKLAŞIMIN KARŞILAŞTIRILMASI

Bu çalışma Çok Katmanlı Algılayıcı (Multilayer Perception) ve Destek Vektör Regresyonu (Support Vector Regression) analiz yöntemleri kullanılarak zaman serisi analizleri ile Türkiye'deki en yüksek turist geceleme sayılarına sahip Antalya, İstanbul ve Muğla illeri için karşılaştırmalı bir çalışmadır. Turizm talebini tahmin edebilmek için yıllık veri aralığı kullanılmıştır. Çok değişkenli verilerle çalışılmıştır. Tahminleme yöntemlerinin uygulandığı bu çalışmada WEKA 3.8 veri madenciliği yazılımı kullanılmış olup illere göre toplam geceleme sayıları üzerinden tahminleme çalışması yapılmış, bu rakamlar iki ayrı regresyon yöntemi ile yerli ve yabancı geceleme sayılarının tahminleme çalışmaları ile karşılaştırılmıştır.

Çok değişkenli verilere ve farklı turizm destinasyonlarına göre yapılan bu çalışmada sonuç olarak; tahminlenen değerlerin, aktüel değerlere en yakın sonucu verebilmesi için farklı regresyon yöntemlerinin kullanıldığı ve turizm destinasyonlarına göre regresyon analizlerinin farklılık gösterdiği, turizm destinasyonuna göre regresyon yönteminin belirlenmesinin aktüel değerlere en yakın tahminlenen değerleri ortaya koyacağı sonucu ortaya çıkmıştır. Bu çalışmada elde edilen sonuçların, veri madenciliği tekniklerini turizm talep tahminleme uygulamaları üzerinde uygulayan turizm sektörü çalışanları, araştırmacılar, yatırımcılar, turizm yöneticileri ve turizm planlaması yapan kuruluşlara yarar sağlayacağı öngörülmektedir.

Anahtar Kelimeler: Konaklama Tahminleri, Turizm Tahminleri, Zaman Serileri Regresyonu, Talep Tahmini

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

OLAP :Online Analytical Processing

MLP :Multilayer Perceptron

SVR :Support Vector Regression

KDD :Knowledge Discovery in Databases

SVM :Support Vector Machines

DBMS :Database Management System

CHAID :Chi-square Automatic Interaction Detection

QUEST :Quick Unbiased and Efficient Statistical Tree

QDA :Quadratic Discriminant Analysis

CPI :Consumer Price Index

WEKA :Waikato Environment for Knowledge Analysis

ECB :European Central Bank

GPR :Gauss Process Regression

ID3 :Iterative Dichotomiser 3

CART :Classification and Regression Tree

INTRODUCTION

This chapter includes the scope of the research. In this chapter; problem status, purpose of the research, the importance of research, limitations, definitions, components of the problem, hypotheses, variables are described.

SCOPE OF THE RESEARCH

Problem Status

Hospitality businesses have to make effective decisions about the various problems they face in order to compete under increasing competition conditions. Since future decisions involve uncertainty for businesses, various estimates need to be developed in making decisions. One of them is demand forecasts. Demand forecasts of the enterprises constitute an important input in the marketing strategies to be determined.

Accommodation numbers have the characteristics of an indispensably important input and are important for the planning works of the enterprises. The purpose of the prediction is to predict future situations that accommodation companies may encounter by using different data and techniques and to take action in advance.

In case the factors affecting the number of tourists change, the number of tourists will also change. In cases of uncertainty, enterprises, which fail to make the planning of personnel, inventory management and cost analysis lose and go bankrupt. Estimation study reveals verifiable and acceptable statistical results to develop personnel, inventory management, cost analysis planning and marketing strategy.

Purpose of the Research

Hospitality businesses will need available data and estimation results when making forward-looking decisions. Accurate forecasts are needed to interpret the market. For the closest estimates to real values, the most appropriate regression method should be determined. In this study, it is aimed to determine

the best regression method while making an estimation study for a tourism destination.

Accommodation numbers have the characteristics of an indispensably important input for accommodation businesses and are important for the planning works of the enterprises. Data mining is required for predictive estimation studies. In the study, which qualifies as a forecasting of the future with time series analyses, the aim is to reveal what kind of analyses the accommodation facilities can make, the important points that they should pay attention to in these analyses and how they can interpret the results. Besides, determining the sales values for the forthcoming years, creating a special sales strategy if the estimations are lower than expected, capacity planning and ensuring the supply-demand balance constitute the sub-objectives of the research. In this study in which regression analysis methods are applied with time series data, estimates of prospective predictive accommodation numbers are expected to be obtained by using the accommodation statistics of the previous years in the provinces of Antalya, Istanbul and Muğla. When the tourism planners and managers of the countries, cities and accommodation enterprises, accommodation facilities operators are faced with uncertainty about the short and long term demands regarding the future, it is thought that the development of the right forecasting tools will be very helpful for planning and management, may provide the infrastructure for the studies to be carried out for research purposes in cause and effect relation in the literature on the issue and that they can reveal what kind of analyses the accommodation facilities can make with their registered data, the important points that they should consider in these analyses and how they can interpret the results. With the work of predictive estimation analysis, in case of the loss making of enterprises regarding stock, cost and personnel planning, they are expected to go downsizing or increase capacity, develop strategies of financing, expenditure management and growth before bankruptcy.

The Importance of Research

Accurate demand estimations form the basis of the business decisions regarding tourism and hotel in terms of pricing and business strategies.

Medium and long term tourism and hotel demand estimations are necessary for investment decisions of private sector actors and state infrastructure investments to prevent accommodation enterprises from facing uncertainties.

Given the high level of data gathering, the macro level hotel demand estimation provides useful information to the lodging industry as a whole, although the contribution of such studies is limited. There is an increasing interest in the demand forecast regarding individual hotels based on hotel-specific data.

Estimates for each hotel will benefit hotel practitioners with operational policy implementation, such as reservations by more valued customers, price discrimination, over-reservation policies, late cancellations and early departures.

In the study, the accommodation enterprises are expected to develop the strategies of downsizing or increasing capacity and growing before bankruptcy in case of loss-making regarding stock, cost, purchasing and personnel planning.

When the tourism planners and managers of the countries, cities and accommodation enterprises, accommodation facilities' operators are faced with uncertainty about the short and long term demands regarding the future, it is thought that the development of the right forecasting tools will be very helpful for planning and management, may provide the infrastructure for the studies to be carried out for research purposes in cause and effect relation in the literature on the issue and that they can reveal what kind of analyses the accommodation facilities can make with their registered data, the important points that they should consider in these analyses and how they can interpret the results.

Limitations

Although there are variables that affect tourism diversity, economic and political reasons and variables affecting the estimated number of tourists, this research was limited with the variables of dollar rate, consumer price index, occupancy rates of accommodation facilities, the number of incoming tourists, the average stay times of incoming tourists, the number of overnight stays in

accommodation facilities for the years 2007-2018. Different variables that affect the results of tourism variations, economic, and political reasons and estimated number of tourists were excluded.

Definitions

Data Mining: It is the retrieval of implicit, unclear, previously unknown but potentially useful information from the available data.

Regression Analysis: It is an analysis method used to measure the relationship between two or more variables.

Time Series: It is the series that provides statistical analysis of the data observed at regular intervals over time and forecasting of the data that can be obtained in the future periods.

Estimation (Forecasting): It is the process conducted by making predictions about the future based on past and current data and by analyzing trends most commonly.

Components of the Problem

- Real values of 3 cities with the highest number of accommodation

- The values estimated by regression analysis models over real values of 3 cities with the highest number of accommodation

- Comparative relationship between predicted values and real values

Hypotheses

Hypotheses I:

H0: The estimation values of the accommodation numbers in accommodation facilities in estimation modelling and different regression analysis methods are not different from real values.

H1: Estimation modelling and estimation values in different regression analysis methods of the accommodation numbers in accommodation facilities are different from real values.

Hypotheses II:

H0: As the numerical values of the total number of accommodations in the accommodation facilities increase, the estimation values in different regression analysis methods approach real values.

H1: As the numerical values of the total number of accommodation in the accommodation facilities increase, the estimation values in the different regression analysis methods move away from the real values.

Hypotheses III:

H0: The accommodation densities of the cities Antalya, İstanbul and Muğla, which are the three cities with the highest tourism concentration according to the number of accommodation increase with predictive figures according to estimated values.

H1: The accommodation densities of the cities Antalya, İstanbul and Muğla, which are the three cities with the highest tourism concentration according to the number of accommodation decrease with predictive figures according to estimated values.

1.8 Variables

Independent Variables

-Dollar Exchange Rate

-Occupancy Rates

-Incoming Tourist Numbers

-Average Stay Periods

-Consumer Price Index

Dependent Variables

- Overnight Numbers at the Facility

CHAPTER 1
INTRODUCTION

In this chaper describes the Definition of Data Mining, Data Mining Tool, Task and Process, Data Science, OLAP, Data Warehouse, Knowledge Discovery in Databases, Histograms, Time Series, Decision Tree, Support Vector Machines, Regression, Prediction, Classification, Clustering, Regression, Multilayer Perceptron, Support Vector Regression, Estimation, Prediction, Forecasting issues are explained and reviewed.

1.1 Data Science

Data Science is the discipline of processing and analyzing data for the purpose of extracting valuable knowledge. The term “Data Science” was coined in the 1960’s. However, it really took shape only recently when technology has become sufficiently mature. Various domains such as commerce, medicine and research are applying data-driven discovery and prediction in order to gain some new insights. Google is an excellent example of a company that applies data science on a regular basis. It is well-known that Google tracks user clicks in an attempt to improve the relevance of its search engine results and its ad campaign management. One of the ultimate goals of data mining is the ability to make predictions about certain phenomena. Obviously, prediction is not an easy task. As the famous quote says, “It is difficult to make predictions, especially about the future” (attributed to Mark Twain and others). Still, we use prediction successfully all the time. For example, the popular YouTube website (also owned by Google) analyzes our watching habits in order to predict which other videos we might like. Based on this prediction, YouTube service can present us with a personalized recommendation which is mostly very effective. In order to roughly estimate the service’s efficiency you could simply ask

yourself how often watching a video on YouTube lead you to watch a number of similar videos that were recommended to you by the system? Similarly, online social networks (OSN), such as Facebook and LinkedIn, automatically suggest friends and acquaintances that we might want to connect with. Google Trends enables anyone to view search trends for a topic across regions of the world, including comparative trends of two or more topics. This service can help in epidemiological studies by aggregating certain search terms that are found to be good indicators of the investigated disease. For example, Ginsberg *et al*. (2008) used search engine query data to detect influenza epidemics. However, a pattern forms when all the flu-related phrases are accumulated. An analysis of these various searches reveals that many search terms associated with flu tend to be popular exactly when flu season is happening. Many people struggle with the question: What differentiates data science from statistics and consequently, what distinguishes data scientist from statistician? Data science is a holistic approach in the sense that it supports the entire process including data sensing and collection, data storing, data processing and feature extraction, data mining and knowledge discovery. As such, the field of data science incorporates theories and methods from various fields including statistics, mathematics, computer science and particularly, its sub-domains: Artificial Intelligence and information technology (Lior, 2014).

1.2 Definition of data mining

Data mining has been called exploratory data analysis, among other things. Masses of data generated from cash registers, from scanning, from topic specific databases throughout the company, are explored, analyzed, reduced, and reused. Searches are performed across different models proposed for predicting sales, marketing response, and profit. Classical statistical approaches are fundamental to data mining. Automated AI methods are also used. However, systematic exploration through classical statistical methods is still the basis of data mining. Some of the tools developed by the field of statistical analysis are harnessed through automatic control (with some key human guidance) in dealing with data. A variety of analytic computer models have been used in data mining. The standard model types in data mining

include regression (normal regression for prediction, logistic regression for classification), neural networks, and decision trees. These techniques are well known (Lior, 2014). Data mining, as we use the term, is the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules (Berry and Linoff, 1997).

Data mining is a term coined to d3escribe the process of shifting through large databases in search of interesting and previously unknown patterns. The accessibility and abundance of data today makes data mining a matter of considerable importance and necessity. The field of data mining provides the techniques and tools by which large quantities of data can be automatically analyzed. Data mining is a part of the overall process of Knowledge Discovery in Databases (KDD) defined below. Some of the researchers consider the term “Data Mining” as misleading, and prefer the term “Knowledge Mining” as it provides a better analogy to gold mining Klosgen and Zytkow (2002). Most of the data mining techniques are based on inductive learning Mitchell (1997), where a model is constructed explicitly or implicitly by generalizing from a sufficient number of training examples. The underlying assumption of the inductive approach is that the trained model is applicable to future unseen examples. Strictly speaking, any form of inference in which the conclusions are not deductively implied by the premises can be thought of as an induction. Traditionally, data collection was regarded as one of the most important stages in data analysis. An analyst (e.g. a statistician or data scientist) would use the available domain knowledge to select the variables that were to be collected. The number of selected variables was usually limited and the collection of their values could be done manually (e.g. utilizing hand-written records or oral interviews). In the case of computer-aided analysis, the analyst had to enter the collected data into a statistical computer package or an electronic spreadsheet.

Due to the high cost of data collection, people learned to make decisions based on limited information. Since the dawn of the Information Age, accumulating and storing data has become easier and inexpensive. It has been estimated that the amount of stored information doubles every 20 months Frawley *et al*. (1991). Unfortunately, as the amount of machine-readable information

increases, the ability to understand and make use of it does not keep pace with its growth (Lior, 2014).

1.3 Data Mining Tasks

Many problems of intellectual, economic, and business interest can be phrased in terms of the following six tasks:

- Classification
- Estimation
- Prediction
- Affinity grouping
- Clustering
- Description and profiling

The first three are all examples of directed data mining, where the goal is to find the value of a particular target variable. Affinity grouping and clustering are undirected tasks where the goal is to uncover structure in data without respect to a particular target variable. Profiling is a descriptive task that may be either directed or undirected (Berry and Linoff, 1997).

1.4 Data Mining Tools

Many good data mining software products are being used, ranging from well-established (and expensive) Enterprise Miner by SAS and Intelligent Miner by IBM, CLEMENTINE by SPSS (a little more accessible by students), PolyAnalyst by Megaputer, and many others in a growing and dynamic industry. WEKA (from the University of Waikato in New Zealand) is an open source tool with many useful developing methods. The Web site for this product (to include free download) is www.cs.waikato.ac.nz/ml/*weka*/. Each product has a well developed Web site. Specialty products cover just about every possible profitable business application. A good source to view current products is www.KDNuggets.com The UCI Machine Learning Repository is a source of very good data mining datasets at http://www.ics.uci.edu/~mlearn/MLOther.html. That site also includes

references of other good data mining sites. Vendors selling data access tools include IBM, SAS Institute Inc., Microsoft, Brio Technology Inc., Oracle, and others. IBM's Intelligent Mining Toolkit has a set of algorithms available for data mining to identify hidden relationships, trends, and patterns. SAS's System for Information Delivery integrates executive information systems, statistical tools for data analysis, and neural network tools (Olson and Delen, 2008).

1.5 Data Mining Process

In order to systematically conduct data mining analysis, a general process is usually followed. There are some standard processes, two of which are described in this chapter. One (CRISP) is an industry standard process consisting of a sequence of steps that are usually involved in a data mining study. The other (SEMMA) is specific to SAS. While each step of either approach isn't needed in every analysis this process provides a good coverage of the steps needed, starting with data exploration, data collection, data processing, analysis, inferences drawn, and implementation (Olson and Delen, 2008).

1.6 Knowledge Discovery in Databases (KDD)

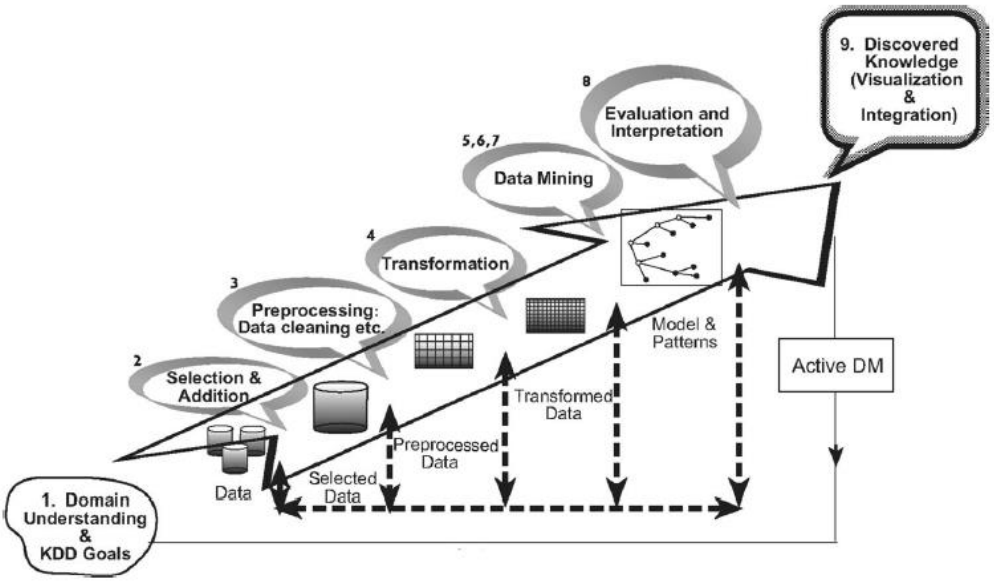


Figure 1: The process of KDD.

KDD process was defined by Fayyad *et al*. (1996) as “the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.” Friedman (1997a) considers the KDD process as an automatic exploratory data analysis of large databases. Hand (1998) views it as a secondary data analysis of large databases. The term “Secondary” emphasizes the fact that the primary purpose of the database was not data analysis. Data Mining can be considered as the central step for the overall process of the KDD process. Because of the centrality of data mining for the KDD process, there are some researchers and practitioners who use the term “data mining” as synonymous with the complete KDD process. Several researchers, such as Brachman and Anand (1994), Fayyad *et al*. (1996) and Reinartz (2002) have proposed different ways of dividing the KDD process into phases. This book adopts a hybridization of these proposals and suggests breaking the KDD process into nine steps as presented in Figure 1. Note that the process is iterative at each step, which means that going back to adjust previous steps may be necessary. The process has many “creative” aspects in the sense that one cannot present one formula or make a complete taxonomy for the right choices for each step and application type. Thus, it is necessary to properly understand the process and the different needs and possibilities in each step. The process starts with determining the goals and “ends” with the implementation of the discovered knowledge. As a result, changes would have to be made in the application domain (such as offering different features to mobile phone users in order to reduce churning). This closes the loop and the effects are then measured on the new data repositories, after which the process is launched again. In what follows is a brief description of the nine-step process, starting with a managerial step:

1.6.1 Developing an understanding of the application domain

This is the initial preparatory step that aims to understand what should be done with the many decisions (about transformation, algorithms, representation, etc.). The people who are in charge of a data mining project need to understand and define the goals of the end-user and the environment in which the knowledge discovery process will take place (including relevant prior

knowledge). As the process proceeds, there may be even revisions and tuning of this step. Having understood the goals, the preprocessing of the data starts as defined in the next three steps (note that some of the methods here are similar to Data Mining algorithms, but these are used in the preprocessing context).

1.6.2 Creating a dataset on which discovery will be performed

Having defined the goals, the data that will be used for the knowledge discovery should be determined. This step includes finding out what data is available, obtaining additional necessary data and then integrating all the data for the knowledge discovery into one dataset, including the attributes that will be considered for the process. This process is very important because the Data Mining learns and discovers new patterns from the available data. This is the evidence base for constructing the models. If some important attributes are missing, then the entire study may fail. For a successful process it is good to consider as many as possible attributes at this stage. However, collecting, organizing and operating complex data repositories is expensive.

1.6.3 Preprocessing and cleansing

At this stage, data reliability is enhanced. It includes data clearing, such as handling missing values and removing noise or outliers. It may involve complex statistical methods, or using specific Data Mining algorithm in this context. For example, if one suspects that a certain attribute is not reliable enough or has too much missing data, then this attribute could become the goal of a data mining supervised algorithm. A prediction model for this attribute will be developed and then, the missing value can be replaced with the predicted value. The extent to which one pays attention to this level depends on many factors. Regardless, studying these aspects is important and is often insightful about enterprise information systems.

1.6.4 Data transformation

At this stage, the generation of better data for the data mining is prepared and developed. One of the methods that can be used here is dimension reduction, such as feature selection and extraction as well as record sampling. Another method that one could use at this stage is attribute transformation, such as discretization of numerical attributes and functional transformation. This step is often crucial for the success of the entire project, but it is usually very project-specific. For example, in medical examinations, it is not the individual aspects/characteristics that make the difference rather, it is the quotient of attributes that often is considered to be the most important factor. In marketing, we may need to consider effects beyond our control as well as efforts and temporal issues such as, studying the effect of advertising accumulation. However, even if we do not use the right transformation at the beginning, we may obtain a surprising effect that hints to us about the transformation needed. Thus, the process reflects upon itself and leads to an understanding of the transformation needed. Having completed the above four steps, the following four steps are related to the Data Mining part where the focus is on the algorithmic aspects employed for each project.

1.6.5 Choosing the appropriate Data Mining task

We are now ready to decide which task of Data Mining would fit best our needs, i.e. classification, regression, or clustering. This mostly depends on the goals and the previous steps. There are two major goals in Data Mining: prediction and description. Prediction is often referred to as supervised Data Mining, while descriptive Data Mining includes the unsupervised classification and visualization aspects of Data Mining. Most data mining techniques are based on inductive learning where a model is constructed explicitly or implicitly by generalizing from a sufficient number of training examples. The underlying assumption of the inductive approach is that the trained model is applicable to future cases. The strategy also takes into account the level of meta-learning for the particular set of available data.

1.6.6 Choosing the Data Mining algorithm

Having mastered the strategy, we are able to decide on the tactics. This stage includes selecting the specific method to be used for searching patterns. For example, in considering precision versus understandability, the former is better with neural networks, while the latter is better with decision trees. Metalearning focuses on explaining wha causes a Data Mining algorithm to be successful or unsuccessful when facing a particular problem. Thus, this approach attempts to understand the conditions under which a Data Mining algorithm is most appropriate.

1.6.7 Employing the Data Mining algorithm

In this step, we might need to employ the algorithm several times until a satisfied result is obtained. In particular, we may have to tüne the algorithm's control parameters such as the minimum number of instances in a single leaf of a decision tree.

1.6.8 Evaluation

In this stage, we evaluate and interpret the extracted patterns (rules, reliability, etc.) with respect to the goals defined in the first step. This step focuses on the comprehensibility and usefulness of the induced model. At this point, we document the discovered knowledge for further usage.

