

DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

**COMPARISON OF PREDICTIVE
PERFORMANCES OF LOGISTIC REGRESSION
AND DECISION TREE ALGORITHMS: AN
APPLICATION OF LIFE SATISFACTION**

by
Arzu YAVUZ

August, 2021
İZMİR

**COMPARISON OF PREDICTIVE
PERFORMANCES OF LOGISTIC REGRESSION
AND DECISION TREE ALGORITHMS: AN
APPLICATION OF LIFE SATISFACTION**

**A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Master of Science in
Statistics, Statistics Program**

**by
Arzu YAVUZ**

**August, 2021
İZMİR**

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled “**COMPARISON OF PREDICTIVE PERFORMANCES OF LOGISTIC REGRESSION AND DECISION TREE ALGORITHMS: AN APPLICATION OF LIFE SATISFACTION**” completed by **ARZU YAVUZ** under supervision of **ASSOC. PROF. DR. ÖZGÜL VUPA ÇİLENGİROĞLU** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

.....
Assoc. Prof. Dr. Özgül VUPA ÇİLENGİROĞLU

Supervisor

.....
Assoc. Prof. Dr. Hakan SAVAŞ SAZAK

(Jury Member)

.....
Assoc. Prof. Dr. İdil YAVUZ

(Jury Member)

Prof.Dr. Özgür ÖZÇELİK

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENT

I would like to thank my supervisor Associate Prof. Dr. Özgül Vupa Çilengiroğlu for her hard work, attention, patience and everything she has given me, for which I have had the opportunity to benefit from his valuable knowledge and experience for a long time during my graduate education.

I would like to express my gratitude to all of my other university professors, who will be less than welcome, for everything they have given me throughout my university life and for providing me with the knowledge that will make me have a say in the future.

And finally, my endless thanks to my family, which is my biggest chance in life, for their support and trust in my work and for bringing me to these days by raising them to know the meaning of the words love and respect.

Arzu YAVUZ

COMPARISON OF PREDICTIVE PERFORMANCES OF LOGISTIC REGRESSION AND DECISION TREE ALGORITHMS: AN APPLICATION OF LIFE SATISFACTION

ABSTRACT

Decision tree algorithms and regression in machine learning create classes of data. Relationships between variables are modeled. Decision trees create classification rules using training data. They also test these rules on test data. Thus, the decision tree determines the success of the algorithm. With the model created in logistic regression, classification is created and classification performance is found. These methods are easy to interpret. They are easily applied to large data sets. They are used in many different fields due to the lack of assumptions.

Satisfaction, which is a part of the concept of life satisfaction, is the fulfillment of needs, desires and wishes. Life satisfaction deals with a person's entire life. Life satisfaction is the whole of processes related to individuals' own life patterns and standards.

The aim of this study is to compare the performances of logistic regression method and decision tree algorithms (CART, CHAID, QUEST) estimators using life satisfaction data ($n = 8430$) obtained by the Turkish Statistical Institute (TURKSTAT) for the year 2017. In this study, performance comparisons (accuracy, sensitivity, selectivity, precision, F-score) were made and it was found that the model that best explains the concept of life satisfaction is the QUEST algorithm.

Keywords: Machine learning, decision tree algorithms, logistic regression, life satisfaction

LOJİSTİK REGRESYON VE KARAR AĞACI ALGORİTMALARININ TAHMİN EDİCİ PERFORMANSLARININ KARŞILAŞTIRILMASI: YAŞAM MEMNUNİYETİ UYGULAMASI

ÖZ

Makine öğrenimindeki karar ağacı algoritmaları ve regresyon, veri sınıfları oluşturur. Değişkenler arasındaki ilişkiler modellenir. Karar ağaçları, eğitim verilerini kullanarak sınıflandırma kuralları oluşturur. Ayrıca bu kuralları test verileri üzerinde test eder. Böylece karar ağacı algoritmanın başarısını belirler. Lojistik regresyonda oluşturulan model ile sınıflandırma oluşturulur ve sınıflandırma performansı bulunur. Bu yöntemlerin yorumlanması kolaydır. Büyük veri kümelerine kolayca uygulanır. Varsayımların olmaması nedeniyle birçok farklı alanda kullanılmaktadır.

Yaşam memnuniyeti kavramının bir parçası olan memnuniyet, ihtiyaçların, arzuların ve isteklerin karşılanmasıdır. Yaşam memnuniyeti, bir kişinin tüm yaşamıyla ilgilenir. Yaşam memnuniyeti, bireylerin kendi yaşam kalıpları ve standartları ile ilişkili süreçler bütünüdür.