1.6.9 Using the discovered knowledge

We are now ready to incorporate the knowledge into another system for further action. The knowledge becomes active in the sense that we can make changes to the system and measure the effects. In fact, the success of this step determines the effectiveness of the entire process. There are many challenges in this step, such as losing the "laboratory conditions" under which we have been operating. For instance, the knowledge was discovered from a certain static snapshot (usually a sample) of the data, but now the data becomes dynamic. Data structures may change as certain attributes become

unavailable and the data domain may be modified (e.g. an attribute may have a value that has not been assumed before.(Lior, 2014)

1.7 Support Vector Machines

Support vector machines (SVMs) are supervised learning methods that generate input-output mapping functions from a set of labeled training data. The mapping function can be either a classification function (used to categorize the input data) or a regression function (used to estimation of the desired output). For classification, nonlinear kernel functions are often used to transform the input data (inherently representing highly complex nonlinear relationships) to a high dimensional feature space in which the input data becomes more separable (i.e., linearly separable) compared to the original input space. Then, the maximum-margin hyperplanes are constructed to optimally separate the classes in the training data. Two parallel hyperplanes are constructed on each side of the hyperplane that separates the data by maximizing the distance between the two parallel hyperplanes. An assumption is made that the larger the margin or distance between these parallel hyperplanes the better the generalization error of the classifier will be. SVMs belong to a family of generalized linear models which achieves a classification or regression decision based on the value of the linear combination of features. They are also said to belong to “kernel methods”. In addition to its solid mathematical foundation in statistical learning theory, SVMs have demonstrated highly competitive performance in numerous real-world applications, such as medical diagnosis, bioinformatics, face recognition, image processing and text mining, which has established SVMs as one of the most popular, state-of-the-art tools for knowledge discovery and data mining. Similar to artificial neural networks, SVMs possess the well-known ability of being universal approximators of any multivariate function to any desired degree of accuracy. Therefore, they are of particular interest to modeling highly nonlinear, complex systems and processes. Generally, many linear classifiers (hyperplanes) are able to separate data into multiple classes. However, only one hyperplane achieves maximum separation. SVMs classify data as a part of a machine-learning process, which “learns” from the historic cases represented as data points. These data points may have more than two

dimensions. Ultimately we are interested in whether we can separate data by an *n*–1 dimensional hyperplane. This may be seen as a typical form of linear classifier. We are interested in finding if we can achieve maximum separation (margin) between the two (or more) classes. By this we mean that we pick the hyperplane so that the distance from the hyperplane to the nearest data point is maximized. Now, if such a hyperplane exists, the hyperplane is clearly of interest and is known as the maximum-margin hyperplane and such a linear classifier is known as a maximum margin classifier (Olson and Delen, 2008).

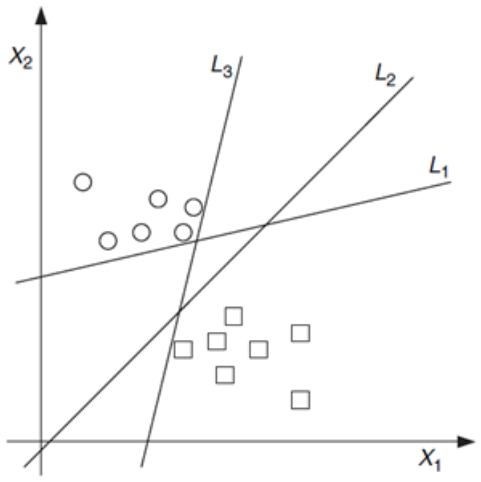


Figure 2: Support Vector Machines

Support vector machines (SVM) is another type of learning system 57, which has many desirable qualities that make it one of most popular algorithms. It not only has a solid theoretical foundation, but also performs classification more accurately than most other algorithms in many applications, especially those applications involving very high dimensional data. For instance, it has been shown by several researchers that SVM is perhaps the most accurate algorithm for text classification. It is also widely used in Web page classification and bioinformatics applications. In general, SVM is a **linear learning system** that builds two-class classifiers. Let the set of training examples *D* be {(**x**1, *y*1), (**x**2, *y*2), …, (**x***n*, *yn*)}, where **x***i* = (*xi*1, *xi*2, …, *xir*) is a *r*-dimensional **input vector** in a real-valued space *X* □□□ *r*,*yi* is its **class label** (output value) and *yi* □□{1, -1}. 1 denotes the positive class and -1 denotes the negative class. We use *y* instead of *c* to represent a class because *y* is commonly used to

represent a class in the SVM literature. Similarly, each data instance is called an **input vector** and denoted by a bold face letter. In the following, we use bold face letters for all vectors. To build a classifier, SVM finds a linear function of the form $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$ so that an input vector $\mathbf{x}_i$ is assigned to the positive class if $f(\mathbf{x}_i) \geq 0$, and to
the negative class otherwise, i.e.,

$$y_i = \begin{cases} 1 & \text{if } \langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \geq 0 \\ -1 & \text{if } \langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b < 0 \end{cases}$$

Hence, $f(\mathbf{x})$ is a real-valued function : $X \to R$. $\mathbf{w} = (w_1, w_2, \ldots, w_r) \in R$ is called the **ias**. $\mathbf{w} \cdot \mathbf{x}$ is the **dot product** of **w** and **x** (or **Euclidean inner product**). Without using vector notation, Equation can be written as: Supervised Learning $f(x1, x2, \ldots, xr) = w1x1+w2x2 + \ldots + wrxr + b$, where xi is the variable representing the ith coordinate of the vector **x**. For convenience, we will use the vector notation from now on. In essence, SVM finds a hyperplane $\mathbf{w} \cdot \mathbf{x} + b = 0$ that separates positive and negative training examples. This hyperplane is called the **decision boundary** or **decision surface**. Geometrically, the hyperplane $\mathbf{w} \cdot \mathbf{x} + b = 0$ divides the input space into two half spaces: one half for positive examples and the other half for negative examples. Recall that a hyperplane is commonly called **a line** in a 2-dimensional space and **a plane** in a 3-dimensional space. Fig. 3(A) shows an example in a 2-dimensional space. Positive instances (also called positive data points or simply positive points) are represented with small filled rectangles, and negative examples are represented with small empty circles. The thick line in the middle is the decision boundary hyperplane (a line in this case), which separates positive (above the line) and negative (below the line) data points. Equation, which is also called the **decision rule** of the SVM classifier, is used to make classification decisions on test instances (Liu, 2007).

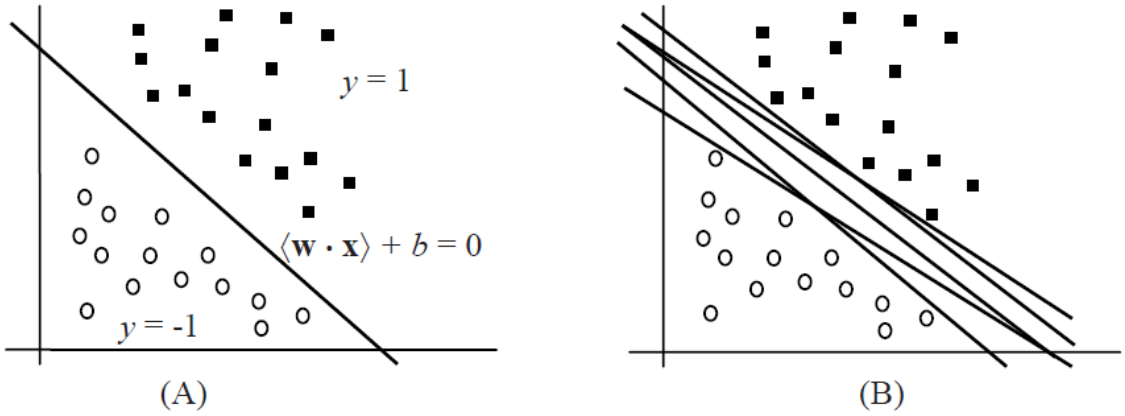


Figure 3: (A) A linearly separable data set and (B) possible decision boundaries (Liu, 2007).

1.8 Regression

A version of a SVM for regression was proposed called support vector regression (SVR). The model produced by support vector classification (as described above) only depends on a subset of the training data, because the cost function for building the model does not care about training points that lie beyond the margin. Analogously, the model produced by SVR only depends on a subset of the training data, because the cost function for building the model ignores any training data that are close (within a threshold)to the model prediction (Olson and Delen, 2008).

1.9 The Multilayer Perceptron

The multilayer perceptron is the most commonly used architecture for predictive data mining. It is a feedforward network, with possibly several hidden layers, one input layer and one output layer, totally interconnected. It can be considered as a highly non-linear generalisation of the linear regression model when the output variables are quantitative, or of the logistic regression model when the output variables are qualitative Giudici(2005) Using a kernel is not the only way to create a nonlinear classifier based on the perceptron. In fact, kernel functions are a recent development in machine learning. Previously, neural network proponents used a different approach for nonlinear classification: they connected many simple perceptron-like models in a

hierarchical structure. This can represent nonlinear decision boundaries. We mentioned there that it is sometimes described as an artificial “neuron.”

Of course, human and animal brains successfully undertake very complex classification tasks—for example, image recognition. The functionality of each individual neuron in a brain is certainly not sufficient to perform these feats. How can they be solved by brain-like structures? The answer lies in the fact that the neurons in the brain are massively interconnected, allowing a problem to be decomposed into subproblems that can be solved at the neuron level. This observation inspired the development of networks of artificial neurons—neural nets. Consider the simple datasets in Figure 5. Figure 5(a) shows a two dimensional instance space with four instances that have classes 0 and 1, represented by white and black dots, respectively. No matter how you draw a straight line through this space, you will not be able to find one that separates all the black points from all the white ones. In other words, the problem is not linearly separable, and the simple perceptron algorithm will fail to generate a separating hyperplane (in this two-dimensional instance space a hyperplane is just a straight line). The situation is different in Figure 5(b) and Figure 5(c): both these problems are linearly separable. The same holds for Figure 5(d), which shows two points in a one-dimensional instance space (in the case of one dimension the separating hyperplane degenerates to a separating point). If you are familiar with propositional logic, you may have noticed that the four situations in Figure 5 correspond to four types of logical connectives. Figure 5(a) represents a logical XOR, where the class is 1 if and only if exactly one of the attributes has value 1. Figure 5(b) represents logical AND, where the class is 1 if and only if both attributes have value 1. Figure 5(c) represents OR, where the class is 0 only if both attributes have value 0. Figure 5(d) represents NOT, where the class is 0 if and only if the attribute has value 1. Because the last three are linearly separable, a perceptron can represent AND, OR, and NOT. Indeed, perceptrons for the corresponding datasets are shown in Figure 5(f) through (h) respectively. However, a simple perceptron cannot represent XOR, because that is not linearly separable. To build a classifier for this type of problem a single perceptron is not sufficient: we need several of them (Witten et al., 2016).

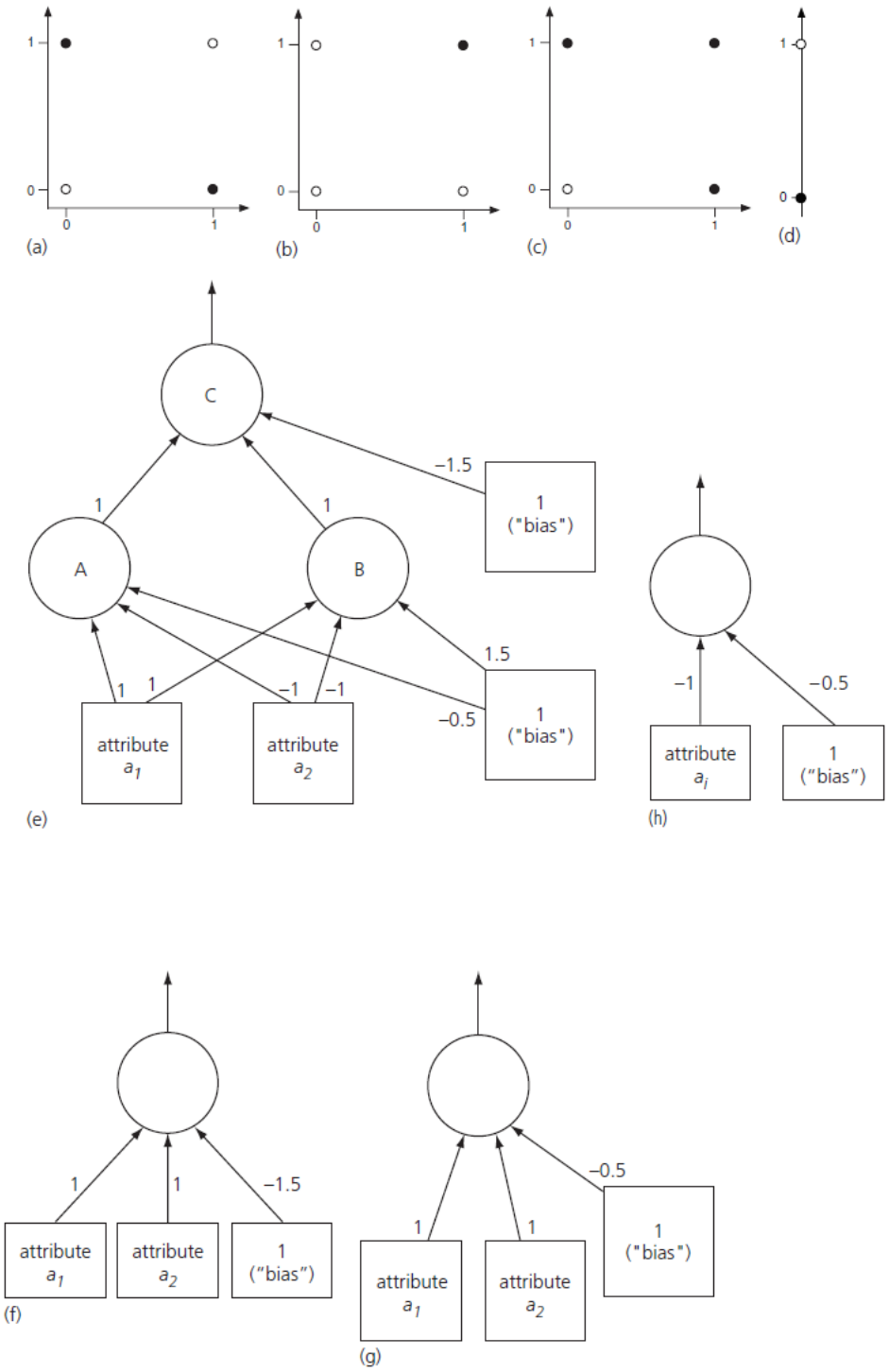


Figure 4: Example datasets and corresponding perceptrons.

Figure 5(e) shows a network with three perceptrons, or *units,* labeled A, B, and C. The first two are connected to what is sometimes called the *input layer* of the network, representing the attributes in the data. As in a simple percep tron, the input layer has an additional constant input called the *bias.* However, the third unit does not have any connections to the input layer. Its input consists of

the output of units A and B (either 0 or 1) and another constant bias unit. These three units make up the *hidden layer* of the multilayer perceptron. They are called “hidden”because the units have no direct connection to the environment. This layer is what enables the system to represent XOR. You can verify this by trying all four possible combinations of input signals. For example, if attribute *a*1 has value 1 and *a*2 has value 1, then unit A will output 1 (because 1 x 1 + 1 x 1 - 0.5 x 1 > 0), unit B will output 0 (because -1 x 1 + -1 x 1 + 1.5 x 1 < 0), and unit C will output 0 (because 1 x 1 + 1 x 0 + -1.5 x 1 < 0). This is the correct answer. Closer inspection of the behavior of the three units reveals that the first one represents OR, the second represents NAND (NOT combined with AND), and the third represents AND. Together they represent the expression (*a*1 OR *a*2) AND (*a*1 NAND *a*3), which is precisely the definition of XOR. As this example illustrates, any expression from propositional calculus can be converted into a multilayer perceptron, because the three connectives AND, OR, and NOT are sufficient for this and we have seen how each can be represented using a perceptron. Individual units can be connected together to form arbitrarily complex expressions. Hence, a multilayer perceptron has the same expressive power as, say, a decision tree. In fact, it turns out that a two-layer perceptron (not counting the input layer) is sufficient. In this case, each unit in the hidden layer corresponds to a variant of AND—a variant because we assume that it may negate some of the inputs before forming the conjunction—joined by an OR that is represented by a single unit in the output layer. In other words, each node in the hidden layer has the same role as a leaf in a decision tree or a single rule in a set of decision rules. The big question is how to learn a multilayer perceptron. There are two aspects to the problem: learning the structure of the network and learning the connection weights. It turns out that there is a relatively simple algorithm for determining the weights given a fixed network structure. This algorithm is called *backpropagation*. However, although there are many algorithms that attempt to identify network structure, this aspect of the problem is commonly solved through experimentation—perhaps combined with a healthy dose of expert knowledge. Sometimes the network can be separated into distinct modules that represent identifiable subtasks (e.g., recognizing different components of an object in an image recognition problem), which opens up a way of incorporating domain

knowledge into the learning process. Often a single hidden layer is all that is necessary, and an appropriate number of units for that layer is determined by maximizing the estimated accuracy (Witten et al.2016).

1.9.1 Supervised learning

Assume that each observation is described by a pair of vectors *(***x***i , ***t***i)* representing the explanatory and response variables, respectively. Let *D* = {*(***x**1, **t**1*) , . . . , (***x***n, ***t***n)*} represent the set of all available observations. The problem is to determine a neural network ***y****i* = *f (***x***i)*, *i* = 1*, . . . , n*, such that the sum of the distances *d(***y***i , ***t***i)* is minimum. Notice the analogy with linear regression models.

1.9.2 Unsupervised learning

Each observation is described by only one vector, with all available variables, *D* = {*(***x**1*) , . . . , (***x***n)*}. The problem is the partitioning of the set *D* into subsets such that the vectors ***x****i* belonging to the same subset are 'close' in comparison to a fixed measure of distance. This is basically a classification problem. We now examine the multilayer perceptron, an example of a supervised network, and the Kohonen network, an example of an unsupervised network (Giudici, 2005).

1.10 Knowledge Discovery From Data

1.10.1 Data cleaning (to remove noise and inconsistent data)

1.10.2 Data integration (where multiple data sources may be combined)

1.10.3 Data selection (where data relevant to the analysis task are retrieved from the database)

1.10.4 Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)

1.10.5 Data mining (an essential process where intelligent methods are applied to extract data patterns)

1.10.6 Pattern evaluation (to identify the truly interesting patterns representing knowledge based on

1.10.7 Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users) (Han et al.2011).

1.11 Database Data

A database system, also called a **database management system** (**DBMS**), consists of a collection of interrelated data, known as a **database**, and a set of software programs to manage and access the data. The software programs provide mechanisms for defining database structures and data storage; for specifying and managing concurrent, shared, or distributed data access; and for ensuring consistency and security of the information stored despite system crashes or attempts at unauthorized access. A **relational database** is a collection of **tables**, each of which is assigned a unique name. Each table consists of a set of **attributes** (*columns* or *fields*) and usually stores a large set of **tuples** (*records* or *rows*). Each tuple in a relational table represents an object identified by a unique *key* and described by a set of attribute values. A semantic data model, such as an **entity-relationship (ER)** data model, is often constructed for relational databases. An ER data model represents the database as a set of entities and their relationships (Han et al.2011).

1.12 Classification and Regression for Predictive Analysis

Classification is the process of finding a **model** (or function) that describes and distinguishes data classes or concepts. The model are derived based on the analysis of a set of **training data** (i.e., data objects for which the class labels are known). The model is used to predict the class label of objects for which the the class label is unknown. *"How is the derived model presented?"* The derived model may be represented in various forms, such as *classification*

rules (i.e., *IF-THEN rules*), *decision trees*, *mathematical formulae*, or *neural networks* A**decisiontree** is a flowchart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions.Decision trees can easily be converted to classification rules. A **neural network**, when used for classification, is typically a collection of neuron-like processing units with weighted connections between the units. There are many other methods for constructing classificationmodels, such as naive Bayesian classification, support vector machines, and *k*-nearest-neighbor classification.Whereas classification predicts categorical (discrete, unordered) labels, **regression** models continuous-valued functions. That is, regression is used to predict missing or unavailable *numerical data values* rather than (discrete) class labels. The term *prediction* refers to both numeric prediction and class label prediction. **Regression analysis** is a statistical methodology that is most often used for numeric prediction, although other methods exist as well. Regression also encompasses the identification of distribution *trends* based on the available data Classification and regression may need to be preceded by **relevance analysis**, which attempts to identify attributes that are significantly relevant to the classification and regression process. Such attributes will be selected for the classification and regression process. Other attributes, which are irrelevant, can then be excluded from consideration (Han et al.2011).

Classification consists of examining the features of a newly presented object and assigning it to one of a predefined set of classes. The objects to be classified are generally represented by records in a database table or a file, and the act of classification consists of adding a new column with a class code of some kind. Examples of classification tasks that have been addressed using the techniques described in this book include:

- Classifying credit applicants as low, medium, or high risk
- Choosing content to be displayed on a Web page
- Determining which phone numbers correspond to fax machines
- Spotting fraudulent insurance claims

- Assigning industry codes and job designations on the basis of free-text job descriptions (Berry and Linoff,1997).

1.13 Estimation

Classification deals with discrete outcomes: yes or no; measles, rubella, or chicken pox. Estimation deals with continuously valued outcomes. Given some input data, estimation comes up with a value for some unknown continuous variable such as income, height, or credit card balance. In practice, estimation is often used to perform a classification task. A credit card company wishing to sell advertising space in its billing envelopes to a ski boot manufacturer might build a classification model that put all of its cardholders into one of two classes, skier or nonskier. Another approach is to build a model that assigns each cardholder a "propensity to ski score." This might be a value from 0 to 1 indicating the estimated probability that the cardholder is a skier. The classification task now comes down to establishing a threshold score. Anyone with a score greater than or equal to the threshold is classed as a skier, and anyone with a lower score is considered not to be a skier. imagine that the ski boot company has budgeted for a mailing of 500,000 pieces. If the classification approach is used and 1.5 million skiers are identified, then it might simply place the ad in the bills of 500,000 people selected at random from that pool. If, on the other hand, each cardholder has a propensity to ski score, it can send the ad to the 500,000 most likely candidates. Examples of estimation tasks include:

- Estimating the number of children in a family
- Estimating a family's total household income
- Estimating the lifetime value of a customer
- Estimating the probability that someone will respond to a balance transfer solicitation.

Regression models and neural networks are well suited to estimation tasks. Survival analysis is well suited to estimation tasks where the goal is to estimate the time to an event, such as a customer stopping (Berry and Linoff, 1997).

1.14 Prediction

Prediction is the same as classification or estimation, except that the records are classified according to some predicted future behavior or estimated future value. In a prediction task, the only way to check the accuracy of the classification is to wait and see. The primary reason for treating prediction as a separate task from classification and estimation is that in predictive modeling there are additional issues regarding the temporal relationship of the input variables or predictors to the target variable. Any of the techniques used for classification and estimation can be adapted for use in prediction by using training examples where the value of the variable to be predicted is already known, along with historical data for those examples. The historical data is used to build a model that explains the current observed behavior. When this model is applied to current inputs, the result is a prediction of future behavior. Examples of prediction tasks addressed by the data mining techniques discussed in this book include:

- Predicting the size of the balance that will be transferred if a credit card prospect accepts a balance transfer offer
- Predicting which customers will leave within the next 6 months
- Predicting which telephone subscribers will order a value-added service such as three-way calling or voice mail

Most of the data mining techniques discussed in this book are suitable for use in prediction so long as training data is available in the proper form. The choice of technique depends on the nature of the input data, the type of value to be predicted, and the importance attached to explicability of the prediction (Berry and Linoff, 1997).