Bu çalışmanın amacı, Türkiye İstatistik Kurumu (TÜİK) tarafından 2017 yılı için elde edilen yaşam memnuniyeti verilerini (n=8430) kullanarak lojistik regresyon yöntemi ve karar ağacı algoritmaları (CART, CHAID, QUEST) tahmin edicilerinin performanslarını karşılaştırmaktır. Bu çalışmada performans karşılaştırmaları (doğruluk, duyarlılık, seçicilik, kesinlik, F-puanı) yapılmış ve yaşam memnuniyeti kavramını en iyi açıklayan modelin QUEST algoritması olduğu bulunmuştur.

Anahtar kelimeler: Makine öğrenimi, karar ağacı algoritmaları, lojistik regresyon, yaşam memnuniyeti

CONTENTS

	Page
M.Sc THESIS EXAMINATION RESULT FORM	ii
ACKNOWLEDGMENT	iii
ABSTRACT	iv
ÖZ	v
LIST OF FIGURES	ix
LIST OF TABLES	x
CHAPTER 1 - INTRODUCTION.....	1
CHAPTER 2 - DATA MINING AND MACHINE LEARNING	3
2.1 Big Data and Data Mining	3
2.2 Machine Learning.....	4
2.2.1 History of Machine Learning.....	5
2.2.2 Machine Learning Application Areas	6
2.3 Machine Learning Methods	8
2.3.1 Supervised Learning.....	9
2.3.1.1 Classification	11
2.3.1.2 Regression	11
2.3.2 Unsupervised Learning.....	12
2.3.3 Semi-supervised learning	12
2.4. Machine Learning Steps	13
CHAPTER 3 - LOGISTIC REGRESSION	15
3.1 Regression Analysis	15
3.2 Literature Reviews.....	16
3.3 Introduction to Logistic Regression Model	19
3.4 Fitting a Simple Logistic Regression Model.....	21
3.4.1 Likelihood Function and Maximum Likelihood Estimation Method	22

3.4.2 Testing for the Significance of the Coefficients	24
3.4.2.1 Likelihood Ratio Test	24
3.4.2.2 Wald Test	25
3.4.2.3 Score Test	26
3.5 Fitting a Multiple Logistic Regression Model	26
3.5.1 Design Variable	27
3.5.2 Testing for the Significance of the Coefficients	27
3.5.2.1 Likelihood Ratio Test	27
3.5.2.2 Wald Test	28
3.5.3 Interpretation of the Coefficients	29
3.6 Comparison of Model Selection Criteria	31
3.6.1 Classification Criterias	31
CHAPTER 4 - DECISION TREE	33
4.1 Decision Tree	33
4.1.1 Structure of the Decision Tree	35
4.1.2 Creating a Decision Tree	36
4.1.3 Accuracy of Classification Process (Train and Test Data).....	38
4.1.4 Decision Trees with Advantages and Disadvantages.....	38
4.2 Decision Tree Algorithms.....	39
4.2.1 CART Algorithm	40
4.2.1.1 Literature Review	40
4.2.1.2 Basic Steps of CART Algorithm.....	41
4.2.1.3 Accuracy Tests of CART Algorithm	45
4.2.2 CHAID Algorithm	46
4.2.2.1 Literature Review	48
4.2.2.2 Steps of the CHAID Algorithm	49
4.2.3 QUEST Algorithm	50
4.2.3.1 Literature Review	50

CHAPTER 5 - APPLICATION	53
5.1 Purpose of Application	53
5.2 Life Satisfaction Definition and Literature Review	53
5.3 Data Definition	55
5.4. Data Analysis	55
5.4.1 Descriptive Statistics	55
5.4.2 Examination of Factors Affecting Life Satisfaction with Logistic Regression Model.....	59
5.4.3 Examination of Factors Affecting Life Satisfaction with CART Algorithm.....	63
5.4.4 Examination of Factors Affecting Life Satisfaction with CHAID Algorithm.....	70
5.4.5 Examination of Factors Affecting Life Satisfaction with QUEST	74
5.5 Performance Comparisons	77
5.6 Model Selection.....	79
CHAPTER 6 - CONCLUSION	83
REFERENCES	88