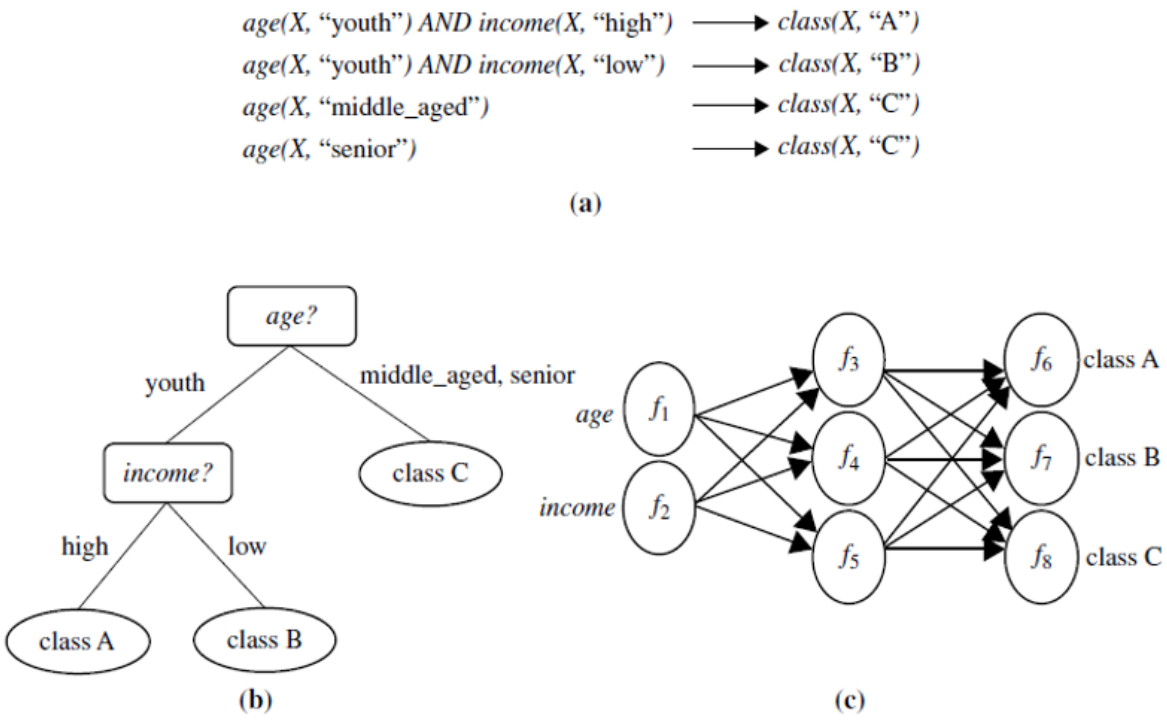


Figure 5: *Data mining: concepts and techniques*. Elsevier . (Han, J., Pei, J., & Kamber, M. 2011).

1.15 Database Systems and Data Warehouses

Database systems research focuses on the creation, maintenance, and use of databases for organizations and end users. Particularly, database systems researchers have established highly recognized principles in data models, query languages, query processingand optimization methods, data storage, and indexing and accessing methods. Database systems are often well known for their high scalability in processing very large, relatively structured data sets. Many data mining tasks need to handle large data sets or even real-time, fast streaming data. Therefore, data mining can make good use of scalable database technologies to achieve high efficiency and scalability on large data sets. Moreover, data mining tasks can be used to extend the capability of existing database systems to satisfy advanced users' sophisticated data analysis requirements. Recent database systems have built systematic data analysis capabilities on database data using data warehousing and data mining facilities. A **data warehouse** integrates data originating from multiple

sources and various timeframes. It consolidates data in multidimensional space to form partially materialized data cubes. The data cube model not only facilitates OLAP in multidimensional databases but also promotes (Han et al., 2011).

1.16 OLAP

Online analytical processing (OLAP) is a category of software technology that enables users to gain insight into data through fast and interactive access to various views of information transformed from raw data, to reflect the real dimensionality of the problem. OLAP is characterized by the function of dynamic dimensional analysis of aggregate enterprise data. A data dimension represents a special perspective of the data along with the data processed, such as time (sales for the last year, by week, month, or quarter), geography (regions or offices), and customers (target market, type, or size). The dimensions are typically hierarchical. Thus, preaggregation can be done logically according to the hierarchies. Preaggregation also allows for a logical drill-down from a large group to a small group. Another way to reduce the data cells in a multidimensional data model is to handle sparse data efficiently. These key techniques of hierarchical dimensions, preaggregation and sparse data management, can reduce the data size and the need for calculation, thus providing quick and direct answers for queries (Ye, 2003).

1.17 Data Warehousing

Data warehouses collect and coalesce data from across an enterprise, often from multiple transaction-processing systems, each with its own database. Analytical systems can Access data warehouses. Data warehousing may be seen as a facilitating technology of data mining. It is not always necessary, as most data mining does not access a data warehouse, but firms that decide to invest in data warehouses often can apply data mining more broadly and more deeply in the organization. For example, if a data warehouse integrates records from sales and billing as well as from human resources, it can be used to find characteristic patterns of effective salespeople (Provost and Fawcett, 2013). A data warehouse is a decision support system; a structured

environment designed to store and analyze all or significant parts of a set of data. The data are logically and physically transformed from multiple source applications into business structure and are updated and maintained over a long time period. The data warehouse is organized around the major subjects in an enterprise such as customer, vendor, product, and activity. The alignment around subject areas affects the design and implementation of data in the data warehouse. Data that will not be used for decision support system processing is excluded from the data warehouse. The architecture of a typical data warehouse and its environment is shown in Fig.7. The main components of a data warehouse include a database; data loading and extracting tools; administration and management platforms; and application tools such as data mining, query, and reporting, visualization and display, and so on (Berson and Smith, 1997). The majör functions of these components are as follows (Zagelow, 1997):

- Access, transform, clean, and summarize data into data warehouse database.
- Store and manage data.
- Manage metadata.
- Display and analyze data, discover useful patterns, and produce reports (Ye, 2003).

To develop such a data warehouse, three steps—mapping from source documentation to enterprise model to tabular model to dimensional model—Source documentation captures the physical data structure of an operational system. Theenterprise model describes all important data in the enterprise at a high level. The outcomes are usually consistent naming conventions and a list of attribute names and definitions, which are useful in mapping source data to the warehouse. The tabular model supports queries and reports for a specific business function such as finance or customer support. The dimensional model supports numerical analysis and represents data as an array. The values in the array are numeric facts that measure business performance, and the dimensions of the array are parameters of the fact, such as date of sale, region, and product. A well-designed data warehouse makes

the right information available to the right people at the right time, and therefore helps a company compete aggressively and sustain leadership (Ye, 2003).

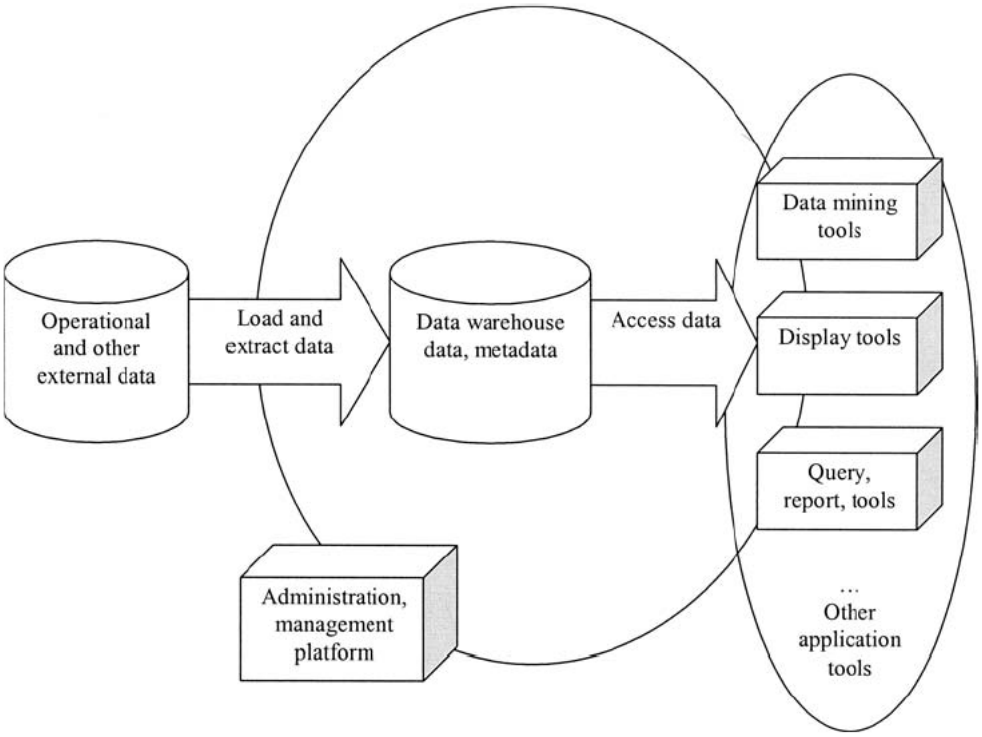


Figure 6: Data warehouse architecture.

1.18 Machine Learning and Data Mining

The collection of methods for extracting (predictive) models from data, now known as machine learning methods, were developed in several fields contemporaneously, most notably Machine Learning, Applied Statistics, and Pattern Recognition. Machine Learning as a field of study arose as a subfield of Artificial Intelligence, which was concerned with methods for improving the knowledge or performance of an intelligent agent over time, in response to the agent's experience in the world. Such improvement often involves analyzing data from the environment and making predictions about unknown quantities, and over the years this data analysis aspect of machine learning has come to play a very large role in the field. As machine learning methods were deployed broadly, the scientific disciplines of Machine Learning, Applied Statistics, and

Pattern Recognition developed close ties, and the separation between the fields has blurred. The field of Data Mining (or KDD: Knowledge Discovery and Data Mining) started as an offshoot of Machine Learning, and they remain closely linked. Both fields are concerned with the analysis of data to find useful or informative patterns. Techniques and algorithms are shared between the two; indeed, the areas are so closely related that researchers commonly participate in both communities and transition between them seamlessly. Nevertheless, it is worth pointing out some of the differences to give perspective. Speaking generally, because Machine Learning is concerned with many types of performance improvement, it includes subfields such as robotics and computer vision that are not part of KDD. It also is concerned with issues of *agency* and *cognition*—how will an intelligent agent use learned knowledge to reason and act in its environment—which are not concerns of Data Mining. Historically, KDD spun off from Machine Learning as a research field focused on concerns raised by examining real-world applications, and a decade and a half later the KDD community remains more concerned with applications than Machine Learning is. As such, research focused on commercial applications and business issues of data analysis tends to gravitate toward the KDD community rather than to Machine Learning. KDD also tends to be more concerned with the entire process of data analytics: data preparation, model learning, evaluation, and so on (Provost and Fawcett, 2013).

1.19 Mining Methodology

Researchers have been vigorously developing new data mining methodologies. This involves the investigation of new kinds of knowledge, mining in multidimensional space, integrating methods fromother disciplines, and the consideration of semantic ties among data objects. In addition, mining methodologies should consider issues such as data uncertainty, noise, and incompleteness. Some mining methods explore how userspecified measures can be used to assess the interestingness of discovered patterns as well as guide the discovery process. Let’s have a look at these various aspects of mining methodology.

1.19.1 Mining various and new kinds of knowledge

Data mining covers a wide spectrum of data analysis and knowledge discovery tasks, fromdata characterization and discrimination to association and correlation analysis, classification, regression, clustering, outlier analysis, sequence analysis, and trend and evolution analysis. These tasks may use the same database in different ways and require the development of numerous data mining techniques. Due to the diversity of applications, new mining tasks continue to emerge, making data mining a dynamic and fast-growing field. For example, for effective knowledge discovery in information networks, integrated clustering and ranking may lead to the discovery of high-quality clusters and object ranks in large networks.

1.19.2 Mining knowledge in multidimensional space

When searching for knowledge in large data sets, we can explore the data in multidimensional space. That is, we can search for interesting patterns among combinations of dimensions (attributes) at varying levels of abstraction. Such mining is known as *(exploratory) multidimensional data mining*. In many cases, data can be aggregated or viewed as a multidimensional data cube. Mining knowledge in cube space can substantially enhance the power and flexibility of data mining.

1.19.3 Data mining—an interdisciplinary effort

The power of data mining can be substantially enhanced by integrating new methods from multiple disciplines. For example, to mine data with natural language text, it makes sense to fuse data mining methods with methods of information retrieval and natural language processing. As another example, consider the mining of software bugs in large programs. This form of mining, known as *bug mining*, benefits from the incorporation of software engineering knowledge into the data mining process.

1.19.4 Boosting the power of discovery in a networked environment

Most data objects reşide in a linked or interconnected environment, whether it be the Web, database relations, files, or documents. Semantic links across multiple data objects can be used to advantage in data mining. Knowledge derived in one set of objects can be used to boost the discovery of knowledge in a “related” or semantically linked set of objects.

1.19.5 Handling uncertainty, noise, or incompleteness of data

Data often contain noise, errors, exceptions, or uncertainty, or are incomplete. Errors and noise may confuse the data mining process, leading to the derivation of erroneous patterns. Data cleaning, data preprocessing, outlier detection and removal, and uncertainty reasoning are examples of techniques that need to be integrated with the data mining process.

1.19.6 Pattern evaluation and pattern- or constraint-guided mining

Not all the patterns generated by data mining processes are interesting. What makes a pattern interesting may vary from user to user. Therefore, techniques are needed to assess the interestingness of discovered patterns based on subjective measures. These estimate the value of patterns with respect to a given user class, based on user beliefs or expectations. Moreover, by using interestingness measures or user-specified constraints to *guide* the discovery process, we may generate more interesting patterns and reduce the search space (Han et al., 2011).

1.20 Classification

Classification is a form of data analysis that extracts models describing important data classes. Such models, called classifiers, predict categorical (discrete, unordered) class labels. For example,we can build a classificationmodel to categorize bank loan applications as either safe or risky. Such analysis can help provide us with a better understanding of the data at large.Many classification methods have been proposed by researchers in machine learning, pattern recognition, and statistics. Most algorithms are

memory resident, typically assuming a small data size. Recent data mining research has built on such work, developing scalable classification and prediction techniques capable of handling large amounts of disk-resident data. Classification has numerous applications, including fraud detection, target marketing, performance prediction, manufacturing, and medical diagnosis. **Regression analysis** is a statistical methodology that is most often used for numeric prediction; hence the two terms tend to be used synonymously, although other methods for numeric prediction exist. Classification and numeric prediction are the two major types of **prediction problems**. This chapter focuses on classification.

Data classification is a two-step process, consisting of a *learning step* (where a classification model is constructed) and a *classification step* (where the model is used to predict class labels for given data) (Han et al., 2011).

1.21 Clustering

Clustering techniques apply when there is no class to be predicted but rather when the instances are to be divided into natural groups. These clusters presumably reflect some mechanism at work in the domain from which instances are drawn, a mechanism that causes some instances to bear a stronger resemblance to each other than they do to the remaining instances. Clustering naturally requires different techniques to the classification and association learning methods (Witten et al., 2016). Clustering is the task of segmenting a heterogeneous population into a number of more homogeneous subgroups or *clusters*. In clustering, there are no predefined classes and no examples. The records are grouped together on the basis of self-similarity. It is up to the user to determine what meaning, if any, to attach to the resulting clusters. Clusters of symptoms might indicate different diseases.

Clusters of customer attributes might indicate different market segments (Berry and Linoff, 1997)**.** *Clustering* refers to the grouping of records, observations, or cases into classes of similar objects. A *cluster* is a collection of records that are similar to one another, and dissimilar to records in other clusters. Clustering differs from classification in that there is no target variable for

clustering. The clustering task does not try to classify, estimate, or predict the value of a target variable. Instead, clustering algorithms seek to segment the entire data set into relatively homogeneous subgroups or clusters, where the similarity of the records within the cluster is maximized and the similarity to records outside the cluster is minimized (Larose, 2014).

1.22 Iterative distance-based clustering

The classic clustering technique is called *k-means.* First, you specify in advance how many clusters are being sought: this is the parameter *k*. Then *k* points are chosen at random as cluster centers. All instances are assigned to their closest cluster center according to the ordinary Euclidean distance metric. Next the centroid, or mean, of the instances in each cluster is calculated—this is the “means” part. These centroids are taken to be new center values for their respective clusters. Finally, the whole process is repeated with the new cluster centers. Iteration continues until the same points are assigned to each cluster in consecutive rounds, at which stage the cluster centers have stabilized and will remain the same forever. This clustering method is simple and effective. It is easy to prove that choosing the cluster center to be the centroid minimizes the total squared distance from each of the cluster’s points to its center. Once the iteration has stabilized, each point is assigned to its nearest cluster center, so the overall effect is to minimize the total squared distance from all points to their cluster centers. But the minimum is a local one; there is no guarantee that it is the global minimum. The final clusters are quite sensitive to the initial cluster centers. Completely different arrangements can arise from small changes in the initial random choice. In fact, this is true of all practical clustering techniques: it is almost always infeasible to find globally optimal clusters.

To increase the chance of finding a global minimum people often run the algorithm several times with different initial choices and choose the best final result—the one with the smallest total squared distance. It is easy to imagine situations in which *k*-means fails to find a good clustering. Consider four instances arranged at the vertices of a rectangle in two dimensional space. There are two natural clusters, formed by grouping together the two vertices

at either end of a short side. But suppose that the two initial cluster centers happen to fall at the midpoints of the *long* sides. This forms a stable configuration. The two clusters each contain the two instances at either end of a long side—no matter how great the difference between the long and the short sides (Witten et al., 2016).

1.23 Faster distance calculations

The *k*-means clustering algorithm usually requires several iterations, each involving finding the distance of *k* cluster centers from every instance to determine its cluster. There are simple approximations that speed this up considerably. For example, you can project the dataset and make cuts along selected axes, instead of using the arbitrary hyperplane divisions that are implied by choosing the nearest cluster center. But this inevitably compromises the quality of the resulting clusters. Here's a better way of speeding things up. Finding the closest cluster center is not so different from finding nearest neighbors in instance-based learning (Witten et al., 2016).

1.24 Support vector regression

The concept of a maximum margin hyperplane only applies to classification. However, support vector machine algorithms have been developed for numeric prediction that share many of the properties encountered in the classification case: they produce a model that can usually be expressed in terms of a few support vectors and can be applied to nonlinear problems using kernel functions. As with regular support vector machines, we will describe the concepts involved but do not attempt to describe the algorithms that actually perform the work. The basic idea is to find a function that approximates the training points well by minimizing the prediction error. The crucial difference is that all deviations up to a user-specified parameter e are simply discarded. Also, when minimizing the error, the risk of overfitting is reduced by simultaneously trying to maximize the flatness of the function. Another difference is that what is minimized is normally the predictions' absolute error instead of the squared error used in linear regression. (There are, however, versions of the algorithm that use the squared error instead.) A user-specified

parameter e defines a tube around the regression function in which errors are ignored: for linear support vector regression, the tube is a cylinder. If all training points can fit within a tube of width 2e, the algorithm outputs the function in the middle of the flattest tube that encloses them. In this case the total perceived error is zero. Figure 8(a) shows a regression problem with one attribute, a numeric class, and eight instances. In this case e was set to 1, so the width of the tube around the regression function (indicated by dotted lines) is 2. Figure 8(b) shows the outcome of the learning process when e is set to 2. As you can see, the wider tube makes it possible to learn a flatter function. The value of e controls how closely the function will fit the training data. Too large a value will produce a meaningless predictor—in the extreme case, when 2e exceeds the range of class values in the training data, the regression line is horizontal and the algorithm just predicts the mean class value. On the other hand, for small values of e there may be no tube that encloses all the data. In that case some training points will have nonzero error, and there will be a trade off between the prediction error and the tube's flatness. In Figure 8(c), e was set to 0.5 and there is no tube of width 1 that encloses all the data. For the linear case, the support vector regression function can be written

$$x = b + \sum_{i \text{ is support vector}} \alpha_i \mathbf{a}(i) \cdot \mathbf{a}.$$

As with classification, the dot product can be replaced by a kernel function for nonlinear problems. The support vectors are all those points that do not fall strictly within the tube—that is, the points outside the tube and on its border. As with classification, all other points have coefficient 0 and can be deleted from the training data without changing the outcome of the learning process. In contrast to the classification case, the a*i* may be negative. We have mentioned that as well as minimizing the error, the algorithm simultaneously tries to maximize the flatness of the regression function. In Figure 8(a) and (b), where there is a tube that encloses all the training data, the algorithm simply outputs the flattest tube that does so. However, in Figure 8(c) there is no tube with error 0, and a tradeoff is struck between the prediction error and the tube's

flatness. This tradeoff is controlled by enforcing an upper limit *C* on the absolute value of the coefficients a*i.* The upper limit restricts the influence of the support vectors on the shape of the regression function and is a parameter that the user must specify in addition to e. The larger *C* is, the more closely the function can fit the data. In the degenerate case e = 0 the algorithm simply performs least-absolute-error regression under the coefficient size constraint, and all training instances become support vectors. Conversely, if e is large enough that the tube can enclose all the data, the error becomes zero, there is no tradeoff to make, and the algorithm outputs the flattest tube that encloses the data irrespective of the value of *C.* (Witten et al., 2016).

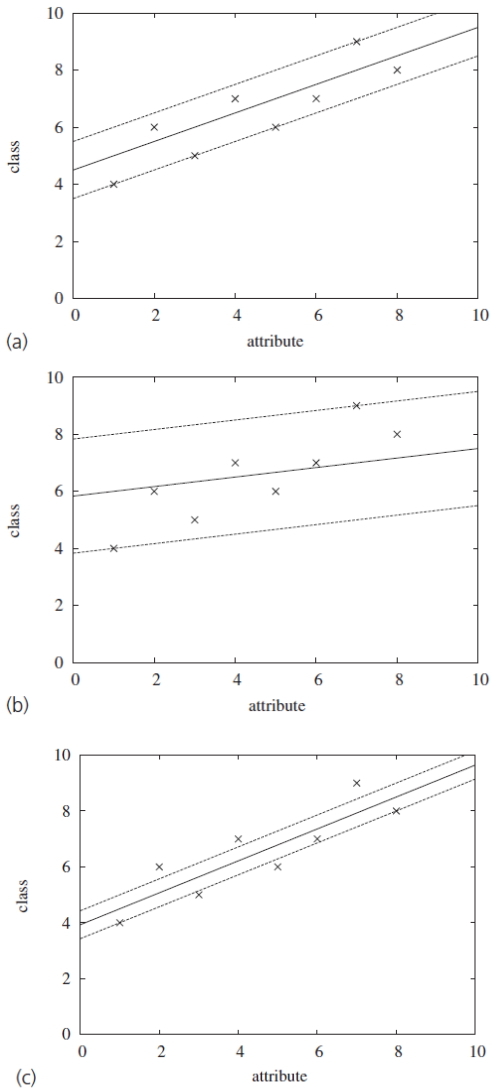


Figure 7: Support vector regression: (a) $\varepsilon = 1$, (b) $\varepsilon = 2$, and (c) $\varepsilon = 0.5$. (Witten et al.,2016)

1.25 Histograms

The most basic descriptive statistic about discrete fields is the number of times different values occur**. Figure 9 shows a *histogram*** of stop reason codes during a period of time. A histogram shows how often each value occurs in the data and can have either absolute quantities (204 times) or percentage (14.6 percent). Often, there are too many values to show in a single histogram such as this case where there are over 30 additional codes grouped into the “other” category. In addition to the values for each category, this histogram also shows the cumulative proportion of stops, whose scale is shown on the left-hand side. Through the cumulative histogram, it is possible to see that the top three codes account for about 50 percent of stops, and the top 10, almost 90 percent. As an aesthetic note, the grid lines intersect both the left- and right-hand scales at sensible points, making it easier to read values off of the chart (Berry and Linoff, 1997).

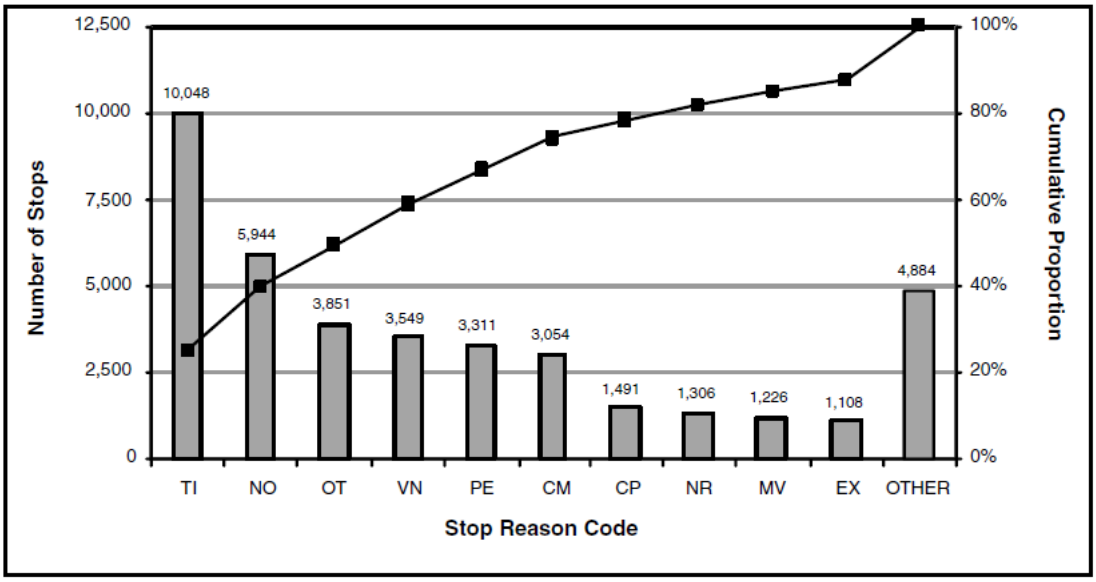


Figure 8: This example shows both a histogram (as a vertical bar chart) and cumulative proportion (as a line) on the same chart for stop reasons associated with a particular marketing effort (Berry and Linoff, 1997).

1.26 Time Series

Time series data contain values that are typically generated by continuous measurement over time. such data typically have implicit dependencies built into the values reveiced over time. (Aggarwal, 2015) Histograms are quite useful and easily made with Excel or any statistics package. However, histograms describe a single moment. Data mining is often concerned with what is happening over time. A key question is whether the frequency of values is constant over time. Time series analysis requires choosing an appropriate time frame for the data; this includes not only the units of time, but also when we start counting from. Some different time frames are the beginning of a customer relationship, when a customer requests a stop, the actual stop date, and so on. Different fields belong in different time frames. For example:

- Fields describing the beginning of a customer relationship—such as original product, original channel, or original market—should be looked at by the customer's original start date.
- Fields describing the end of a customer relationship—such as last product, stop reason, or stop channel—should be looked at by the customer's stop date or the customer's tenure at that point in time.
- Fields describing events during the customer relationship—such as product upgrade or downgrade, response to a promotion, or a late payment—should be looked at by the date of the event, the customer's tenure at that point in time, or the relative time since some other event.

The next step is to plot the time series as shown in Figure 10. This figure has two series for stops by stop date. One shows a particular stop type over time (price increase stops) and the other, the total number of stops. Notice that the units for the time axis are in days. Although much business reporting is done at the weekly and monthly level, we prefer to look at data by day in order to see important patterns that might emerge at a fine level of granularity, patterns that might be obscured by summarization. In this case, there is a clear up and down wiggling pattern in both lines. This is due to a weekly cycle in stops. In addition, the lighter line is for the price increase related stops. These clearly show a marked increase starting in February, due to a change in pricing. A

time series chart has a wealth of information. For example, fitting a line to the data makes it possible to see and quantify long term trends, as shown in Figure 10. Be careful when doing this, because of seasonality. Partial years might introduce inadvertent trends, so include entire years when using a bestfit line. The trend in this figure shows an increase in stops. This may be nothing to worry about, especially since the number of customers is also increasing over this period of time. This suggests that a better measure would be the stop rate, rather than the raw number of stops (Berry and Linoff, 1997).

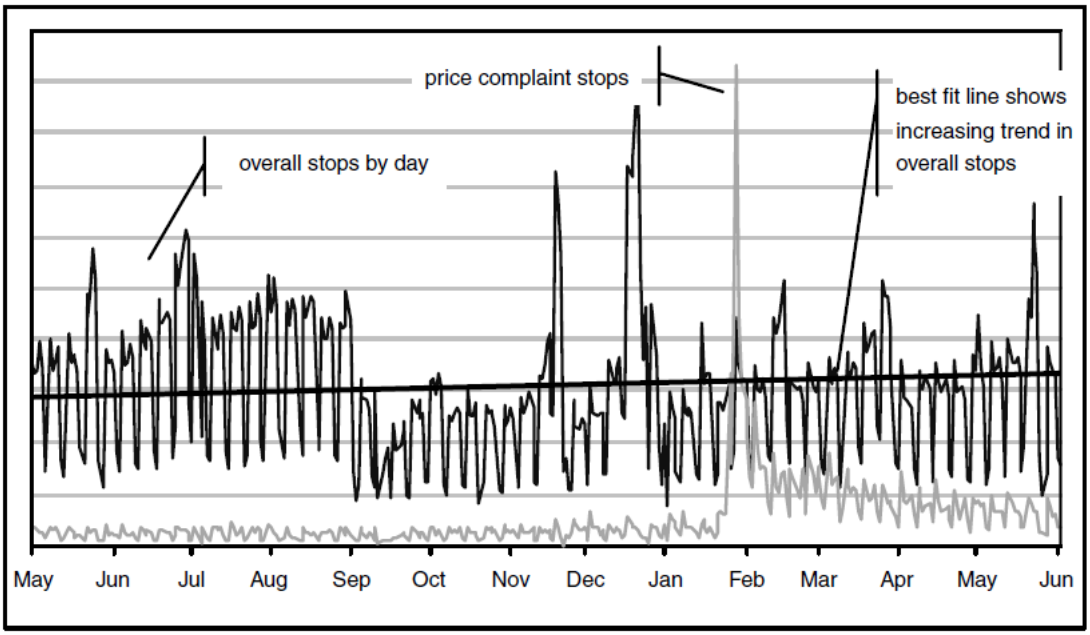


Figure 9. This chart shows two time series plotted with different scales. The dark line is for overall stops; the light line for pricing related stops shows the impact of a change in pricing strategy at the end of January.

1.27 Decision Tree Induction Algorithms

1.27.1 ID3

The ID3 algorithm is considered to be a very simple decision tree algorithm Quinlan (1986). Using information gain as a splitting criterion, the ID3 algorithm ceases to grow when all instances belong to a single value of a target feature or when best information gain is not greater than zero. ID3 does not apply any pruning procedure nor does it handle numeric attributes or missing values. The main advantage of ID3 is its simplicity. Due to this reason, ID3 algorithm is

frequently used for teaching purposes. However, ID3 has several disadvantages:

(1) ID3 does not guarantee an optimal solution, it can get stuck in local optimums because it uses a greedy strategy. To avoid local optimum, backtracking can be used during the search.

(2) ID3 can overfit to the training data. To avoid overfitting, smaller decision trees should be preferred over larger ones. This algorithm usually produces small trees, but it does not always produce the smallest possible tree.

(3) ID3 is designed for nominal attributes. Therefore, continuous data can be used only after converting them to nominal bins. Due to the above drawbacks, most of the practitioners prefer the C4.5 algorithm over ID3 mainly because C4.5 is an evolution of the ID3 algorithm that tries to tackle its drawbacks.

1.27.2 C4.5

C4.5, an evolution of ID3, presented by the same author Quinlan (1993), uses gain ratio as splitting criteria. The splitting ceases when the number of instances to be split is below a certain threshold. Error-based pruning is performed after the growing phase. C4.5 can handle numeric attributes. It can also induce from a training set that incorporates missing values by

using corrected gain ratio criteria as described. C4.5 algorithm provides several improvements to ID3. The most important improvements are:

(1) C4.5 uses a pruning procedure which removes branches that do not contribute to the accuracy and replace them with leaf nodes.

(2) C4.5 allows attribute values to be missing (marked as ?).

(3) C4.5 handles continuous attributes by splitting the attribute’s value range into two subsets (binary split). Specifically, it searches for the best threshold that maximizes the gain ratio criterion. All values above the threshold constitute the first subset and all other values constitute the second subset. C5.0 is an updated, commercial version of C4.5 that offers a number of improvements: it is claimed that C5.0 is much more efficient than C4.5 in terms

of memory and computation time. In certain cases, it provides an impressive speedup from hour and a half (that it took to C4.5 algorithm) to only 3.5 seconds. Moreover, it supports the boosting procedure that can improve predictive performance 9.4.1. J48 is an open source Java implementation of the C4.5 algorithm in the Weka data mining tool. Because J48 algorithm is merely a reimplementation of C4.5, it is expected to perform similarly to C4.5. Nevertheless, a recent comparative study that.

1.27.3 CART

CART stands for Classification and Regression Trees. It was developed by Breiman *et al*. (1984) and is characterized by the fact that it constructs binary trees, namely each internal node has exactly two outgoing edges. The splits are selected using the Twoing Criteria and the obtained tree is pruned by Cost-Complexity Pruning. When provided, CART can consider misclassification costs in the tree induction. It also enables users to provide prior probability distribution. An importa nt feature of CART is its ability to generate regression trees. In regression trees, the leaves predict a real number and not a class. In case of regression, CART looks for splits that minimize the prediction squared error (the least-squared deviation). The prediction in each leaf is based on the weighted mean for node.

1.27.4 CHAID

Starting from the early Seventies, researchers in applied statistics developed procedures for generating decision trees (Kass, 1980). Chi-squared-Automatic-Interaction-Detection (CHIAD) was originally designed to handle nominal attributes only. For each input attribute *ai*, CHAID finds the pair of values in *Vi* that is least significantly different with respect to the target attribute. The significant difference is measured by the *p* value obtained from a statistical test. The statistical test used depends on the type of target attribute. An *F* test is used if the target attribute is continuous; a Pearson chi-squared test if it is nominal; and a likelihood ratio test if it is ordinal. For each selected pair of values, CHAID checks if the *p* value obtained is greater than a certain merge threshold. If the answer is positive, it merges the values and searches for an additional potential pair to be merged. The process is repeated

until no significant pairs are found. The best input attribute to be used for splitting the current node is then selected, such that each child node is made of a group of homogeneous values of the selected attribute. Note that no split is performed if the adjusted *p* value of the best input attribute is not less than a certain spl threshold. This procedure stops also when one of the following conditions is fulfilled:

(1) Maximum tree depth is reached.

(2) Minimum number of cases in a node for being a parent is reached, so it cannot be split any further.

(3) Minimum number of cases in a node for being a child node is reached. CHAID handles missing values by treating them all as a single valid category. CHAID does not perform pruning.

1.27.5 QUEST

The Quick, Unbiased, Efficient Statistical Tree (QUEST) algorithm supports univariate and linear combination splits Loh and Shih (1997). For each split, the association between each input attribute and the target attribute is computed using the ANOVA F-test or Levene's test (for ordinal and continuous attributes) or Pearson's chi-square (for nominal attributes). An ANOVA F-statistic is computed for each attribute. If the largest F-statistic exceeds a predefined threshold value, the attribute with the largest F-value is selected to split the node. Otherwise, Levene's test for unequal variances is computed for every attribute. If the largest Levene's statistic value is greater than a predefined threshold value, the attribute with the largest Levene value is used to split the node. If no attribute exceeded either threshold, the node is split using the attribute with the largest ANOVA F-value. If the target attribute is multinomial, two-means clustering is used to create two super-classes. The attribute that obtains the highest association with the target attribute is selected for splitting. Quadratic Discriminant Analysis (QDA) is applied to find the optimal splitting point for the input attribute. QUEST has negligible bias and yields a binary decision tree. Ten-fold cross-validation is used to prune the trees.

1.28 Reference to Other Algorithms

Table 1 describes other decision tree algorithms available in the literatüre. A profound comparison of the above algorithms and many others has been conducted in (Lim *et al*., 2000), (Lior, 2014).

Table 1:

Additional decision tree inducers (Lior, 2014).

Additional decision tree inducers.

Algorithm	Description	Reference
CAL5	Designed specifically for numerical-valued attributes.	[Muller and Wysotzki (1994)]
FACT	An earlier version of QUEST. Uses statistical tests to select an attribute for splitting each node and then uses discriminant analysis to find the split point.	[Loh and Vanichsetakul (1988)]
LMDT	Constructs a decision tree based on multivariate tests which are linear combinations of the attributes.	[Brodley and Utgoff (1995)]
T1	A one-level decision tree that classifies instances using only one attribute. Missing values are treated as a "special value". Support both continuous an nominal attributes.	[Holte (1993)]
PUBLIC	Integrates the growing and pruning by using MDL cost in order to reduce the computational complexity.	[Rastogi and Shim (2000)]
MARS	A multiple regression function is approximated using linear splines and their tensor products.	[Friedman (1991)]

CHAPTER 2
THEORETICAL FRAMEWORK AND LITERATURE REVIEW

2.1 Theoretical Framework

Increasing competitive conditions and market conditions affect accommodation companies. In cases of uncertainty; Businesses who are unable to plan personnel, inventory management and cost analysis may face the risk of loss and bankruptcy.

Hospitality businesses need to make decisions to create a marketing strategy for the next year, manage stocks, plan personnel management, and conduct growth or contraction analysis. In case of uncertainty, the decisions to be taken will be inaccurate. Incorrect decisions will lead to difficult problems to compensate. Uncertainty causes business functions not to work properly. See the Reaches the maximum cost and minimum profit or loss level. Wrong strategies are determined. It causes conflicts between units. With an effective demand forecast, business functions increase to optimum profit with minimum cost. Strategies are determined correctly, conflicts between units are reduced. The main purpose of accommodation enterprises; to meet the demands and needs of domestic and foreign tourists who are consumers. Demand forecasting; It concerns not only the unit making predictive estimates, but all units of the accommodation establishments. In addition to producing a sales and marketing strategy, stock planning unit in which the forecasting results will be conveyed will have the status of planning forecasting according to the forecasting results. It will be possible for the human resources unit to plan the human power.

It is not possible to adapt to changing market trends without predictive prediction. It will allow accommodation businesses to adapt to changing market trends and keep up with competition. The differences between forecasting and the current situation are indicative of the change in the market. One of the most important aspects for accommodation enterprises is customer satisfaction.

Customer satisfaction can also be possible by meeting expectations. Demand forecasting is important as it will give businesses a preliminary idea to ensure customer satisfaction. For large-scale hospitality businesses, keeping stock is costly. Keeping the right amount of stock; It will reduce inventory costs such as storage and transportation, and capital will be able to prevent unnecessary stock expenses.

When the periods where the demand will decrease and increase will be uncertain, situations that will prevent the planning of financing will arise. Predictive demand forecasts allow for the planning of periods in which demand will decrease and increase. It prevents possible damages. It allows financing planning during periods of rising income, provides an opportunity to evaluate cash and turn to new investments.

Demand forecasting is done by using the existing data of the accommodation establishments of previous years and the data of the variables that affect the demand.

Since tourism demand is one of the main factors that determine the profitability of all businesses and institutions operating in the tourism sector, forecasting the future demand is the most important part of planning studies. Reliable and accurate demand forecasts are essential for effective planning of all activities related to the tourism sector, especially accommodation, transportation and travel businesses (Song and Witt, 2000).

It is not possible to stock up rooms for accommodation that cannot be sold. The demand for accommodation varies depending on economic, social and political factors. The driving force of touristic investments creates demand and investments emerge as a function of the numerical and qualitative

characteristics of the demand. The success of the tourism projects where huge investments are made depends on the forecast of the future demand and the market structure, and therefore the adaptation of the supply sources to the demand (Chu, 2004).

Economic factors have an important position in the tourism sector as in every sector. One should not expect anyone who does not have enough purchasing power to travel to another country for tourism purposes. Tourism is a product that consumers (tourists) have to physically go to the place of production to take advantage of the product. The economic variables in the region where mostly tourists want to go are quite different from those in the regions where tourists come. This is particularly evident in international tourism and long-term travel (Bull, 1995).

Data mining is the science of deducing useful data from huge amount of data. Estimations can be made on data types and data trends by data mining methods. This is especially useful to make conscious decisions on future of the enterprises. Data mining is used to estimate customer demand trends and increase sale performance by existing data on sale details (Hemambika, 2005). Correlation analysis is a statistical method used to test the linear relationship between two variables or relationship of one variable with two or more variables and measure level of this relationship if any. Data mining is generally used for statistical methods (Kadambi, 2005). Models used for data mining are examined in two main titles, estimative and descriptive (Lim, 2013).

Data mining techniques have a wide range by options coming from many disciplines. These selections include techniques such as support vector machines, correlation, linear regression, non-linear regression, genetic algorithms, artificial neural networks, decision trees et cetera. Although size of the database is significant for natural solution of the problem, selecting a data mining technique is essential.(Adriaans and Zantinge 1996, 57) In some cases, data should have higher scales. Following amore complex algorithm running efficiently on a data set of which dimension is reduced, a combined procedure a la scalable algorithms can be predicted (Cloyd, 2002). Time series analysis is one of the most popular approaches used for estimates. Past data

and trends are taken into consideration and used for estimating the future (Al-Alwani, 2014). Time Series are digital quantities in which values of variables can be observed successively. It is not only a source of information used for estimating the future but also a method serving the same purpose. Time series consist of trends, seasonal fluctuations, cyclical fluctuations and irregular movements (error term), (Ibrahim, 1995). Time series data are one of best recognized types of data. Generally, data are formed instantly and measured by real-time global measurements (Rakthanmanon, 2012). Time series estimates are equally spaced data which estimates by historical values formed in time and express these historical values by certain time intervals such as monthly sale data and daily electricity consumptions (Seliem, 2006). There are two different time series analyses which are the most popular ones: Modelling time series: time series forming mechanism to obtain insight and prediction time series: to estimate future values of the time series variable (Yunyue, 2004). Regression methods are probably the oldest and most popular approach for exploring the problem of functional dependency of data (Draper & Smith. 1998; Maimon & Last, 2000). Diversity of regression methods, including the simple one, is related to the number of linear, multilinear and non-linear models. Multiple regression analysis is one of the most popular regression methods. This method includes studies related to linear relationship between univariate and multivariate clusters (Bello, 1995).

Theoretically, increase in the quantity of data ensuring predictive regression estimate also increases score quality of equation estimation. It ensures better estimations. On other hand, missing values create inadequate and weak predictive estimations (Frane, 1976; Sutarso, 1995).

It shall be better to estimate parameters obtained from more observations for regression analysis. In case of a perfect match, estimated value is on par with the real value. However, if this is not the case, a difference between a certain real value and estimated value, in other words deviation shall be observed (Draper and Smith, 1998) requested; (Freund Wilson, 1998); Statsoft.com, 2002**).**

(Siripitayananon, 2002). Data mining shall be used for statistical predictive estimates and studies such as price, sale, cost and demand in hotel and accommodation facility applications (Ding, 2014). In the studies on sale estimates, time series data analyses have been used and positive reflections have been observed in development of monthly sale method for sale and marketing (Gaojun and Boxue, 2009).

Regression models based on sample data sets have been emphasised in previous approaches for sale estimations. However, overly unrealistic results have arisen in this regression model. In most recent theoretical studies on statistics, support vector regression (SVR) method has been offered as a new method to overcome the problem of unrealistic (fabricated) results. In contrast with the traditional regression model, the aim of SVR is to obtain empiric risk with minimum structural risk. In conclusion, it has been seen in the study that SVR is a superior method (Yu and Zhao, 2013). Support vector machine for data regression (SVR) has become a more powerful learning model by better generalization ability (Kong and Lee, 2015). Based on statistical learning theory, SVR has been used for an efficient neural network algorithm and successful sale estimation methods to solve the problem of non-linear regression estimates (Dai and Lu, 2015). Internal factors (seasonality, trend, promotions, price variances etc.) and external factors (cross sale, cannibalization effect, promotions of the competitor etc.) have been taken into consideration in the sale estimation studies performed by time series (García et al., 2012). The results of sale estimates have generally been used to support purchasing and production decision. Long term stability is a significant concern for making a decision on purchasing and production. However, the sale departments generally want to react and update the estimation model more frequently in more efficient markets. There are two significant issues, accuracy of estimation model and currency of the estimation model, while making sale decisions related to the entire market. Long term stability and accuracy/currency are always significant for different decision types. Therefore, these two issues should be dealt in a more comprehensive research (Chern et al., 2015).

Better estimation models have been developed day by day. A new demand estimation system based on estimation model has been developed in more recent studies (to help wholesalers and retailers for issues like logistics distribution). Such data mining studies are significant to improve customer satisfaction (Fang and Weng, 2011). Performances of the estimation model have been compared by using the machine learning techniques known as ANN (Artificial Neural Network), SVM (Support Vector Machine) and LM (Linear Modelling) (Moon et al., 2014). Researches have been carried out on support vector machines in the discipline of medicine (Wu and Shen, 2016). Actual estimative studies have been made by using the support vector regression model, most developed estimation method for data mining (Aggarwal, 2015). Support vector regression has been firstly developed to solve the problem of classification within the frame of statistical learning theory (Burges, 1998), (Cristianini and Shawe-Taylor, 2000). It has been used in estimation and predictive estimation researches due to its terrific estimation qualities (DiPillo et al., 2016). Support vector regression is a useful method for creating an efficient sale estimation structure (Dai et al., 2014). Support vector regression is an artificial intelligence estimation tool based on statistical learning theory and structural risk minimization principle (Lu, 2014).

Support vector regression (SVR) is also used for non-linear regression solution in time series problems and estimations (Zhang et al., 2016). Support vector regression is a significant tool for time series estimations. Researches have been made on accuracy and reliability of estimations and SVR has been tested in terms of accuracy and reliability (Zhao et al., 2013). Support vector machines have yielded positive results for studies aiming to increase analysis verification rate. It has drawn the attention of the scientific community by these results (Shawe-Taylor & Cristianini, 2000), (Gomez et al., 2011). It has been seen that time series analyses and support vector regression method have been used in researches on sale estimations and estimation methods for data mining and, SVR (support vector regression) has had a better performance in multivariate time series analyses. (Khalil et al., 2014), (Cankurt and Subasi, 2015)

Tourism has provided a continuous growth and diversity in the last 60 years despite some obstacles having significant effects on tourist movements in the short term such as wars, regional epidemic and financial crises. Accurate demand estimates form the basis on which decisions on tourism and hotels are made regarding pricing and corporate strategies. At the same time, medium and long term tourism and hotel demand estimates are required for investment decisions of the private actors and public infrastructure investments. Demand modelling and estimation is a significant area for tourism and accommodation researches (Wu and Shen, 2016).

Arrival of tourists at a destination is the traditional and most commonly used indicator of the tourism demand. Other two popular indicators are tourism expenditures (Cortés-Jiménez and Blake, 2011), (Smeral, 2010) and number of overnight stays (Athanasopoulos and Hyndman, 2008), (Baggio and Sainaghi, 2016). Demand for accommodation at hotels is measured by different variables from various perspectives. Some variables such as arrivals of visitors (Guizzardi and Stacchini, 2015), number of overnight stays (Falk, 2014), (Lim et al., 2009) number of rooms sold and occupancy rate (Kouppriouchina and Schwartz, 2014), (Wu et al., 2010) are related to the demand scale (Song et al, 2011). In terms of data set, annual data have been used by many researchers for tourism and hotel demand modelling and estimation studies. These studies normally focus on factors affected by tourism (or hotel) demandand/or long term relationships between medium and long term trend estimates (Guizzardi and Stacchini, 2015), (Song et al., 2011).