LIST OF FIGURES

	Page
Figure 2.1 Machine Learning Algorithm	9
Figure 3.1 Logistic Regression Plot.....	20
Figure 4.2 N-Fold Cross Validation	45
Figure 5.1 CART Analysis (80% Training Data) Decision Tree Numbering.....	64
Figure 5.2 CHAID Analysis (80% Training Data) Decision Tree Numbering	71
Figure 5.3 QUEST Analysis (80% Training Data) Decision Tree Numbering.....	75
Figure 5.4 Performance of the Models	79
Figure 5.5 QUEST Analysis (100% Training Data) Decision Tree Numbering.....	80
Figure 5.6 SPSS Output for QUEST	80

LIST OF TABLES

	Page
Table 3.1 Number of Subjects Classified by Two Classifications	31
Table 5.1 Chi-Square p-values for Some Variables	56
Table 5.2 Frequencies, Percentages and Chi-Square p-values for Independent Variables	57
Table 5.3 Frequencies and Chi-Square p-values for Satisfaction Scales (S: Satisfied, DS: Dissatisfied).....	59
Table 5.4 Logistic Reg. Model (80% Trainig Data): Wald Chi-Square p-Values and OR(CI)	60
Table 5.5 Performance Evaluation for Logistic Regression Model	63
Table 5.6 CART Analysis (80% Training Data).....	65
Table 5.7 CART Analysis Normalized Significance Values	66
Table 5.8 Nodes and Results for CART Algorithm	67
Table 5.9 Performance Evaluation for CART Algorithm.....	69
Table 5.10 k-fold Cross Validation.....	70
Table 5.11 CHAID Analysis (80% Training Data)	72
Table 5.12 Nodes and Results for CHAID Algorithm.....	73
Table 5.13 Performance Evaluation for CHAID Algorithm	74
Table 5.14 Quest Analysis (80% Training Data)	75
Table 5.15 Nodes and Results for Quest Algorithm.....	76
Table 5.16 Performance Evaluation for QUEST Algorithm.....	77
Table 5.17 Node Numbers for CART CHAID and QUEST Algorithms	77
Table 5.18 Performance Evaluation According to Algorithms of Models	78
Table 5.19 Quest Analysis (100% Training Data).....	81
Table 5.20 Nodes and Results for Quest Algorithm.....	82

CHAPTER 1

INTRODUCTION

Developments in computer technologies and cheaper computer hardware have enabled large data to be stored. With the use of these data stored in large databases, the concept of information discovery in databases has emerged. Data mining is the process that expresses the extraction of this meaningful and useful information from the data in databases using statistical methods and various computer algorithms. Understanding this process and the classification methods and algorithms used constitute the main subject of this study.

One of the processes that makes unprocessed information stored in large data sets understandable and interpretable is data mining. It is also possible to find undiscovered data in data mining. For this, a decision process must be carried out.

Decisions are a daily part of human life. There are a wide variety of outcomes, both good and bad, as a result of these decisions. Finding ways to improve the decision-making process helps us achieve more positive results. Marks (2018) "taking advice (Hütter & Ache, 2016) quantitative aids (Diab et al., 2011; Highhouse, 2008), decision aids (Diab et al., 2011; Highhouse, 2008) and heuristics. (Artinger, Petersen, Gigerenzer, & Weibler, 2015)" as an example.

Multiple regression, logistic regression, neural networks and decision trees increase the accuracy of decisions and improve decision making processes (Coussement, Van den Bossche, & De Bock, 2014; Stauffer & Ree, 1996). Of these approaches, decision trees are gaining interest by researchers Coussement et al., (2014); Liu et al., (2011); Sinha & May, (2004); Stewart, Tuerk, Metzger, Davidson, & Young, (2016).

One of the most common methods of data mining, which reveals important classes by classifying large data sets in sample studies or used to predict data estimates, is classification and regression models. Among these models, the preferred methods are techniques such as Naive Bayes, support vector machine, k-nearest

neighborhood, decision trees, linear and logistic regression analysis. Although especially the classification used in models is an important resource for multivariate analysis, it has a wide range of uses in various disciplines, including health.

In this study, in the second chapter following the introduction, basic concepts such as big data, databases, model and learning types were investigated. In addition, data mining processes such as data preparation, model establishment and evaluation were explained.

Literature studies on different sectors were also included. In the third chapter, general information about simple and multiple logistic regression models, interpretation of coefficients, and modeling strategies and methods for logistic regression model were given. In the fourth chapter, the way decision trees work, the separation and stopping criteria in the trees were explained, and the CART, CHAID and QUEST tree algorithms were discussed.