Artificial intelligent based models and artificial intelligent techniques have been maintained to be applied on tourism and hotel demand estimations and empiric evidences have displayed satisfying performances. Most of these studies are published in the magazines on other disciplines such as science and statistics calculation. A possible reason is that these studies have mostly focused on methodically development and assessment of estimation accuracy instead of tourism-specific practises. Moreover, establishing a model based on artificial intelligent is deprived of strong theoretical foundations and it is difficult to measure effect of economic factors on tourism and hotel demand by using such models. These explain why artificial intelligence based models have been

limited to tourism and hotel demand analysis and number of publications on artificial intelligence is limited in tourism and accommodation magazines. The artificial neutral network (ANN) is the artificial intelligence based technique most commonly seen in the recent literature. Other techniques such as support vector regression (SVR), rough set model, fuzzy system methods, genetic algorithms and Gauss process regression (GPR) have also been used for tourism and hotel demand estimated albeit in a less frequent way (Wu and Shen, 2016).

Artificial Neural Network which is a nonparametric and data oriented technique has attracted great attention due to its ability of mapping the linear or nonlinear function without an assumption forced by the modelling process. Layers simulating the biological neural system, especially the human brain including input and output have one or more neurons.

These neurons are related to each other in information and data processing process (Cuhadar et al., 2014). Different ANN models have been implemented for tourism and hotel estimation applications such as multilayer perceptron(MLP), radial basis function (RBF), general regression neural network (GRNN) and Elman neural network (Elman NN). MLP is the most commonly usedANN (Artificial Neural Network) model and has more three or more neural layers with non-linear activation function. Chen and Yeh(2012), Claveria and Torra(2014), Lin et al.(2011) SVR is another artificial intelligent based model. It minimizes the learning error by applying the structural risk principle in contrast with the ANN adopting the risk minimization principal. SVR solves the problems of linear regression by mapping input data on a large area non-linearly. Theoretically, SVR, a la YSA model, could yield a global optimum level instead of limited optimisation (Hong et al., 2011). SVR has been used in various studies on tourism and hotel estimations. (Cang, 2014),(Chen and Wang, 2007), (Hong et al., 2011), (Xu et al.2009)

Fuzzy system model is appropriate in cases in which data have been formed in terms of language or consists of less than 50 data points (Tsaur and Kuo, 2011). Different versions of the fuzzy stem model are used for tourism and hotel demand estimations. For example (Aladag et al., 2014), a fuzzy system

model has been used to estimate the international tourism demand in Turkey by seasonal time series (Chen et al.,2010).

It has been observed in the studies with SSCI and SCI indexes that SVR and MLP models have been comparatively studies and these models have been used in the recent studies on tourism accommodation estimations and modelling. However, it has been found out as a conclusion that there is not a method providing clear and precise results in these estimation studies (Cankurt and Subasi, 2015). Studies comparing Linear, MLP, and SVR models are available in the literature.

Linear and non-linear models were compared in studies in the literature. Studies comparing regression analysis in tourism forecasts are available in the literatüre (Cang, 2014), (Cankurt and Subasi, 2015).
In the research on predictive estimation according to the number of domestic and foreign tourists coming to the accommodation establishments; Turkey's most popular tourist areas have been studied comparative three tourist cities.

Turkey's most popular tourist areas in the study, three tourist city of Antalya, Istanbul and Mugla province were selected. In data collected between the years 2007-2018 Antalya, Istanbul and Mugla province, according to data of the accommodation seems to be in the top three in the ranking of Turkey.

Turkey's most popular tourist area between the years 2007-2018, according to data the number of accommodation of the tourist city; Antalya has become Turkey's most popular tourist attractions in the city area.

According to the same data, Turkey's most popular tourist areas of Istanbul has become the second tourist city. The area of Turkey's most popular tourist attractions of the city is Muğla.

Table 2.

Accommodation Data for the Years 2007-2009 regarding Turkey's Touristic Cities with the Highest Number of Incoming Tourists.

	CITIES	*2007*	*CITIES*	*2008*	*CITIES*	*2009*
1	ANTALYA	8.760.026	ANTALYA	7.545.620	ANTALYA	8.840.502
2	İSTANBUL	4.820.073	İSTANBUL	4.409.978	İSTANBUL	4.256.312
3	MUĞLA	2.033.461	MUĞLA	2.154.641	MUĞLA	2.087.705
4	ANKARA	1.516.163	ANKARA	1.451.215	ANKARA	1.494.764
5	İZMİR	1.331.929	İZMİR	1.285.285	İZMİR	1.394.599
6	AYDIN	1.011.720	DENİZLİ	768.466	DENİZLİ	832.934

Table 3.

Accommodation Data for the Years 2010-2012 regarding Turkey's Touristic Cities with the Highest Number of Incoming Tourists.

	CITIES	*2010*	*CITIES*	*2011*	*CITIES*	*2012*
1	ANTALYA	10.952.694	ANTALYA	11.726.601	ANTALYA	12.786.923
2	İSTANBUL	4.641.209	İSTANBUL	5.588.545	İSTANBUL	6.157.578
3	MUĞLA	2.533.635	MUĞLA	2.477.016	MUĞLA	2.489.086
4	ANKARA	1.370.326	İZMİR	1.668.356	İZMİR	1.876.734
5	İZMİR	1.305.486	ANKARA	1.644.528	ANKARA	1.769.454
6	DENİZLİ	893.513	DENİZLİ	1.061.242	AYDIN	1.077.784

Table 4.

Accommodation Data for the Years 2013-2015 regarding Turkey's Touristic Cities with the Highest Number of Incoming Tourists

	CITIES	*2013*	*CITIES*	*2014*	*CITIES*	*2015*
1	ANTALYA	13.794.072	ANTALYA	14.657.471	ANTALYA	14.513.510
2	İSTANBUL	6.314.969	İSTANBUL	7.048.722	İSTANBUL	7.969.371
3	MUĞLA	2.686.304	MUĞLA	3.016.624	MUĞLA	3.411.274
4	İZMİR	1.728.975	İZMİR	1.794.228	İZMİR	2.099.569
5	ANKARA	1.709.556	ANKARA	1.657.617	ANKARA	1.643.621
6	AYDIN	1.135.494	AYDIN	1.170.672	AYDIN	1.264.021

Table 5.

Accommodation Data for the Years 2016-2018 regarding Turkey's Touristic Cities with the Highest Number of Incoming Tourists

	CITIES	*2016*	*CITIES*	*2017*	*CITIES*	*2018*
1	ANTALYA	11.328.410	ANTALYA	13.852.873	ANTALYA	16.615.773
2	İSTANBUL	7.015.399	İSTANBUL	7.823.925	İSTANBUL	9.013.444
3	MUĞLA	2.488.887	MUĞLA	2.083.647	MUĞLA	2.713.132
4	İZMİR	1.899.276	ANKARA	2.046.073	ANKARA	2.446.238
5	ANKARA	1.614.943	İZMİR	1.882.062	İZMİR	2.395.446
6	AYDIN	1.223.751	AYDIN	1.463.182	AYDIN	1.566.340

In the literature research conducted, it was observed that the variables of Dollar Exchange Rate, Occupancy Rates, Incoming Tourist Numbers, Average Stay Periods, Consumer Price Index, The Number of Overnight Stays in the Facility affected the results of demand estimates, the data were collected by years and used in the study.

Table 6.

Data between 2007 and 2018 for Dollar Exchange Rates

Years	Dollar Buying Rate	Dollar Selling Rate
2007	1,4086	1,4154
2008	1,1566	1,1622
2009	1,5293	1,5367
2010	1,481	1,4881
2011	1,5445	1,5519
2012	1,8619	1,8709
2013	1,78	1,7886
2014	2,1687	2,1726
2015	2,3449	2,3491
2016	3,0475	3,0597
2017	3,8149	3,8302
2018	3,8072	3,8224

Dollar Exchange Rates in Turkey by Years between the Years 2007-2018 (http://www.tcmb.gov.tr/)

While the dollar exchange rate data in Turkey were being collected for the years between 2007-2018, the statistics regarding the data of the Central Bank of the Republic of Turkey were utilized.

Table 7.

The Occupancy Rates of the Accommodation Facilities in Antalya, İstanbul and Muğla Provinces by Years for the Years between 2007-2018

	ANTALYA			İSTANBUL			MUĞLA		
	Domestic	Foreign	Total	Domestic	Foreign	Total	Domestic	Foreign	Total
2007	7,52	54,59	62,11	13,24	34,55	47,79	10,05	41,42	51,46
2008	6,29	57,08	63,37	12,41	32,42	44,83	10,51	44,71	55,22
2009	7,27	51,63	58,90	11,73	29,41	41,13	10,01	42,92	52,92
2010	6,94	51,60	58,54	9,88	34,89	44,77	9,89	40,64	50,53
2011	6,91	51,50	58,41	12,34	37,95	50,29	11,25	42,40	53,65
2012	7,12	56,10	63,22	11,58	42,37	53,94	11,02	44,26	55,28
2013	8,00	53,13	61,13	11,24	39,35	50,59	11,64	43,08	54,72
2014	6,46	53,26	59,71	10,58	39,59	50,17	11,03	42,47	53,49
2015	7,69	51,86	59,55	12,60	37,23	49,83	11,96	42,98	54,94
2016	12,38	34,39	46,77	11,19	30,68	41,87	17,53	24,08	41,60
2017	12,28	49,36	61,64	16,76	34,37	51,12	20,15	32,18	52,34
2018	9,82	57,45	67,27	16,51	40,75	57,26	18,57	37,63	56,20

While the data regarding the Occupancy Rates of the Accommodation Facilities in Antalya, İstanbul and Muğla by Years were being collected for the years between 2007-2018, statistics related to the data of the Ministry of Culture and Tourism were utilized.

Table 8.

The Number of Arrivals to Accommodation Facilities by Years in Antalya, İstanbul and Muğla provinces for the Years between 2007-2018.

	ANTALYA			İSTANBUL			MUĞLA		
	Domestic	Foreign	Total	Domestic	Foreign	Total	Domestic	Foreign	Total
2007	1.689.949	7.070.077	8.760.026	1.618.228	3.201.845	4.820.073	640.510	1.392.951	2.033.461
2008	1.371.459	6.174.161	7.545.620	1.479.959	2.930.019	4.409.978	643.916	1.510.725	2.154.641
2009	1.871.527	6.968.975	8.840.502	1.452.070	2.804.242	4.256.312	652.093	1.435.612	2.087.705
2010	2.257.463	8.695.231	10.952.694	1.269.257	3.371.952	4.641.209	819.990	1.713.645	2.533.635
2011	2.272.239	9.454.362	11.726.601	1.756.510	3.832.035	5.588.545	804.244	1.672.772	2.477.016
2012	2.603.361	10.183.562	12.786.923	1.740.970	4.416.608	6.157.578	827.611	1.661.475	2.489.086
2013	2.966.418	10.827.654	13.794.072	1.839.131	4.475.838	6.314.969	1.024.851	1.661.453	2.686.304
2014	2.712.991	11.944.480	14.657.471	1.969.773	5.078.949	7.048.722	1.081.520	1.935.104	3.016.624
2015	3.256.199	11.257.311	14.513.510	2.701.091	5.268.280	7.969.371	1.301.686	2.109.588	3.411.274
2016	4.887.490	6.440.920	11.328.410	2.521.797	4.493.602	7.015.399	1.578.391	910.496	2.488.887
2017	3.813.811	10.039.062	13.852.873	3.282.657	4.541.268	7.823.925	1.046.401	1.037.246	2.083.647
2018	3.655.224	12.960.549	16 615 773	3 334 055	5 679 389	9 013 444	1 163 037	1 550 095	2 713 132

While the data regarding the Number of Arrivals to the Accommodation Facilities by Years in Antalya, İstanbul and Muğla provinces for the years between 2007-2018 were being collected, the statistics of the Ministry of Culture and Tourism were utilized.

Table 9.

The Average Stay Periods of Tourists Coming to the Accommodation Facilities in Antalya, İstanbul and Muğla Provinces between the Years 2007-2018.

	ANTALYA			İSTANBUL			MUĞLA		
	Domestic	Foreign	Total	Domestic	Foreign	Total	Domestic	Foreign	Total
2007	2,9	5,0	4,6	1,7	2,3	2,1	2,7	5,2	4,4
2008	2,9	5,8	5,3	1,7	2,3	2,1	2,8	5,1	4,5
2009	2,9	5,5	5,0	1,8	2,3	2,1	2,9	5,6	4,7
2010	2,9	5,7	5,1	1,7	2,3	2,2	2,7	5,2	4,4
2011	3,0	5,4	4,9	1,7	2,4	2,2	2,9	5,3	4,5
2012	2,9	5,9	5,3	1,7	2,5	2,3	2,9	5,8	4,9
2013	2,9	5,3	4,8	1,7	2,5	2,2	2,7	6,2	4,9
2014	2,8	5,3	4,8	1,7	2,5	2,3	2,7	5,8	4,7
2015	2,8	5,5	4,9	1,6	2,5	2,2	2,5	5,5	4,3
2016	3,1	6,5	5,0	1,6	2,5	2,2	3,1	7,3	4,6
2017	2,93	4,47	4,05	1,74	2,58	2,23	2,88	4,64	3,75
2018	2,94	4,86	4,43	1,81	2,63	2,33	2,78	4,22	3,60

While the data regarding the Average Stay Periods of Tourists Coming to the Accommodation Facilities in Antalya, İstanbul and Muğla provinces between the Years 2007-2018 were being collected, the statistics of the data of the Ministry of Culture and Tourism were utilized.

Table 10.

Consumer Price Indexes by Years between 2007-2018

Years	Consumer Price Index
2007	8,39
2008	10,06
2009	6,53
2010	6,40
2011	10,45
2012	6,16
2013	7,40
2014	8,17
2015	8,81
2016	8,53
2017	11,92
2018	20,3

Source: Central Bank of the Republic of Turkey (http://www.tcmb.gov.tr/)

While data regarding the consumer price indices in Turkey according to the years between 2007-2018 were being collected, the statistics regarding the data of the Central Bank of the Republic of Turkey were utilized.

Table 11.

The Number of Stays of the Tourists Coming to the Accommodation Facilities in Antalya, İstanbul and Muğla by Years for the Years between 2007-2018.

	ANTALYA			İSTANBUL			MUĞLA		
	Domestic	Foreign	Total	Domestic	Foreign	Total	Domestic	Foreign	Total
2007	4.873.596	35.354.560	40.228.156	2.765.667	7.218.526	9.984.193	1.748.894	7.210.768	8.959.662
2008	3.949.280	35.857.931	39.807.211	2.547.188	6.652.379	9.199.567	1.825.709	7.765.768	9.591.477
2009	5.408.876	38.418.577	43.827.453	2.592.780	6.500.709	9.093.489	1.873.795	8.036.645	9.910.440
2010	6.626.632	49.264.226	55.890.858	2.219.366	7.839.170	10.058.536	2.185.224	8.981.194	11.166.418
2011	6.836.384	50.978.182	57.814.566	2.960.632	9.102.455	12.063.087	2.349.251	8.856.012	11.205.263
2012	7.664.494	60.373.209	68.037.703	2.989.465	10.940.248	13.929.713	2.415.408	9.698.075	12.113.483
2013	8.685.148	57.691.550	66.376.698	3.134.169	10.971.911	14.106.080	2.786.949	10.311.491	13.098.440
2014	7.606.770	62.739.573	70.346.343	3.347.550	12.530.462	15.878.012	2.918.072	11.238.477	14.156.549
2015	9.108.898	61.418.288	70.527.186	4.438.885	13.117.799	17.556.684	3.210.924	11.535.722	14.746.646
2016	14.978.938	41.621.666	56.600.604	4.102.811	11.253.206	15.356.017	4.852.135	6.666.101	11.518.236
2017	11.174.350	44.922.472	56.096.822	5.719.053	11.729.842	17.448.895	3.010.338	4.807.971	7.818.309
2018	10.760.725	62.928.381	73.689.106	6.051.060	14.932.763	20.983.823	3.231.490	6.546.690	9.778.180

While the data regarding the Number of Stays of Tourists Coming to the Accommodation Facilities by Years in Antalya, İstanbul and Muğla provinces for the years between 2007-2018 were being collected, the statistics of the Ministry of Culture and Tourism were utilized.

Annual time series analyses were used to estimate future periods by regression methods of data sets, time series analyses and historical figures. The data regarding the variables that affect the outcome were collected to be used in the study. Multiple variables associated with the estimation method, dollar rate, occupancy rates, incoming tourist numbers, average length of stay of tourists, number of stays in the facility, consumer price index data were used in the study by being analyzed using time series methods.

Based on the variables that affect the number of stays, an estimation study was made by entering the data of the Dollar Rate, Occupancy Rates of Accommodation Facilities, Number of Incoming Tourist, Average Duration of Stay in the Facilities, Consumer Price Indices for the years between 2007-2014 in WEKA 3.8 software. Using the data for the years 2007-2014, the next year, 2015 was estimated, the result was compared with real values and the percentage error margin regarding the number of tourists who stay at night was determined.

2.2 LITERATURE REVIEW

Tourism provided continuous expansion and diversity for the past 60 years despite various obstacles such as wars, regional epidemics and financial crises, which have important effects on tourist flows in the short term. Accurate demand estimations form the basis of the business decisions regarding tourism and hotel in terms of pricing and business strategies. At the same time, medium and long term tourism and hotel demand forecasts are required for investment decisions of private sector actors and state infrastructure investments. Demand modeling and forecasting is an important area in tourism and accommodation researches. Recent studies published in the fields of modeling and forecasting tourism and hotel demand from 2007 until today, the hotel demand modelling and forecasting studies regarding accommodation estimates in science citation index and social science citation index journals revealing hotel demand modelling and forecasting and future research orientations aimed at determining tourism and developing issues and the methods studied are generally related to hotel revenue management. In addition, estimation models were used in subjects of study such as hotel management, business management, business planning and purchasing. A total of 170 articles were listed with the results obtained using the key words "tourism estimation," "hotel estimation," "tourism modelling" and "hotel model" in the SSCI and SCI databases. The majority of 170 articles focused on tourism demand (145 studies), while some of the other studies (26 studies) were on hotel estimates. Regarding the distribution of these articles, 130 articles were published in tourism and hotel management magazines, the rest in non-

tourism and accommodation magazines in areas such as prediction, economics, statistics and computer science. In terms of the frequency of data used for model prediction, respectively, annual data were used in 37 studies, data of three months were used in 42 studies and monthly data were used in 61 studies. It was also determined that five studies used weekly data and data of six days. Meanwhile, ten studies used mixed frequency data, and ten sections used cross-sectional data that focused on demand analysis without an estimation. It was found that studies focusing on hotel demand are relatively less considering tourism demand. While different market researches are gradually increasing, it seems remarkable that there is a move away from the analysis of total tourism demand. Some studies went beyond neoclassical economic theory and more sophisticated techniques were introduced such as explaining tourism dynamics and hotel demand, additional dynamics such as environmental factors, online behavior of tourists and consumer confidence indicators, nonlinear soft transition regression to this research area, the technique of modelling data with different frequencies and non-parametric singular spectrum analysis (Han, J., Pei, J., & Kamber, M. 2011). Between 1960 and 2002, 420 studies were published on modeling tourism demand and estimation (Witten et al., 2016).

In the period 2000-2006, only three studies were published on hotel demand estimation. 119 more studies were reviewed on the subject published between 2000 and 2007 (Berry and Linoff, 1997). 155 studies on the methodological development of the tourism demand estimation published between 1995 and 2009 were reviewed (Lior, 2014). In the hotels published between 1985 and 2013, 26 research topics on estimation subjects were studied. Relatively few researches focused on hotel modeling and prediction (Provost and Fawcett, 2013). Hotel demand modeling and estimation are generally related to hotel revenue management. In addition, hotel management has been used for business management, business planning, purchasing decisions and stock control (Ye, 2003). Regarding the variables, the measurement of tourism demand and market segmentation is the tourist demand aimed at a specific destination and the amount of tourism goods and services that consumers want to buy for a certain period of time under certain conditions(Liu, 2007).

The arrival of tourists to a destination is the traditional and mostly used measure of tourism demand. Two other popular measures are tourism expenditures (Pyle, 1999), (Berry and Linoff, 1997) and the number of nights s/he stayed (Larose, 2014), (Aggarwal, 2015). These three variables reflect the magnitude of tourism demand from different perspectives and their analysis can directly contribute to policy recommendations for target governments and administrative decisions in private tourism businesses. Instead of focusing on the total tourism demand at one point, some new studies examine the disaggregated demand for a particular market segment or for a particular type of tourism. Subcategories based on arrivals are usually tourists who come for holiday purposes, tourists coming for business purpose and for friend and relative visits. Spending-based sub-categories are categories such as food expenses, travel expenses, shopping expenses, game expenses etc.(Pyle, 1999) They modeled tourism expenditures for four purposes: these are expenditures for holiday, business, friends and relatives. Some researches are focused on sub-categories according to the type of travel such as cruise (Hemambika, 2005) and air travel (Kadambi, 2005), (Lim, 2013) .

Recently, researchers have given more importance to this sector. In this review, a total of 25 studies investigating hotel demand modeling and forecasting were determined. Hotel accommodation demand is measured with various variables from different perspectives. Some variables such as arrivals of the guests (Cloyd, 2002)the number of nights they stayed(Al-Alwani, 2014)**,** (Ye, 2003), the number of rooms sold (Ibrahim, 1995) are related to demand scale, (Rakthanmanon, 2012) and occupancy rates (Provost and Fawcett, 2013)**,** (Seliem, 2006) Some variables measure hotel revenue according to the financial performance perspective such as sales revenue (Yunyue, 2004)current income per room (Siripitayananon, 2002) and current profit per room (Ding, 2014) . Given the high level of data gathering, the macro level hotel demand estimation provides useful information to the lodging industry as a whole, although the contribution of such studies is limited. There is an increasing interest in the demand forecast for individual hotels based on hotel-specific data (Gaojun and Boxue, 2009), (Provost and Fawcett, 2013).

Estimates for each hotel will benefit hotel practitioners with operational policy implementation, such as reservations by more valued customers, price discrimination, over-reservation policies, late cancellations and early departures (Provost and Fawcett, 2013) The choice of factors that affect tourism demand is quite different from the measurement considering different research objectives of different studies. According to the neo-classical economic theory, price and income are two important influence factors of demand for a product. In empirical studies, tourists' income is frequently used in explaining and estimating tourism prices at a destination and tourist demand at places of residence with substitution prices. Tourist income is expected to affect tourism demand positively and is generally measured in gross national product. Others also include the industrial production index (Yu and Zhao, 2013) and gross disposable income (Kong and Lee, 2015) .

Tourism prices in a destination are expected to affect tourism demand negatively. It is usually measured by the relative consumer price index between the target price and its origin and adjusted by exchange rates. The replacement price refers to the tourism price at a residential destination or group of residences, measured by the weighted average of the consumer price index of the replacement destination or the consumer price indices of a group residence. Therefore, the predicted positive coefficient indicates a substitution relationship, while the negative coefficient indicates an additional relationship between the target and the substitution. Other traditional determinants usually include oil price, advertising expenditures, exchange rate, trade volume between origin and destination, population in the market of origin, unemployment rate and transportation costs measured by other social, cultural, geographical and political factors. In addition, the impact of unique events such as dummy variables, the emergence of seasonality and diseases, terrorist attacks and the tourism demand of the Olympics are also used in forecasting studies. When analyzing hotel demand from a macro level perspective (hotel accommodation demand at a destination), the important determinants of hotel demand are similar to the factors that affect tourist demand: tourist / guest income, arrival tourism price, replacement rate, tourism price, exchange rates, transportation cost, one-time events and seasonal

variables. Other important determinants such as room price, unemployment rate, inflation rate, money supply, industrial production increase and stock market return were also examined (Yunyue, 2004), (Dai and Lu, 2015). The above-mentioned economic variables still dominate recent research on econometric modeling and the forecast of tourist and hotel demand. Meanwhile, recent explanatory variables have emerged in recent researches explained and some are particularly strong at explaining tourism and hotel demand trends and changes. These include climate variables and tourist online behavior variables (Han, J., Pei, J., & Kamber, M. 2011). The climate is thought to affect tourism and hotel demand in the long run, as tourists prefer certain climatic conditions. This variable is relatively constant and does not show the high variations required for tourism demand modeling. Climate, tourism and hotel demand are rarely considered in previous studies on modeling and forecasting. Due to the growing concerns about climate change and the growing interest in research on climate issues and their impact on tourism, some new experimental studies have included climate variables in tourism and hotel demand models and have had a significant impact on tourism and hotel demand. Inclusion of temperature alone as a determining climate variable tended to become widespread. However, it is known that temperature alone does not fully represent a target's climate. There are also other climate variables such as relative humidity, temperature waves, frost days, sunlight duration and seasonal changes (García et al., 2012).

Other new explanatory variables; in reality, not all variables can be included in a single model, given the degree of freedom for data availability and research purposes and model prediction. For this reason, researchers try to find the appropriate determinants of tourism and hotel demand and their optimal representatives according to specific research objectives. For example, (Chernin et al., 2015) measured the relative income and determined the significant effect of the variable using the distance between individual income and the average income of a province (Yu and Zhao, 2013). S/he included the leisure index and climate index in the monthly demand estimation and discovered that they had a stronger impact on tourist arrivals compared to economic factors, (Fang and Weng, 2011) and (Moon et al., 2014) revealed

that the visa restriction had a significant negative impact on incoming tourist flows, with the legal use of those coming to South Korea, Japan, China, and Hong Kong. (Tezel and Buyukyildiz, 2016). S/he found that the termination of the Turkish government's visa requirement policy increased tourist entry. Data set; annual data for tourism and hotel demand modeling and forecasting studies were used by many studies. The focus of these studies is normally long-term relationships between tourism (or hotel) demand and the factors affected and / or medium to long-term trend forecasts. Using annual data cannot analyze seasonal variability in the tourism (or hotel) demand model; the disadvantage of this is that such an analysis cannot capture seasonal characteristics or predict seasonal variations of demand. On the other hand, the focus of a seasonal study is the seasonal data, including quarterly and monthly data that should be taken into account during the modeling process (Cloyd, 2002), (Rakthanmanon, 2012).

Methodological development; it is observed that non-causal time series models, causal econometric approaches and artificial intelligence-based methods still dominate the tourism and hotel demand forecasting area. In particular, some advanced models, such as the almost ideal demand system and panel data analysis, have applied more to this field or entered this field and have exhibited their superiority over some traditional methods. In addition, the combination of different techniques continue being the main aspect of methodological development. Artificial intelligence based methods; artificial intelligence techniques continued to be applied to tourism and hotel demand forecasts, and empirical evidences demonstrated their satisfactory performance. Most of these studies are published in journals in other disciplines such as science and statistical computing. One possible reason is that the majority of these studies focus primarily on methodological development and evaluation of forecasting accuracy rather than on practices specific to tourism. In addition, the establishment of an artificial intelligence-based model lacks strong theoretical foundations and it is difficult to measure the impact of economic factors on tourism and hotel demand using such models. These explain the limitation of artificial intelligence-based models with

the analysis of tourism and hotel demand, and the scarcity of publications on artificial intelligence methods in tourism and accommodation magazines.
The most common artificial intelligence-based technique observed in the recent literature is the artificial neural network (ANN) model. Other techniques such as support vector regression (SVR), coarse embankment model, fuzzy system methods, genetic algorithms, and Gauss process regression (GPR) are also used in tourism and hotel demand forecasts, although less than others (Han, J., Pei, J ., & Kamber, M. 2011). Artificial Neural Network, which is a non-parametric and data-oriented technique, has attracted great attention due to its ability to map linear or nonlinear function without any assumptions imposed by the modeling process. Layers that simulate biological nervous systems, especially human brain, including entry, latent and output; Each layer contains one or more neurons. These neurons are interrelated in the information processing and computing process (Hemambika, 2005).

Different ANN models have been applied to tourism and hotel prediction applications such as Multi Layer Perceptron (MLP), radial base function (RBF), generalized regression neural network (GRNN) and Elman neural network (Elman NN). MLP is the most used ANN (Artificial Neural Network) model; it contains three or more layers of neurons with nonlinear activation function (Dipillo et al., 2016),(Dai et al., 2014),(Lu, 2014). Another artificial intelligence-based model is SVR. Unlike ANN, which adopts the empirical risk minimization principle, SVR minimizes the training error by applying the structural risk principle. SVR solves linear regression problems by nonlinearly mapping input data to a high-dimensional area. Theoretically, SVR, like the ANN model, can achieve a global optimum instead of stuck optimization (Zhang et al., 2016). SVR has been applied with various studies in tourism and hotel prediction (Khalil et al., 2014), (Wu and Shen, 2016), (Zhang et al., 2016), (Athanasopoulos and Hyndman, 2008).

The fuzzy system model is suitable when the data is formed in terms of language or at the point of less than 50 data (Gomez et al., 2011). Different versions of the fuzzy system model are used in tourism and hotel demand forecasts. For example, (Cankurt and Subasi, 2015) used fuzzy system model

with the data of seasonal time series in order to estimate the international tourism demand in Turkey (Kouppriouchina and Schwartz, 2014). They applied the adaptive network-based fuzzy inference system model to predict the tourists coming to Taiwan, and the fuzzy time series model, the gray prediction model and Markov effluent modified model were shown to exhibit superior prediction performance. The fuzzy system model is often combined with genetic algorithms, another AI-based technique used to calculate data. The idea of genetic algorithm stems from natural selection and the theory of genetic evolution. A hybrid method based on fuzzy system and genetic algorithms has been used in various studies (Cortés-Jiménez and Blake, 2011), (Smeral, 2010), (Gomez et al., 2011). Genetic algorithms have also been applied to an SVR model (Wu and Shen, 2016), (Zhang et al., 2016), (Lim et al., 2009).

The fuzzy system has incorporated the SVR technique and genetic algorithms into a new model with superior prediction performance over a number of other models. Different advantages of artificial intelligence-based methods have been taken into consideration and researchers have achieved high predictive performance and satisfactory results. Artificial intelligence-based techniques have been used to determine how individual estimates are combined. For example, (Khalil et al., 2014) combined individual time series estimates based on two ANN models and an SVR model, and empirical results showed that the combined predictions based on three AI-based techniques provide satisfactory prediction performance. Regarding combined forecast performance, it is generally accepted that a combination of forecasts obtained from different forecasting techniques can help improve forecasting accuracy. In particular, (Baggio and Sainaghi, 2016) have shown that combination estimates do not exceed the best single estimate but always outperform the worst single estimate. Therefore, switching to combined prediction techniques is less risky (Çuhadar et al., 2014). Combined estimates have generally proved to be better than the best single estimate (Zhao et al., 2013).

The statistical evidences have been provided; but combined estimates, the best single estimate, are higher than the average of the corresponding single estimates regarding their estimation accuracy (Guizzardi and Stacchini, 2015).

Then he combined the estimates obtained from the tourism demand data to capture the information of the time series with different frequencies. The results show that the prediction combination provides better performance than individual models. Given the potential to reduce estimation risks and improve estimation accuracy, in future studies, there should be more discussion and consideration of combination predictions such as selection criteria of individual models for the pool, optimal number of individual models to be combined and innovative combination methods (Han, J., Pei , J., & Kamber, M. 2011). In the studies conducted in SSCI and SCI indexed publications, current studies on tourism accommodation estimation and modeling, it is seen that comparative researches of SVR and MLP models are conducted and that SVR and MLP models are used. However, as a result, it is stated in these estimation studies that there is no method that gives clear and precise results (Falk, 2014).

CHAPTER 3
THE METHODOLOGY OF THE RESEARCH

3.1 Research Method

WEKA 3.8 data mining software was used in the research. Regarding the accommodation, occupancy rates, the number of arrivals at the facility, average stay periods and overnights in Antalya, İstanbul and Muğla, which are the three touristic cities of Turkey with the highest number of incoming tourists, the data were collected for previous years.

The data of the variables that affect the estimation results were collected. Using the data of Dollar Exchange Rates, Facility Occupancy Rates, Arrivals in the Facility, Average Length of Stay in the Facility, Consumer Price Indices and Overnight Numbers in the Facility between 2007 and 2014, regarding the tourists whose arrival in Antalya, İstanbul and Muğla in 2015 was planned, the Overnight Numbers in the Facility were estimated on the basis of individual provinces. Real values were compared with estimated figures for 2015 and an evaluation was made on the basis of error margin in percentage. It was aimed to determine the best regression method while making prediction studies for tourism destination.

With the same method, data for the years between 2007-2015 was collected and an estimation study for 2016 was carried out. Real values were compared with estimated figures for 2016 and an evaluation was made on the basis of error margin in percentage.

Data was collected between 2007-2016 and the estimation study for 2017 was made. Real values were compared with estimated figures for 2017 and an evaluation was made on the basis of error margin in percentage.

Data was collected between 2007-2017 and estimation study was made for 2018. Real values were compared with the estimated figures for 2018 and an evaluation was made on the basis of error margin in percentage.

3.2 Research Model

The research qualifies as a prediction of the future period with time series analysis. Demand forecasts were made with quantitative techniques in the study. It is thought that the study may provide an infrastructure for other studies to be conducted in the literature to investigate the cause-effect relationship.

3.3 Population and Sample

The research was carried out in Antalya, İstanbul and Muğla. The population of the research consists of domestic and foreign tourists visiting Antalya, Istanbul and Muğla. The sample of the study consists of domestic and foreign tourists staying in the accommodation establishments in Antalya, Istanbul and Muğla between 2007-2018.

3.4 Data Set and Multivariate Approach for Tourism Demand Modeling

As used in many studies, annual time series analyses have been used to make future estimations by data set time series analyses and regression methods of past figures. The multiple variables related to estimation method, USD exchange rate, occupancy rates, arriving tourist numbers, average accommodation duration of tourists, number of overnight stays and CPI, have been used and implemented in the study by analysis via time series methods. (Naik and Samant, 2016).

In the study, annual time series data of the variables have been cited from the websites of the following institutions: Republic of Turkey Ministry of Tourism(www.die.gov.tr) Turkish State Institute of Statistics (http://evds.tcmb.gov.tr), Central Bank of Republic of Turkey (https://www.ecb.europa.eu), European Central Bank (ECB) (www.tursab.org.tr) and TURSAB.

3.5 Data and Modelling Approaches

As seen from the studies performed on tourism demand estimation, the main factors affecting the tourism demand estimation reveal the relation between variables (USD exchange rate, occupancy rates, arriving tourist numbers, average accommodation duration of tourists, number of overnight stays, CPI) and demand and determinants explaining the data mining methods.

3.6 Data Collection Tools and Data Analysis

The data regarding dollar exchange rate, occupancy rates, the number of incoming tourists, the average stay periods of tourists, the number of overnights in the facility and consumer price index were needed and the data obtained from the Republic of Turkey Culture and Tourism Ministry, Turkey State Institute of Statistics, Central Bank of the Republic of Turkey, the European Central Bank and Turkey Travel Agencies Union were used as primary sources. In the literature review carried out on the Web of Science, scientific articles on the subject of research and scientific books published by internationally known publishers were used as secondary sources. In the research process, data were collected using appropriate methods and normality test was applied to analyze the collected data.

The analysis of the research data was performed with WEKA 3.8 data mining software. In the application, the Multilayer Perception and Support Vector Regression analysis methods, which are among the regression analysis methods of WEKA 3.8 data mining software were used; regarding Antalya, İstanbul and Muğla, the cities with the highest number of overnight tourists, correlation coefficient, relative absolute error, root relative square error, predicted values, real values were determined and results of regression analysis methods were compared.

3.7 Contribution of the Research to the Literature

When the tourism planners and managers of countries and cities, operators of accommodation facilities are faced with uncertainty about the short and long term demands regarding the future, it is thought that using the right estimation tools will be very helpful for planning and management, may provide the

infrastructure about issues regarding the determination of the regression method in order to obtain the best result in estimation studies about tourism and the other studies that will be conducted in the literature about the issue in cause-effect relation with the purpose of research and may reveal the important points that they should consider and how they can interpret the results in these analyses.

CHAPTER 4
FINDINGS AND DISCUSSIONS

In this section, statistical analysis results of study data are presented.

Graph 1: Comparison of SVR and MLP methods for the province of Antalya.

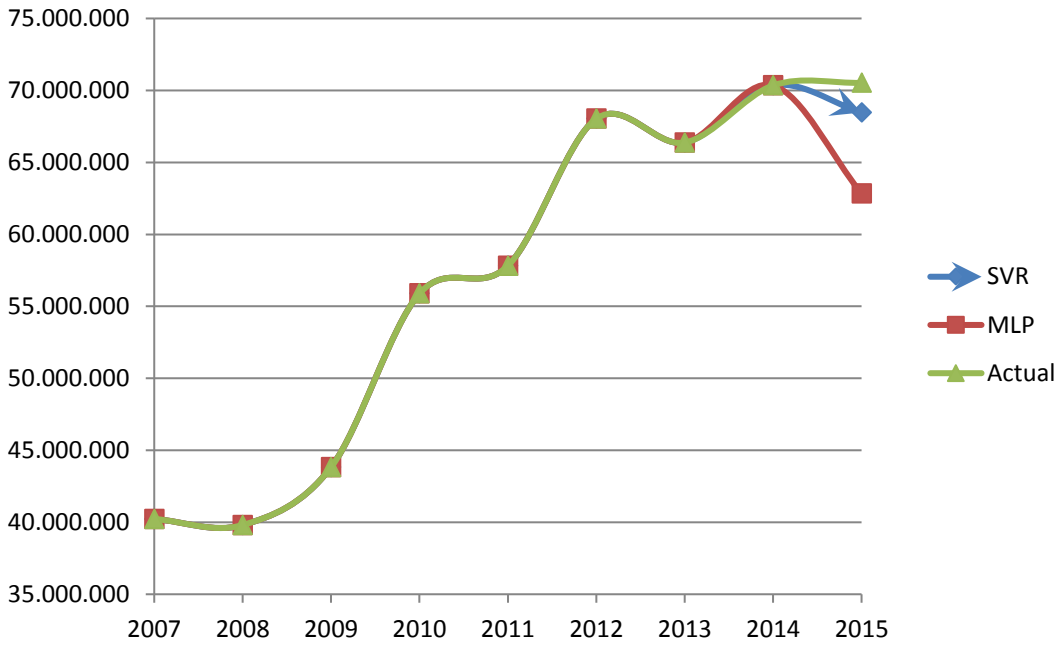


Table 12.

The comparison of estimated and real overnight numbers in Antalya for 2015.

Model	Percentage of Error	Correlation coefficient	Relative Absolute error	Root relative squared error	Forecasting Values	Actual Values
SVR	2,9203 %	0.9887	18.6332 %	18.5094 %	68.467.596	70.527.186
MLP	10,9095 %	1	0.0669 %	0.0783 %	62.833.047	70.527.186

Using the data of the years 2007-2014 for the variables that affect the tourist overnight numbers of the accommodation establishments in Antalya province, the number of tourist overnight stays in 2015 was estimated by the Multilayer Perceptron regression method. Whereas the real figure for 2015 is 70,527,186 the predicted figure was found to be 62,833,047. Percentage of error between real and predicted numbers was calculated as 10,9095 %.

Using the data of the years 2007-2014 for the variables that affect the tourist overnight numbers of the accommodation establishments in Antalya province, the number of tourist overnight stays in 2015 was estimated by the SVR regression method. Whereas the real figure for 2015 was 70,527,186, the predicted figure was found to be 68,467,596. Percentage of error between real and predicted numbers was calculated as 2.9203%.

Graph 2: Comparison of SVR and MLP methods for the province of Istanbul.

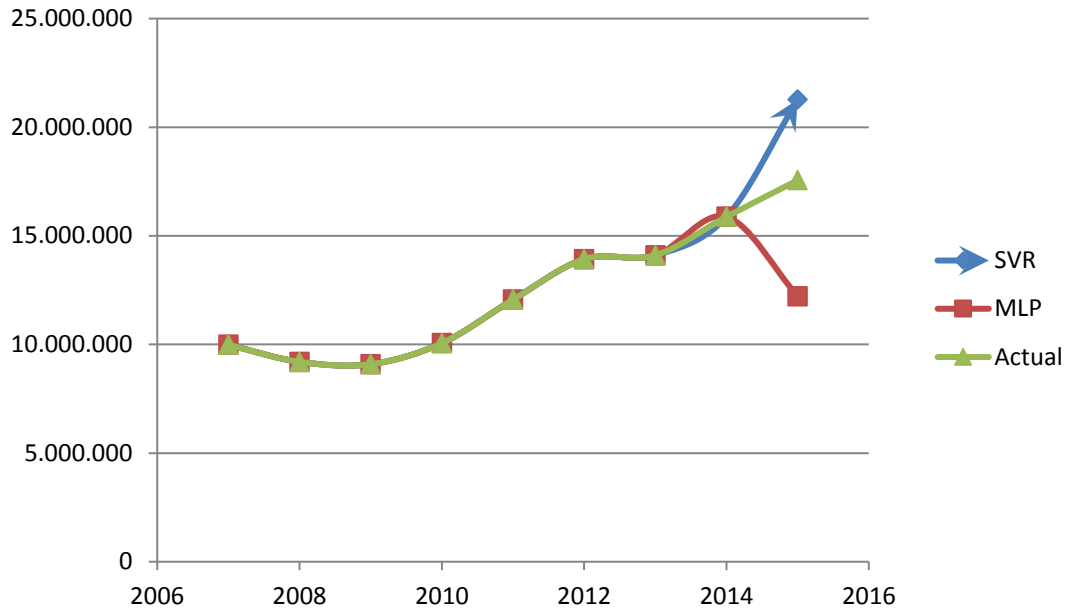


Table 13.

The comparison of estimated and real overnight numbers in Istanbul for 2015.

Model	Percentage of Error	Correlation coefficient	Relative Absolute error	Root relative squared error	Forecasting Values	Actual Values
SVR	-21,1661 %	0.9974	3.3235 %	7.7393 %	21.272.753	17.556.684
MLP	30,3951 %	1	0.183 %	0.2364 %	12.220.308	17.556.684

Using the data of the years 2007-2014 for the variables that affect the tourist overnight numbers of the accommodation establishments in İstanbul province, the number of tourist overnight stays in 2015 was estimated by the Multilayer Perceptron regression method. While the real figure for 2015 was 17,556,684

the predicted figure was found to be 12,220,308. Percentage of error between real and predicted numbers was calculated as 30,3951 %.

Using the data of the years 2007-2014 for the variables that affect the tourist overnight numbers of the accommodation establishments in İstanbul province, the number of tourist overnight stays in 2015 was estimated by the SVR regression method. Whereas the real figure for 2015 was 17,556,684 the predicted figure was found to be 21,272,753. Percentage of error between real and predicted numbers was calculated as -21,1661 %.

Graph 3: Comparison of SVR and MLP methods for the province of Muğla.

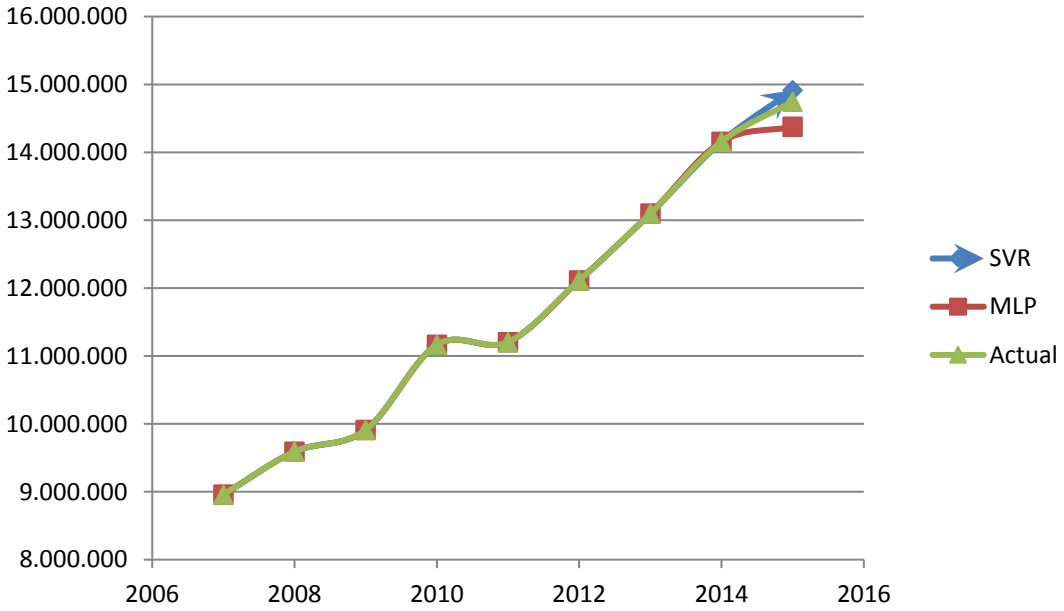


Table 14.

The comparison of estimated and real overnight numbers in Muğla for 2015.

Model	Percentage of Error	Correlation coefficient	Relative Absolute error	Root relative squared error	Forecasting Values	Actual Values
SVR	-1,1217 %	0.9987	3.5979 %	5.2698 %	14.912.065	14.746.646
MLP	2,5153 %	1	0.0113 %	0.0123 %	14.375.728	14.746.646

Using the data of the years 2007-2014 for the variables that affect the tourist overnight numbers of the accommodation establishments in Muğla province, the number of tourist overnight stays in 2015 was estimated by the Multilayer Perceptron regression method. While the real figure for 2015 was 14,746,646 the predicted figure was found to be 14,375,728. Percentage of error between real and predicted numbers was calculated as 2,5153 %.

Using the data of the years 2007-2014 for the variables that affect the tourist overnight numbers of the accommodation establishments in Muğla province, the number of tourist overnight stays in 2015 was estimated by the SVR regression method. While the real figure for 2015 was 14,746,646 the estimated figure was found to be 14,912,065. Percentage of error between real and predicted numbers was calculated as -1,1217 %.

Graph 4: Comparison of SVR and MLP methods for the province of Antalya.

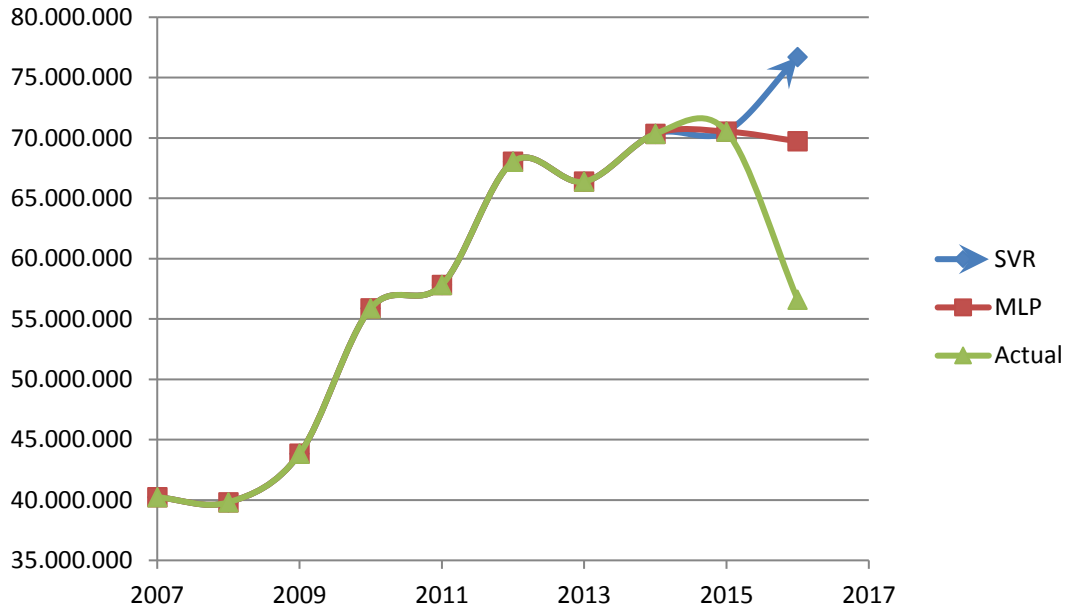


Table 15.