In the fifth chapter, the application is made for Turkey Statistics Institution of life satisfaction derived from 2017 data. Using this information, models that estimate life satisfaction have been developed and compared. The purpose of establishing the model was to determine the variables that affect the decision of life satisfaction and to help the model to be used for new data coming to the system with an additional study that can be done later. In this context, logistic regression, decision tree models were applied, the results were given, and comparisons were made between them. This study was concluded with the conclusion and the suggestion, which is the sixth chapter, where the comparison and explanation of the analyzes made are included.

CHAPTER 2

DATA MINING AND MACHINE LEARNING

2.1 Big Data and Data Mining

The term “big data” was first used in an article by Cox & Ellsworth (1997). In this article, he mentioned that the data sets are very large. For this reason, he stated that the computer system fills its memory, disks and even external disks. He named this problem the “Big Data Problem” (Cox & Ellsworth, 1997). According to Gahi et al. (2016) defined big data as a concept created from various contents that describe heterogeneous data of different sizes that cannot be processed provided that traditional database methods are used (Gahi, 2016). The data are divided into three as structural, semi-structural or non-structural. Structured data is a rigid structure that includes modeling, processing, storing, querying, processing and visualizing data as input. There are many different methods and models in this data in which structurality is defined. Semi-structured data includes models in which structurality is defined, as well as various meta models such as labels and signs used to describe a particular element and a hierarchical representation of different areas in the data. Non-structured data are types of records that are presented and stored, except for a defined format such as books, articles, documents.

The concept of big data dates back to the 1970s when relational databases were developed. This term, which simply describes a large amount of data, gained popularity in the early 2000s when data generated by users began to be researched and analyzed via social media.

The usage area of data mining is very wide. Data mining, which is frequently used in database analysis and decision support mechanisms, in market research (similarity detection, basket analysis, cross market review, customer profile determination, campaign analysis, sales forecast), risk analysis (quality control, competition analysis, foresight and fraud detection, credit card analysis, banking and insurance analysis, customer credit risk research), inter-document similarity (e-mail analysis) and institutional analysis (optimal use of resources, future prediction in the presence of past and current data).

Data mining is a process study. The data owner (company, people, institution...) is first collected and uploaded to the data warehouse and then stored and managed in internal presentations. Business analysts, information technology experts and management teams are assigned to access the data. The steps followed in the data mining process are: (1) Identification of the problem, (2) Preparation of data (collection and harmonization, assembly and cleaning, selection), (3) Model establishment and evaluation, (4) Model usage, (5) Model monitoring. There are many methods used at various stages of the data mining process. These methods are listed as statistical methods (classification, regression, clustering ...), memory-based methods, genetic algorithms, neural networks and decision trees.

2.2 Machine Learning

Machine learning has become the foundation of information technologies in the past years. However, machine learning has become a part of our lives even if it is not noticed. Data Today is the technological advance of effective data analysis. It is thought that data will become more common as a necessary component. Machine learning obtained from data can be expressed as automatic obtaining of meaningful patterns (TDK, 2019).

Machine learning using sample data set or using from past experience is a computer program that optimizes performance criteria by making use of a problem. We need an algorithm to solve it on the computer. The algorithm is a series of commands that convert its output (Alpaydm, 2010). The purpose of machine learning is learning that automatically makes learning without human intervention or assistance designing algorithms (Schapire, 2008). Learning algorithm data is called "input" and information obtained from learning is called "output" (Shwartz & Ben-David, 2014).

The machine learning that obtained from large data sets for the past few decades it has become one of the most common tools used to obtain information. In our diarly lives, we are intertwined with technology based on machine learning. Detecting credit card frauds, digital cameras and face recognition of smartphones, voice commands of cars, accident prevention of cars systems are examples of these uses.

Also machine learning astronomy, it is also frequently used in scientific applications such as medicine and bioinformatics. Human most of their skills are learned or developed by taking examples from intelligent beings and experiencing them. Machine learning is also about programming with the ability to learn and adapt (Shalev-Shwartz & Ben-David, 2014).

With the development of computer technology large amounts of data storage are offered. In recent years, databases its growth is incredible.

2.2.1 History of Machine Learning

Machine Learning is the general name of computer algorithms that models a given problem according to the data obtained from the problem environment. Since it is an intensively studied subject, there are many proposed approaches and algorithms. Some of these approaches have the ability to make prediction and estimation, and some of them have the ability to classification.

In 1949, Donald Hebb developed a learning rule that modifies the values of an artificial neural network made up of artificial cells. This rule, called the "Hebbian learning" rule, forms the basis of many learning rules today. At the same time, studies have emerged from laboratories and become systems used in daily life.

Turing created the Turing Test, which bears his name, in 1950. The purpose of the test was to determine whether a computer had human intelligence. However, the Article did not express the words "thinking" and "machine" in a clear way to satisfy people. Turing advises that we "replace the question with another that is closely related to it and is phrased in relatively straightforward words." Turing created the "Turing Test". Rosenblatt (1958), the first artificial neural network "perceptron" pattern and shape recognition designed for the purpose. Perceptron is a neural network type.

"Nearest neighbor algorithm" was written in 1967. With this algorithm, computers are enabled to recognize simple patterns (Marr, 2016; Foote, 2019).

In the early periods of the artificial intelligence field (1950s and 1960s), the dominant positive perspective was reflected in the definitions of artificial intelligence. According to this approach, a machine can emulate human intelligence.

With the 1970s, the disappointments and heavy criticisms experienced in the projects caused a change in the field of interest of the artificial intelligence discipline. New themes emerging in the field; knowledge-based systems, expert systems, and connective networks. This situation has brought with it a new understanding of artificial intelligence.

2.2.2 Machine Learning Application Areas

Machine learning is very important in the digital age. Image recognition, one of the most common uses of machine learning, can detect faces and photographs with the help of advanced machine learning systems. Today, many channels such as banks and social media platforms identify people with image recognition systems. Airports also use facial recognition systems to question the authenticity and reliability of passengers.

Machine learning techniques used on the Internet, such as web pages related to a word entered in a search engine opening applications, automatic translation of a text into the language we want, or facial recognition programs that can identify who a particular photo belongs to. (Smola & Vishwanathan, 2008). It is known that the human face is not unique in face recognition programs and is identified with many factors. The basis of face tagging in social networks is done with machine learning algorithms (Das et al., 2015).

The routing system is pretty heavily implemented in machine learning. Referral systems introduce us to products based on previous search systems. Such systems are implemented in various online retail environments, including Amazon and Netflix (Shutz, 2017).

Machine learning algorithms are developed in customer relationship management systems, business intelligence analytics, human resources systems, self-driving

vehicles, learning to identify the characteristics of effective employees in areas such as creating models and finding the best candidates for open positions (Oluk, 2019).

The collection and storage of more data in the field of health, the new discovery of diseases, the development of new diagnostic methods and the complexity of data types have increased the need for machine learning methods.

Machine learning is also frequently used in the field of health. For machine learning in the field of medicine, it reduces the production time of drugs, identifying suitable patients in clinical trials and defining diseases more clearly are given as an example.

In addition to these, machine learning also contributes greatly to the biomedical field. Because it increases the sensitivity and selectivity in the diagnosis of the disease, as well as making the decision-making process more objective (Sajda, 2006).

Today, it has become possible for various artificial intelligence algorithms such as machine learning to obtain meaningful results by processing large numbers of data. With the inferences obtained from these data, machines or computers can make accurate predictions on new data.

The most basic thing in the treatment of a disease is to detect it correctly. With the help of some technological developments in image processing with artificial intelligence, sensitive diagnostic studies independent of subjective results can be made nowadays. Capper and his team are working on these machine learning-based technologies and have presented a new classification of brain tumors based on molecular models.

The classification studies for correct diagnosis related to male infertility were made with different machine learning algorithms (Palechor et al., 2016).

Analysis of ECG results to classification studies in the field of health was used for detection of a lung nodule and a chest x-ray in radiology (Deo, 2018).

Medical care starts with the correct diagnosis. Machine learning is most prominent area is the diagnosis of the correct disease (Sharma, 2019).

In social sciences such as economics, psychology and sociology, the machine learning techniques were used (Korkmaz et al., 2015). The life satisfaction was investigated with the machine learning techniques using 2013 TURKSTAT data (Şehribanoğlu & Diler, 2018).

Patel et al. (2014) tried to predict stock and stock price index movements with ML techniques. In this study, two approaches of Artificial Neural Networks (ANN), Support Vector Machine (SVM), random forests and naive Bayes were applied and the results were compared.

Bajari et al. (2015) investigated the effectiveness of demand forecasting with machine learning methods. In this article, standard econometric methods with ML methods were used.

2.3 Machine Learning Methods

There are three general machine learning algorithms. There are: supervised, unsupervised, and reinforcement.

In supervised learning, it is known what the dataset is and what the desired output should be from this data. Supervised learning feeds the data and the results from that data back to the machine. It allows a function (a match between the input data and the result data) to be extracted from this information. Thus, the machine learns the relationship between the data. Supervised learning problems are divided into two as “regression” and “classification”.