The comparison of estimated and real overnight numbers in Antalya for 2016.

Model	Percentage of Error	Correlation coefficient	Relative Absolute error	Root relative squared error	Forecasting Values	Actual Values
SVR	-35,5111 %	0.9909	9.5074 %	13.651 %	76.700.105	56.600.604
MLP	-23,1885 %	0.9999	2.5867 %	2.7262 %	69.725.459	56.600.604

Using the data of the years 2007-2015 for the variables that affect the tourist overnight numbers of the accommodation establishments in Antalya province, the number of tourist overnight stays in 2016 was estimated by the Multilayer Perceptron regression method. While the real figure for 2016 was 56,600,604 the estimated figure was found to be 69,725,459. Percentage of error between real and predicted numbers was calculated as -23.1885%.

Using the data of the years 2007-2015 for the variables that affect the tourist overnight numbers of the accommodation establishments in Antalya province, the number of tourist overnight stays in 2016 was estimated by the SVR regression method. Whereas the real figure for 2016 was 56,600,604 the estimated figure was found to be 76,700,105. Percentage of error between real and predicted numbers was calculated as -35,5111%.

Graph 5: Comparison of SVR and MLP methods for the province of Istanbul.

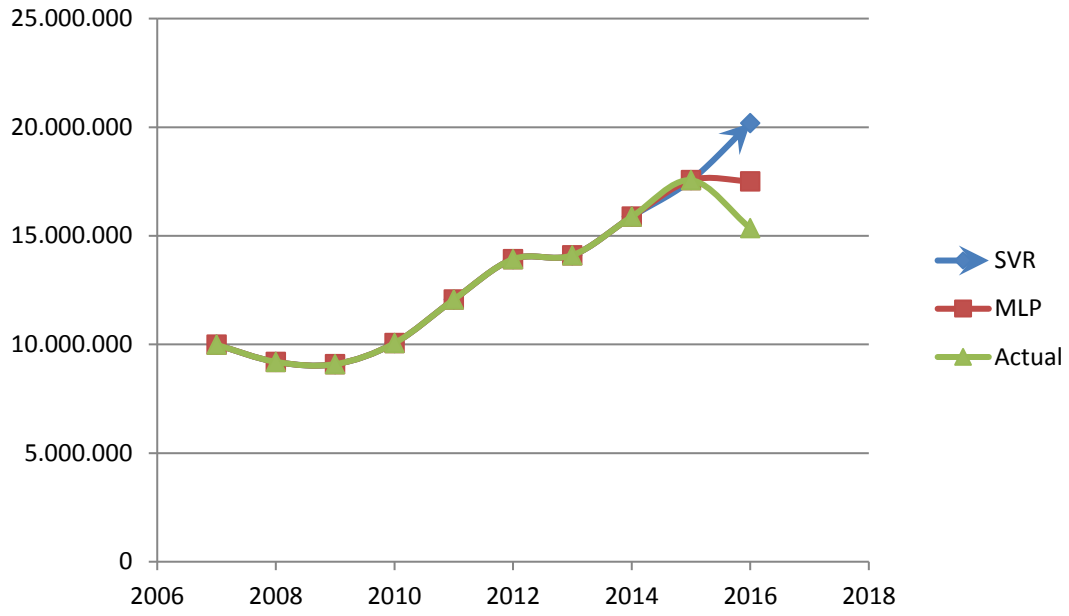


Table 16.

The comparison of estimated and real overnight numbers in Istanbul for 2016.

Model	Percentage of Error	Correlation coefficient	Relative Absolute error	Root relative squared error	Forecasting Values	Actual Values
SVR	-31,4256 %	0.9978	3.0905 %	6.8888 %	20.181.735	15.356.017
MLP	-14,0236 %	0.9999	1.4348 %	1.6703 %	17.509.483	15.356.017

Using the data of the years 2007-2015 for the variables that affect the tourist overnight numbers of the accommodation establishments in İstanbul province, the number of tourist overnight stays in 2016 was estimated by the Multilayer Perceptron regression method. Whereas the real figure for 2016 was 15,356,017 the estimated figure was found to be 17,509,483. Percentage of error between real and predicted numbers was calculated as -14,0236%.

Using the data of the years 2007-2015 for the variables that affect the tourist overnight numbers of the accommodation establishments in İstanbul province, the number of tourist overnight stays in 2016 was estimated by the SVR regression method. While the real figure for 2016 was 15,356,017 the estimated figure was found to be 20,181,735. Percentage of error between real and predicted numbers was calculated as -31,4256%.

Graph 6: Comparison of SVR and MLP methods for the province of Muğla.

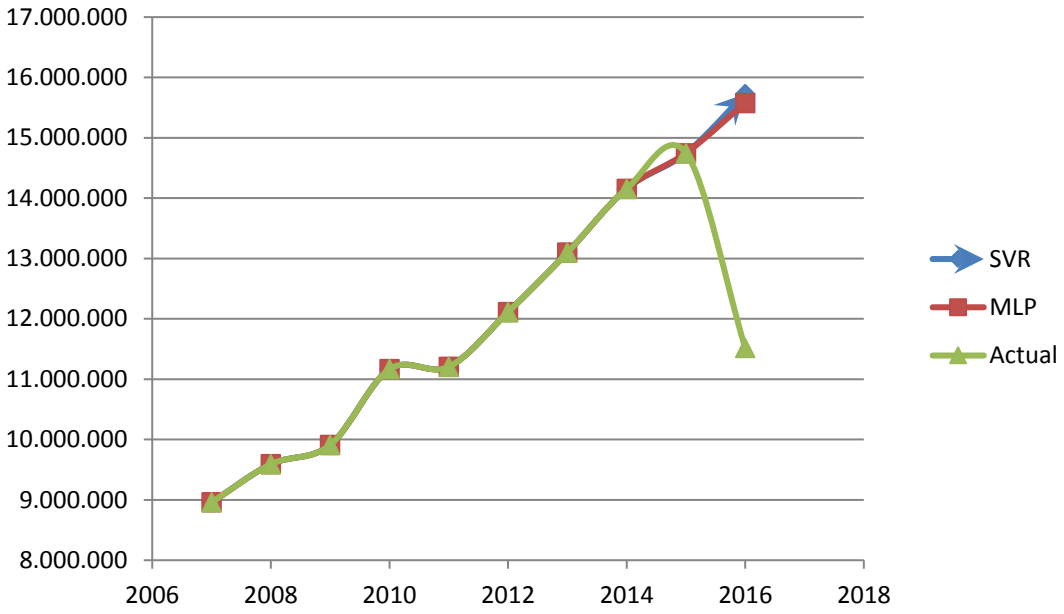


Table 17.

The comparison of estimated and real overnight numbers in Muğla for 2016.

Model	Percentage of Error	Correlation coefficient	Relative Absolute error	Root relative squared error	Forecasting Values	Actual Values
SVR	-36,5426 %	0.9985	3.7947 %	5.8367 %	15.727.301	11.518.236
MLP	-35,2513 %	0.9998	3.0142 %	3.03 %	15.578.567	11.518.236

Using the data of the years 2007-2015 for the variables that affect the tourist overnight numbers of the accommodation establishments in Muğla province, the number of tourist overnight stays in 2016 was estimated by the Multilayer Perceptron regression method. While the real figure for 2016 was 11,518,236

the predicted figure was found to be 15,727,301. The percentage of error between the real and predicted numbers was calculated as -36,5426 %.

Using the data of the years 2007-2015 for the variables that affect the tourist overnight numbers of the accommodation establishments in Muğla province, the number of tourist overnight stays in 2016 was estimated by the SVR regression method. While the real figure for 2016 was 11,518,236 the predicted figure was found to be 15,578,567. Percentage of error between real and predicted numbers was calculated as -35,2513%.

Graph 7: Comparison of SVR and MLP methods for the province of Antalya.

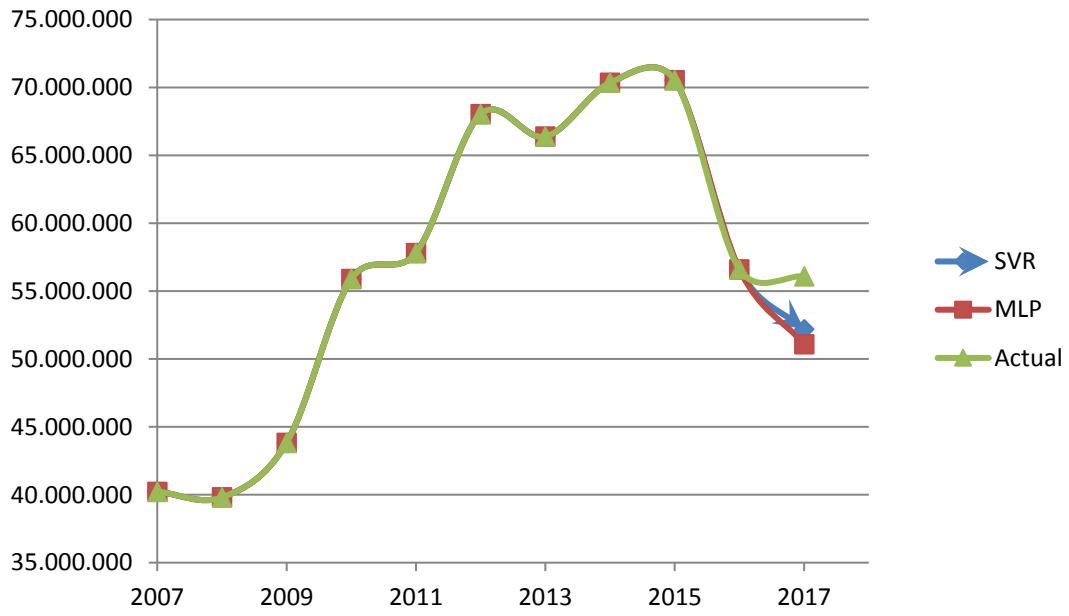


Table 18.

The comparison of estimated and real overnight numbers in Antalya for 2017.

Model	Percentage of Error	Correlation coefficient	Relative Absolute error	Root relative squared error	Forecasting Values	Actual Values
SVR	6,9656 %	0.9931	7.2917 %	11.817 %	52.189.329	56.096.822
MLP	8,9186 %	0.9998	3.0094 %	2.8432 %	51.093.757	56.096.822

Using the data of the years 2007-2016 for the variables that affect the tourist overnight numbers of the accommodation establishments in Antalya province,

the number of tourist overnight stays in 2017 was estimated by the Multilayer Perceptron regression method. While the real figure for 2017 was 56,096,822, the estimated figure was found to be 51,093,757. Percentage of error between real and predicted numbers was calculated as 8.9186%.

Using the data of the years 2007-2016 for the variables that affect the tourist overnight numbers of the accommodation establishments in Antalya province, the number of tourist overnight stays in 2017 was estimated by the SVR regression method. The real figure for 2017 was 56,096,822, while the estimated figure was found to be 52,189,329. The percentage of error between real and predicted numbers was calculated as 6.9656%.

Graph 8: Comparison of SVR and MLP methods for the province of Istanbul.

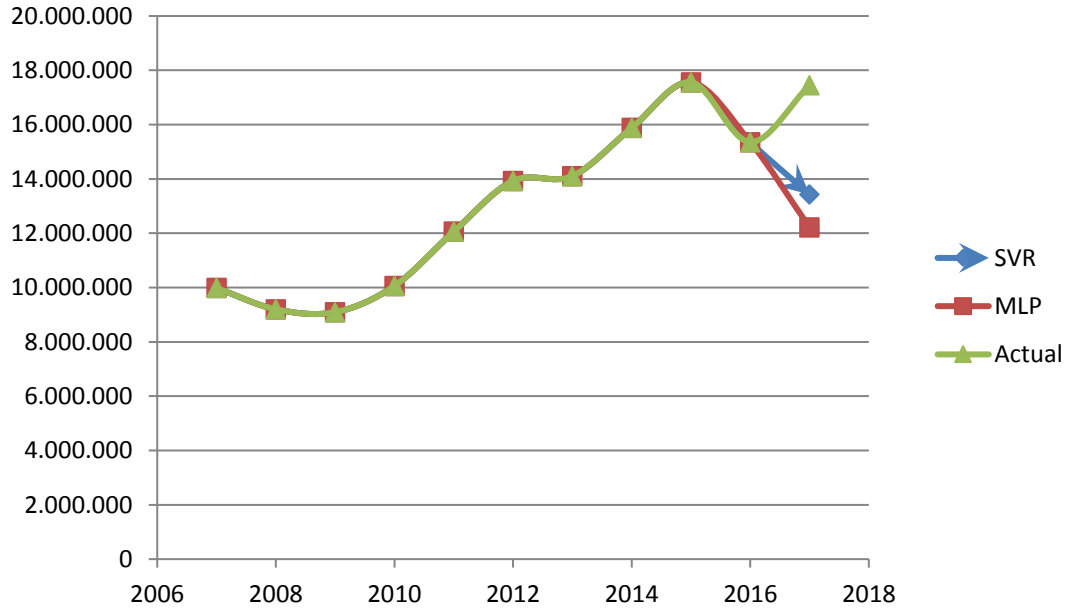


Table 19.

The comparison of estimated and real overnight numbers in Istanbul for 2017.

Model	Percentage of Error	Correlation coefficient	Relative Absolute error	Root relative squared error	Forecasting Values	Actual Values
SVR	23,037 %	0.9981	5.1491 %	7.9339 %	13.429.190	17.448.895
MLP	29,9727 %	1	0.0987 %	0.1145 %	12.218.994	17.448.895

Using the data of the years 2007-2016 for the variables that affect the tourist overnight numbers of the accommodation establishments in İstanbul province, the number of tourist overnight stays in 2017 was estimated by the Multilayer Perceptron regression method. While the real figure for 2017 was 17,448,895, the estimated figure was found to be 12,218,994. The percentage of error between real and predicted numbers was calculated as 29,9727%.

Using the data of the years 2007-2016 for the variables that affect the tourist overnight numbers of the accommodation establishments in İstanbul province, the number of tourist overnight stays in 2017 was estimated by the SVR regression method. The real figure for 2017 was 17.448.895, while the estimated figure was 13.429.190. Percentage of error between real and predicted numbers was calculated as 23,037%.

Graph 9: Comparison of SVR and MLP methods for the province of Muğla.

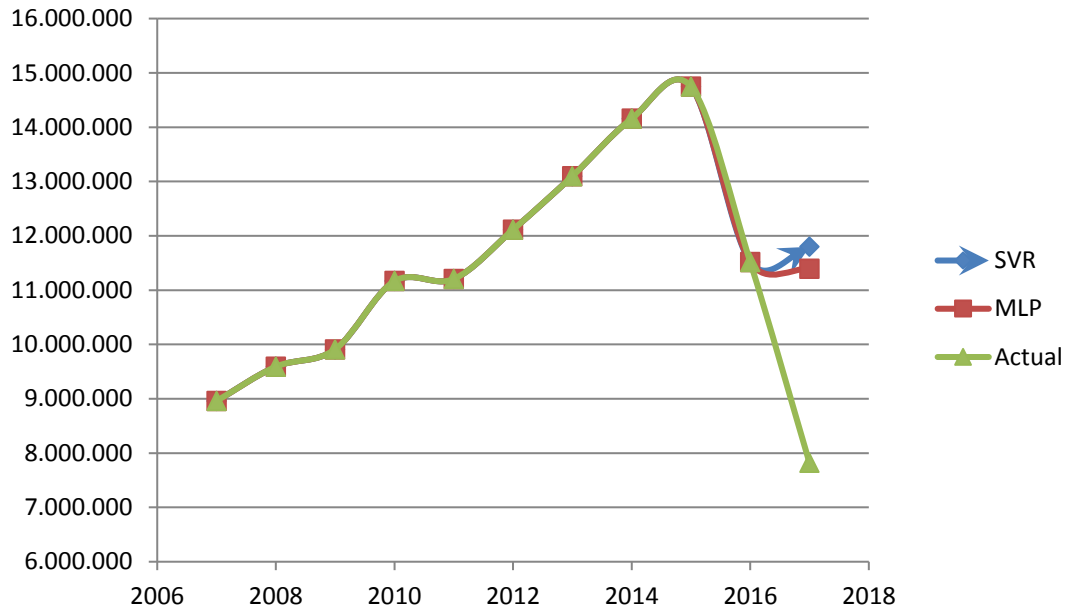


Table 20.

The comparison of estimated and real overnight numbers in Muğla for 2017.

Model	Percentage of Error	Correlation coefficient	Relative Absolute error	Root relative squared error	Forecasting Values	Actual Values
SVR	-50,8999 %	1	0.3609 %	0.3461 %	11.797.823	7.818.309
MLP	-45,7136 %	1	1.193 %	1.0705 %	11.392.338	7.818.309

Using the data of the years 2007-2016 for the variables that affect the tourist overnight numbers of the accommodation establishments in Muğla province, the number of tourist overnight stays in 2017 was estimated by the Multilayer Perceptron regression method. The real figure for 2017 was 7,818,309, while the estimated figure was found to be 11,392,338. Percentage of error between real and predicted numbers was calculated as -45,7136%.

Using the data of the years 2007-2016 for the variables that affect the tourist overnight numbers of the accommodation establishments in Muğla province, the number of tourist overnight stays in 2017 was estimated by the SVR

regression method. The real figure for 2017 was 7,818,309, while the estimated figure was found to be 11,797,823. Percentage of error between real and predicted numbers was calculated as -50,8999%.

Graph 10: Comparison of SVR and MLP methods for the province of Antalya.

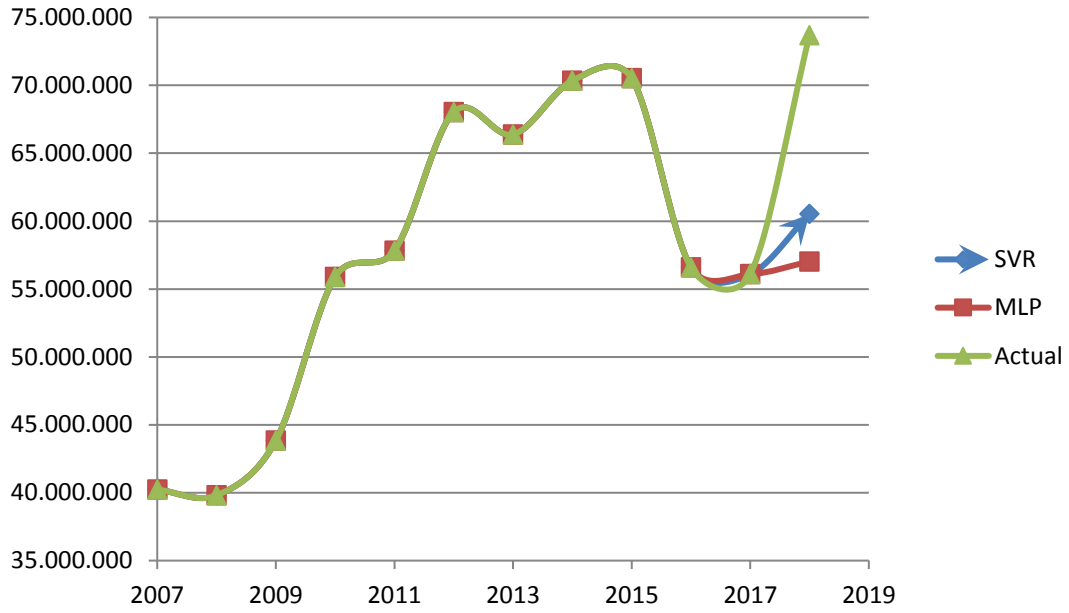


Table 21.

The comparison of estimated and real overnight numbers in Antalya for 2018.

Model	Percentage of Error	Correlation coefficient	Relative Absolute error	Root relative squared error	Forecasting Values	Actual Values
SVR	17,866 %	0.9953	8.0293 %	9.9676 %	60.523.781	73.689.106
MLP	22,621 %	0.9999	1.5152 %	1.5346 %	57.019.910	73.689.106

Using the data of the years 2007-2017 for the variables that affect the tourist overnight numbers of the accommodation establishments in Antalya province, the number of tourist overnight stays in 2018 was estimated by the Multilayer Perceptron regression method. While the real figure for 2018 was 73,689,106

the predicted figure was found to be 57,019,910. Percentage of error between real and predicted numbers was calculated as 22,621%.

Using the data of the years 2007-2017 for the variables that affect the tourist overnight numbers of the accommodation establishments in Antalya province, the number of tourist overnight stays in 2018 was estimated by the SVR regression method. Whereas the real figure for 2018 was 73,689,106 the predicted figure was found to be 60,523,781. Percentage of error between real and predicted numbers was calculated as 17,866%.

Graph 11: Comparison of SVR and MLP methods for the province of Muğla.

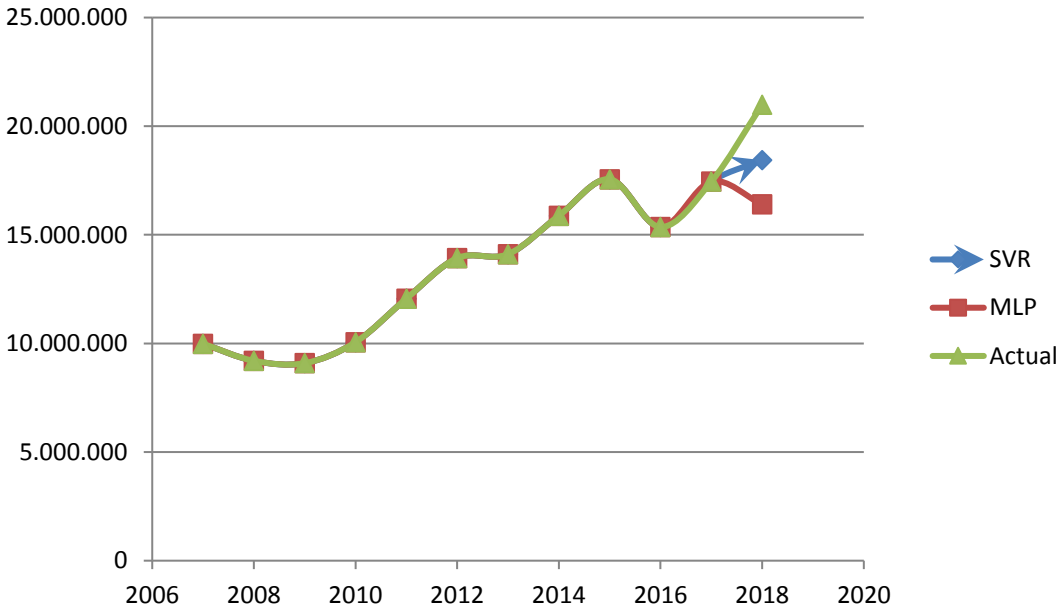


Table 22.