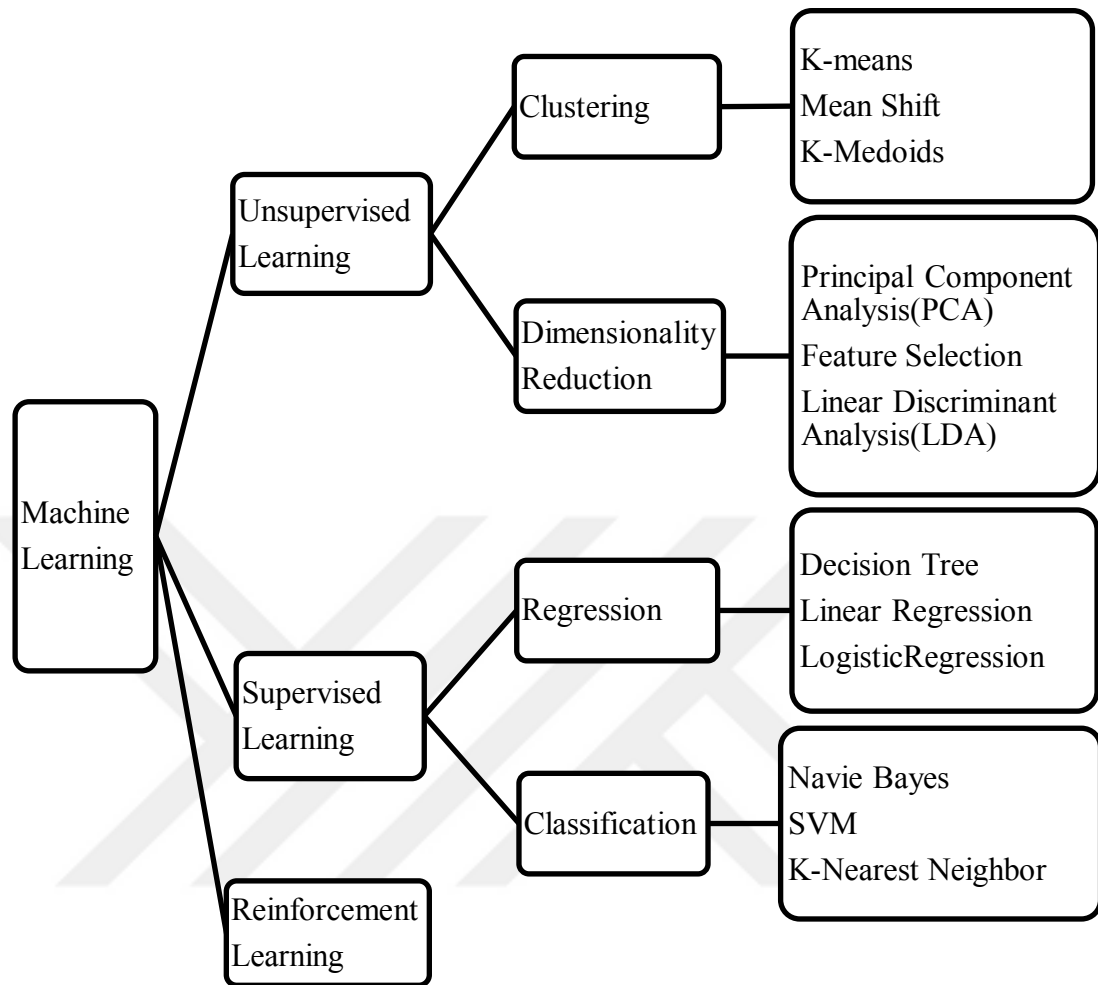


Figure 2.1 Machine learning algorithms

2.3.1 Supervised Learning

Supervised learning is a type of system in which input and output data are provided in artificial intelligence (AI) and machine learning. Input and output data are labeled for classification to provide a basis for future data processing. Labeled data assigned to samples. For example, in binary classification problems, it is spam/non-spam categories (Mohri et al., 2012). The most commonly used supervised learning algorithm to determine the class label is logistic regression. This analysis provides us with information about the probability of observations to which class the prediction belongs to (Géron, 2019)

Unsupervised learning algorithms take a data chunk containing only inputs and learn from test data whose data is unclassified or uncategorized. Rather than responding to feedback, unsupervised learning algorithms identify commonalities in the data. It reacts based on the presence or absence of such common features in each new piece of data.

Supervised learning is not just in cases where the class variable is categorical. It can also work in situations where there is a numerical variable. If the variable is numerical, trying to guess the learning variable is called regression.

The frequently used supervised learning algorithms that are frequently used:

- K-nearest neighbor algorithm (KNN)
- Linear Regression
- Logistic Regression
- Support Vector Machines
- Algorithms based on decision trees
- Artificial Neural Networks

Classification and regression are the most popular and useful machine learning methods. (Murphy, 2012). As stated earlier, classification is used to predict categories in a data. For example, the method of determining whether a picture belongs to a dog or a cat can be called classification. Likewise, the fact that the patient is healthy or sick is an example in breast cancer data. If there are two classes stipulated, it is called “binary classification”. When the number of classes is large, it is called "multiple classification". Regression does not define categories. It tries to predict a certain value. It is the classification form for continuous values (Lantz, 2015). For example, to predict the price of a product or to estimate the level of glucose in the blood can be done by regression.

2.3.1.1 Classification

Classification is a form of supervised learning. Classification is the process of dividing a given dataset into different different classes. This classification process begins with estimating the class of generated data points. Predictive modeling is done in classification. As a result of the modeling, a function is obtained from input variable to output variable. The main goal of classification is to identify which class/category the new data will fall into.

In another words, In this supervised learning method, which is most commonly used for prediction, determining which category a sample belongs to is made by classification. There are two types of classification: binomial and multi-class. Examples for classification, whether an e-mail is unnecessary, whether a person is cancer, whether a football team wins or not can be given (Lantz, 2015). This type of classification indicates binary classification. Multiple class the problems are to determine which of the multiple categories an input belongs to. Classification is a supervised learning method where output is having defined labels (discrete value). The main algorithms used in classification are decision trees, support machines, logistic regression and Naive bayes.

2.3.1.2 Regression

Another supervised learning method used for prediction is regression. Regression dataset consists of input and output variables. In other words, every time to check the algorithm has an output value in response to the observation (Elitedatascience, 2019). Here, if the desired output value consists of continuous variables, this learning problem called regression.

The main goal for regression algorithms is discrete or continuous variables is to guess. In some cases, the estimates show the linear relationship between the attributes can be used to explain. For example, if the sales increase, the budget should be increased (Polamuri, 2014).

Between input data and output data, the functional relationship is determined by using training data. When output data are known parameters of the function are

learned. However, it can also make predictions when the data of the machines are not known.

Regression is a supervised learning method where output is having continuous value. The types of regression in a supervised learning technique are linear regression, decision tree regression, random forest regression.

2.3.2 Unsupervised Learning

There is no label defining the data set in unsupervised learning. This system tries to be learned before it can change. The unsupervised learning approach is clustering, anomaly and novelty detection, and visualization. It can be used for different purposes such as size reduction, learning the association rule (Meltem, 2019).

In supervised learning, the goal is to match the input with the output. Unsupervised learning does not have such a supervisor and only input data is available. The aim is to detect differences in input. For input data a certain structure is formed and this structure is tried to be determined (Abdallh et al., 2016). Briefly, input data is not labeled and has no known results. It is prepared to extract structures in input data (Brownlee, 2016).

2.3.3 Semi-supervised learning

In the semi-supervised learning, some of the data has a class label while some of it is no information is given about which class it belongs to. Most semi-controlled algorithms. It is a combination of supervised and unsupervised learning algorithms. For example, deep belief networks (DBN) are called constrained Boltzmann machines (RBM). It is based on the unification of unsupervised learning components. Then merged the whole system is evaluated with a supervised learning approach (Meltem, 2019).

2.4. Machine Learning Steps

The purpose of machine learning is to obtain information from the data. For this reason, "data" is the most important resource for machine learning. Machine learning steps recommended in the literature were always evaluated using this data (Van Rijmenam, 2019).

Nath (2016), Brownlee (2013) and Yufeng (2017) suggested machine learning steps with different step numbers in literature. Five steps are mentioned by Nath (2016). These are data collection, data preparation, training of the model, evaluation of the model and performance improvement.

Machine learning, which is considered a sub-branch of artificial intelligence, is basically implemented in 7 steps.

Step 1: Data Collection: In this step, we collect the data we will use depending on the machine learning we will choose.

Step 2: Preparing the Data: After the data is collected, it is necessary to prepare the data for processing in order to minimize the possibility of false-positive results.

Step 3: Model Selection: In this step of machine learning, the right model is selected that is suitable for our project.

Step 4: Training: In this step, the categorization process is made according to the parameters determined in the given data and the model is learned. The training step is a phase that requires patience.