The comparison of estimated and real overnight numbers in Muğla for 2018.

Model	Percentage of Error	Correlation coefficient	Relative Absolute error	Root relative squared error	Forecasting Values	Actual Values
SVR	12,1506 %	0.9992	3.2123 %	4.7479 %	18.434.153	20.983.823
MLP	21,8438 %	1	0.1636 %	0.186 %	16.400.167	20.983.823

Using the data of the years 2007-2017 for the variables that affect the tourist overnight numbers of the accommodation establishments in İstanbul province, the number of tourist overnight stays in 2018 was estimated by the Multilayer Perceptron regression method. The real figure for 2018 was 20,983,823, while the estimated figure was 16,400,167. Percentage of error between real and predicted numbers was calculated as 21,8438%.

Using the data of the years 2007-2017 for the variables that affect the tourist overnight numbers of the accommodation establishments in İstanbul province, the number of tourist overnight stays in 2018 was estimated by the SVR regression method. The real figure for 2018 was 20,983,823, while the estimated figure was 18,434,153. Percentage of error between real and predicted numbers was calculated as 12,1506%.

Graph 12: Comparison of SVR and MLP methods for the province of Muğla.

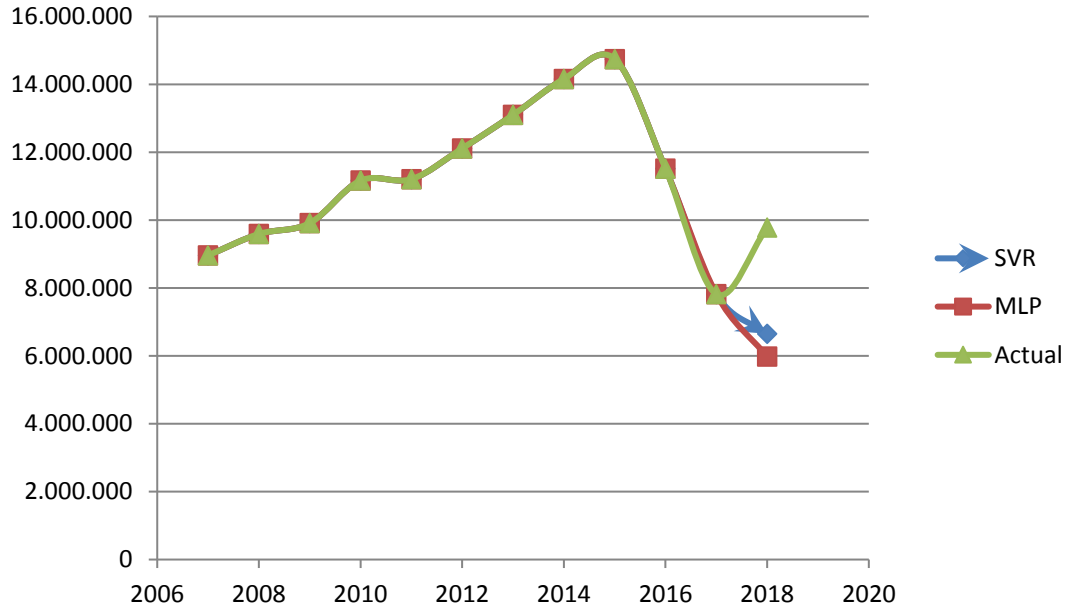


Table 23.

The comparison of estimated and real overnight numbers in Muğla for 2018.

Model	Percentage of Error	Correlation coefficient	Relative Absolute error	Root relative squared error	Forecasting Values	Actual Values
SVR	31,9587 %	0.9994	2.7962 %	3.6278 %	6.653.205	9.778.180
MLP	38,8590 %	0.9998	2.3448 %	2.2902 %	5.978.475	9.778.180

Using the data of the years 2007-2017 for the variables that affect the tourist overnight numbers of the accommodation establishments in Muğla province, the number of tourist overnight stays in 2018 was estimated by the Multilayer Perceptron regression method. While the real figure for 2018 was 73,689,106 the predicted figure was found to be 57,019,910. Percentage of error between real and predicted numbers was calculated as 22,621%.

Using the data of the years 2007-2017 for the variables that affect the tourist overnight numbers of the accommodation establishments in Muğla province, the number of tourist overnight stays in 2018 was estimated by the SVR regression method. Whereas the real figure for 2018 was 73,689,106 the predicted figure was found to be 60,523,781. Percentage of error between real and predicted numbers was calculated as 17,866%.

CHAPTER 5
CONCLUSION AND RECOMMENDATION

Different tourism destinations of Turkey have been dealt in this study. SVR analysis method have yielded the results closest to the actual values compared to Multilayer Perceptron analysis methods in the estimation study performed on number of overnight stays for the province of Antalya, Istanbul and Mugla.

In conclusion, it has been found in this study performed by multivariate data and different tourism destinations that different regression methods have been used for estimated values to yield the results closest to the actual values and regression analyses have shown variance by tourism destinations and determining the regression method according to the tourism destination planned to be estimated has yielded the estimated values closest to the actual values. In the graphs with high density of cyclical fluctuation, it was observed that SVR method gave the closest result to real values. In the graphs with less density from high of cyclical fluctuation, it was observed that MLP method gave the closest result to real values. In cases where the fluctuation intensity in the graphs increased, it was seen that SVR method's estimation success increased. In cases where the fluctuation intensities in the graphs of the data planned to be estimated decreased, the MLP regression analysis method showed results close to real values.

While determining the method for estimation in tourism destinations, was concluded that the selection of the appropriate regression method according to the cyclical fluctuations in the graphs would obtain the closest estimates to the real values.

REFERENCES

Aggarwal, C. C. (2015). *Data mining: the textbook*. Springer.

Aladag, C. H.,Egrioglu, E., Yolcu, U., & Uslu, V. R. (2014). A high order seasonal fuzzy time series model and application to international tourism demand of Turkey. *Journal of Intelligent&Fuzzy Systems*, *26*(1), 295-302.

Al-Alwani, M. A. (2014). *Data mining and statistical analysis of completions in the Canadian Montney formation.*

Athanasopoulos, G.,& Hyndman, R. J. (2008). Modelling and forecasting Australian domestic tourism. *Tourism Management*, *29*(1), 19-31.

Baggio, R., & Sainaghi, R. (2016). Mapping time series in to networks as a tool to assess the complex dynamics of tourism systems. *Tourism Management*, *54*, 23-33.

Berry, M. & Linoff, G. & Lucas, B. (2009). *Data Mining Techniques: Theory and Practice*

Berry, M. J., & Linoff, G. (1997). *Data mining techniques: for marketing, sales, and customer support*. John Wiley & Sons, Inc..

BULL, A., The Economics of Travel and Tourism, 2nd. Edition, Addison Wesley Longman Australia Pty Ltd., Melbourne, 1995, s. 28

Cang, S. (2014). A Comparative Analysis of Three Types of Tourism Demand Forecasting Models: Individual, Linear Combination and Non-linear Combination. *International Journal of Tourism Research*, *16*(6), 596-607.

Cankurt, S., & Subaşı, A. (2016). Tourism demand modelling and forecasting using data mining techniques in multivariate time series: a case study in Turkey. *Turkish Journal of Electrical Engineering & Computer Sciences*, *24*(5), 3388-3404.

Chen, C. F.,Lai, M. C., & Yeh, C. C. (2012). Forecasting tourism demand based on empirical mode decomposition and neural network. *Knowledge-Based Systems*, *26*, 281-287.

Chen, K. Y.,& Wang, C. H. (2007). Support vector regression with genetic algorithms in forecasting tourism demand. *Tourism Management*, *28*(1), 215-226.

Chen, M. H. (2013). Determinants of theTaiwanese tourist hotel industry cycle. *Tourism Management*, *38*, 15-19.

Chen, M. S.,Ying, L. C., & Pan, M. C. (2010). Forecasting tourist arrivals by using the adaptive network-based fuzzy inference system. *Expert System swith Applications*, *37*(2), 1185-1191.

Chern, C. C.,Wei, C. P., Shen, F. Y., & Fan, Y. N. (2015). A salesforecasting model for consumer products based on theinfluence of online word-of-mouth. *Information Systems and e-Business Management*, *13*(3), 445-473.

Chu, F. L., “Forecasting Tourism Demand: A Cubic Polynomial Approach”, Tourism Management, Volume 25, Issue 2, 2004, s. 209

Claveria, O.,& Torra, S. (2014). Forecasting tourism demand to Catalonia: Neural networks vs. time series models. *Economic Modelling*, *36*, 220-228.

Cloyd,J. D. (2002). *Data Mining with Newton’s Method (Doctoral dissertation,* East Tennessee State University*).*

Corgel, J.,Lane, J., & Walls, A. (2013). How currency Exchange rates affect the demand for US hotel rooms. *International Journal of Hospitality Management*, *35*, 78-88.

Cortés-Jiménez, I.,& Blake, A. (2011). Tourism demand modeling by purpose of visit and nationality. *Journal of Travel Research*, *50*(4), 408-416.

Croes, R.,& Semrad, K. J. (2012). Discounting works in the hotel industry: a structural approach to understanding why. *TourismEconomics*, *18*(4), 769-779.

Cuhadar, M.,Cogurcu, I., & Kukrer, C. (2014). Modelling and forecasting cruise tourism demand to Izmir by different artificial neural network architectures. *International Journal of Business and Social Research*, *4*(3), 12-28.

Dai, W.,Chuang, Y. Y., & Lu, C. J. (2015). A Clustering-based Sales Forecasting Scheme Using Support Vector Regression for Computer Server. *Procedia Manufacturing*, *2*, 82-86.

Dai, W.,Wu, J. Y., & Lu, C. J. (2014). Applying different in dependent component analysis algorithms and support vector regression for IT chain store sales forecasting. *The Scientific World Journal*, *2014*.

Ding, Y. (2014). *Estimating truncated hotel demand: A comparison of low computational cost forecasting methods* (Doctoral dissertation, Purdue University).

DiPillo, G.,Latorre, V., Lucidi, S., & Procacci, E. (2016). An application of support vector machines to sales forecasting underpromotions. *4OR*, 1-17.

Duriqi, R., Raca, V., & Cico, B. (2016, June). Comparative analysis of classification algorithms on three different datasets using WEKA. In *Embedded Computing (MECO), 2016 5th Mediterranean Conference on*(pp. 335-338). IEEE.

Falk, M. (2014). Impact of weather conditions on tourism demand in the peak summer season over the last 50 years. *Tourism Management Perspectives*, *9*, 24-35.

Fang, D.,& Weng, W. (2011). Sales Forecasting System for Chinese Tobacco Wholesalers. *Procedia Environmental Sciences*, *11*, 380-386.

Gaojun, L.,& Boxue, L. (2009, December). There search on Combination forecasting model of the automobile sales forecasting system. In *Computer Science-Technologyand Applications, 2009. IFCSTA'09. International Forum on*(Vol. 3, pp. 82-85). IEEE.

García, F. T.,Villalba, L. J. G., & Portela, J. (2012). Intelligent system for time series classification using support vector machines applied to supply-chain. *Expert Systems with Applications*, *39*(12), 10590-10599.

Giudici, P. (2005). *Applied data mining: Statistical methods for business and industry*. John Wiley & Sons.

Gomez, D. D.& Agudo, D.&Castroman,L.J.&Santacruz, C. Rodriguez,A.A (2011). Improving sale performance prediction using support vector machines. *Expert Systems with Applications*, *2011*

Guizzardi, A.,& Stacchini, A. (2015). Real-time forecasting regional tourism with busines ssentiment surveys. *Tourism Management*, *47*, 213-223.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hemambika, P. (2005). *Data Mining Framework Dissertation*

Hong, W. C.,Dong, Y., Chen, L. Y., & Wei, S. Y. (2011). SVR with hybrid chaotic genetic algorithms for tourism demand forecasting. *Applied Soft Computing*, *11*(2), 1881-1890.

Ibrahim, A. M. (1995).*Forecasting Hotel Occupancy Ratesin Egypt Using Time Series Models.*

Kadambi,R.R. (2005) *Analysis of Data Mining Techniques for Customer Segmentation and Predictive Modeling-A Case Study*

Khalil Zadeh, N.,Sepehri, M. M., & Farvaresh, H. (2014). Intelligent Sales Prediction for Pharmaceutical Distribution Companies: A Data Mining Based Approach. *Mathematical Problems in Engineering*, *2014*.

Kong, X.,Liu, X., Shi, R., & Lee, K. Y. (2015). Wind speed prediction using reduced support vector machines with feature selection. *Neurocomputing*, *169*, 449-456.

Kouprioucha, L.,van der Rest, J. P., & Schwartz, Z. (2014). On revenue management and the use of occupancy forecasting error measures. *International Journal of Hospitality Management*, *41*, 104-114.

Larose, D. T., & Larose, C. D. (2014). *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons.

Lim, C.,Chang, C., & McAleer, M. (2009). Forecasting h (m) otel guest nights in New Zealand. *International Journal of Hospitality Management*, *28*(2), 228-235.

Lim, J. (2013). *Anefficient approach to clustering datasets with mixed type attributes in data mining*.

Lin, C. J.,Chen, H. F., & Lee, T. S. (2011). Forecasting tourism demand using time series, artificial neural networks and multivariate adaptive regression splines: Evidence from Taiwan. *International Journal of Business Administration*, *2*(2), 14.

Lior, R. (2014). *Data mining with decision trees: theory and applications* (Vol. 81). World scientific.

Liu, B. (2007). *Web data mining: exploring hyperlinks, contents, and usage data*. Springer Science & Business Media.

Lu, C. J. (2014). Sales forecasting of computer products based on variables election scheme and support vector regression. *Neurocomputing*, *128*, 491-499.

Moon, S.,Bae, S., & Kim, S. (2014, August). Predicting the Near-Weekend Ticket Sales Using Web-Based External Factorsand Box-Office Data. In *Proceedings of the 2014 IEEE/WIC/ACM International Joint*

Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 02(pp. 312-318). IEEE Computer Society.

Naik, A., & Samant, L. (2016). Correlation review of classification algorithm using data mining tool: WEKA, Rapidminer, Tanagra, Orange and Knime. *Procedia Computer Science*, *85*, 662-668.

Olson, D. L., & Delen, D. (2008). *Advanced data mining techniques*. Springer Science & Business Media.

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*. " O'Reilly Media, Inc.".

Pyle, D. (1999). *Data preparation for data mining* (Vol. 1). morgan kaufmann.

Rakthanmanon, T. (2012). *Efficient Algorithms for High Dimensional Data Mining*

Seliem,M. L. (2006) .*Foreign Exchange Forecasting Using Artificial Neural Networkas Data Mining Tool*.

Siripitayananon, P. (2002). *Data mining techniques for handling a missing data problem*. The University of Alabama.

Smeral, E. (2010). Impacts of the world recession and economic crisis on tourism: Forecasts and potential risks. *Journal of Travel Research*, *49*(1), 31-38.

Song, H. And WITT, S. F. (a), Tourism Demand Modelling and Forecasting: Modern Econometric Approach, Elsevier Science - Pergamon, 2000, s. 9

Song, H., Lin, S., Witt, S. F., & Zhang, X. (2011). Impact offinancial/economic crisis on demand for hotel rooms in Hong Kong. *Tourism Management*, *32*(1), 172-186.

Tezel, G.,& Buyukyildiz, M. (2016). Monthly evaporation forecasting using artificial neural networks and supportvectormachines. *Theoretical and Applied Climatology*, *124*(1-2), 69-80.

Tsaur, R. C.,& Kuo, T. C. (2011). The adaptive fuzzy time series model with an application to Taiwan's tourism demand. *Expert systems with Applications*, *38*(8), 9164-9171.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wu, D.,Song, H., & Shen, S. (2016). New Developments in Tourismand Hotel Demand Modeling and Forecasting. *International Journal of Contemporary Hospitality Management*.

Wu, E. H.,Law, R., & Jiang, B. (2010). Data mining for hotel occupancy rate: an independent component analysis approach. *Journal of Travel &Tourism Marketing*, *27*(4), 426-438.

Xu, X.,Law, R., & Wu, T. (2009). Support vector machines with manifold learning and probabilistic space projection for tourist expenditure analysis. *International Journal of Computational IntelligenceSystems*, *2*(1), 17-26.

Ye, N. (Ed.). (2003). *The handbook of data mining* (Vol. 24). Mahwah, NJ/London: Lawrence Erlbaum Associates, Publishers.

Yu, X.,Qi, Z., & Zhao, Y. (2013). Support Vector Regression for Newspaper/Magazine Sales Forecasting. *Procedia Computer Science*, *17*, 1055-1062.

Yunyue, Z. (2004). *High Performance Data Mining in Time Series: Techniques and Case Studies* (Doctoral dissertation, PHD. Thesis). New York City: NY University, 2004.138 139).

Zhang, F.& Deb,C.& Lee, E.S.& Yang, J.& Shah, W.K. (2016). Time series forecasting for building energy consumption using weighted Support

Vector Regression with differential evolution optimization technique. *Energy and Building*, *2016*

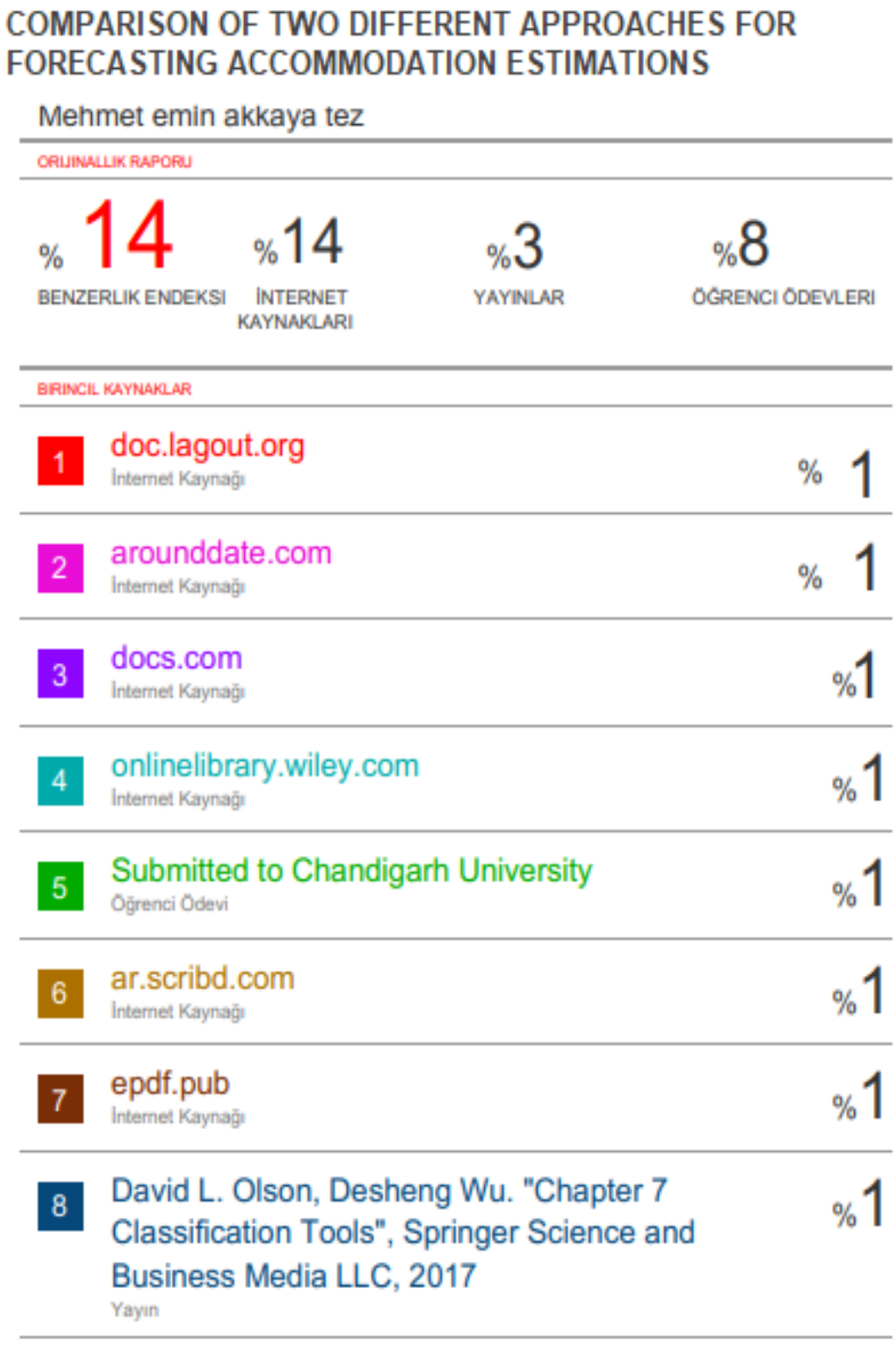
Zhao, W.& Tao, T.& Zio, E.(2013).Parameters Tuning in Support Vector Regression for Reliability Forecasting. *Chemical Engineering Transactions*, *2013*

Zheng, T. (2014). What caused the decrease in RevPAR during the recession? An ARIMA within tervention analysis of room supply and market demand. *International Journal of Contemporary Hospitality Management*, *26*(8), 1225-1242.

BIOGRAPHY

Mehmet Emin AKKAYA was born on January 29, 1986 in Sanliurfa, Turkey. He completed his primary, secondary and high school education in Sanliurfa. He graduate of Harran University Computer Technologies and Programming Associate's Degree. He graduated from Anadolu University, Faculty of Business Administration with a bachelor's degree. He graduated from Gazi University, Department of Management Information Systems with a master's degree. He completed his PhD in Near East University, Department of Business Administration. He works as a public employee in the Ministry of Culture and Tourism. He has published articles in international refereed journals and papers presented at international congresses. He is a journal referee in an SSCI indexed journal. He is the editor-in-chief of an international refereed journal.

PLAGIARISM REPORT



COMPARISON OF TWO DIFFERENT APPROACHES FOR FORECASTING ACCOMMODATION ESTIMATIONS

Mehmet emin akkaya tez

ORIJINALLIK RAPORU

%14	%14	%3	%8
BENZERLIK ENDEKSI	İNTERNET KAYNAKLARI	YAYINLAR	ÖĞRENCI ÖDEVLERI

BIRINCIL KAYNAKLAR

1	doc.lagout.org İnternet Kaynağı	% 1
2	arounddate.com İnternet Kaynağı	% 1
3	docs.com İnternet Kaynağı	%1
4	onlinelibrary.wiley.com İnternet Kaynağı	%1
5	Submitted to Chandigarh University Öğrenci Ödevi	%1
6	ar.scribd.com İnternet Kaynağı	%1
7	epdf.pub İnternet Kaynağı	%1
8	David L. Olson, Desheng Wu. "Chapter 7 Classification Tools", Springer Science and Business Media LLC, 2017 Yayın	%1

ETHICS COMMITTEE APPROVAL

YAKIN DOĞU ÜNİVERSİTESİ

BİLİMSEL ARAŞTIRMALAR ETİK KURULU

18.09.2019

Dear Mehmet Emin Akkaya

Your project **"Comparison of two different approaches for forecasting accommodation estimations"** has been evaluated. Since only secondary data will be used the project it does not need to go through the ethics committee. You can start your research on the condition that you will use only secondary data.

Assoc. Prof. Dr. Direnç Kanol

Rapporteur of the Scientific Research Ethics Committee

Direnc Kanol

Note: If you need to provide an official letter to an institution with the signature of the Head of NEU Scientific Research Ethics Committee, please apply to the secretariat of the ethics committee by showing this document.