Step 5: Calculation: Once a model has been identified and trained, it needs to be tested to determine its response in real-world conditions. For this, a data set that has not been used before should be preferred. In this step, we need to use the 20% data that we reserved in step 2.

Step 6: Setting the Parameters: In this step, the model is retrained by making minor changes in the parameters and it is ensured that it can make more precise predictions.

Step 7: Estimate In this step, where machine learning comes to an end, the model becomes free from the human factor and can make predictions on its own. At this stage, the response of the model to different judgments is measured.



CHAPTER 3

LOGISTIC REGRESSION

3.1 Regression Analysis

Regression analysis is a widely used data analysis method in the field of statistical analysis. It is an excellent technique that has both descriptive and predictive purposes. The most well-known among these methods are simple and multiple linear regression methods. As it is known, it is a numerical data type obtained by dependent variable measurement in simple and multiple linear regression. Also, regression analysis is often encountered when the dependent variable is of a categorical/qualitative data type. When the dependent variable is of two or more categorical qualitative data types, logistic regression method is used (Alpar, 2013).

In regression analysis, predictions are made about the outcome variable by using predictor variable values. With the regression analysis, information can be obtained about the existence of the relationship between the variables and, if there is, the strength of this relationship. While performing the regression analysis, it is necessary to show the observation values and the affected events with a mathematical representation, that is, with the help of a function. This model is called the regression model.

Regression analysis also takes into account inference and estimation. According to Campbell & Campbell (2008), regression analysis is a statistical technique used to determine the baseline relation between two or more variables, used for inference and estimation.

These definitions and estimates are made by estimating the unknown parameters of the model. In other words, "Regression Analysis is an estimation process."

Regression models differ according to the structure of the dependent variable. For this reason, it is very important to construct the dependent variable well. Especially in recent years, the categorical dependent variable structure has been widely applied in the literature. In different regression applications, the categorical dependent variable takes the values 0 and 1 and represents a binary indicator variable. Binary

response variables are measured on a binary scale. For example, the responses may be alive or dead, or present or absent. These outcomes can be coded 0 and 1, respectively (Freund & Wilson, 1996).

3.2 Literature Reviews

There are many researchers and studies about logistic regression and these studies contain many different areas. In a study by Altunoz, he has used logistics regression and discriminant analysis, which are multivariate statistical methods, to estimate the financial failure of banks in Turkey. For this purpose, it was tried to estimate the probability of banks to fail financially one and two years in advance by using financial ratios belonging to 36 private banks. As a result of the study, both models predicted the failures experienced in the Turkish banking system at a high rate a year in advance (Altunoz, 2015).

In their studies, Akbulut and Rencber used logistic regression analysis to determine the factors affecting the activities of enterprises operating in the cement sector and the impact levels of these factors. 65.8% of the total effectiveness of the enterprises is explained with the variables of Equity Multiplier, Receivable Turnover Rate, Net Profit Margin, and Debt Rate. In their studies, it has been demonstrated that these factors determine the efficiency of businesses at a high rate (Akbulut & Rencber, 2015).

In their studies, Kaygin and his colleagues tried to estimate the financial success and failure of the enterprises by making use of the annual balance sheet and income table data of 143 manufacturing industry companies that were traded on the Istanbul Stock Market between 2010 and 2013. In the research, CHAID and C 5.0 methods, which are decision tree algorithms, and logistic regression analysis were used. Based on 2013, models for predicting one year, two years, and three years in advance were developed and the predictive power of the models was compared. In the CHAID model, 30 were classified correctly and 6 were classified incorrectly for 2012. The success rate of the model was calculated as 83.33%. In C5.0 Model, 32 were classified correctly and 4 were classified incorrectly for 2012. The success rate of the model was determined as 88.89%. In the logistic regression model, 34 were

classified correctly and 2 were incorrectly classified for 2012. The success rate of the model was calculated as 94.44%. In other words, when comparing the classification accuracy of models, the predictive power of the models, the logistic regression model was found to be more successful (Kaygin et al., 2016).

In his study, Ari tried to determine the relationship between individuals' level of hope and their thoughts about the future with the multinomial logistic regression model. To this end, it was benefited from the 2014-Life Satisfaction Survey data applied by TUIK to 7984 individuals. In the first stage of the application, the chi-square independence test was applied to many variables that are thought to be related to the level of hope included in the questionnaire and worked with 4 variables by removing the insignificant variables from the data set. The validity of the model examined in the study was found to be significant by calculating it with a likelihood estimator. Odds ratios of the variables that make up the model were obtained and 3 categories were compared according to the coefficient 22 odds ratios according to the selected reference category. Coefficient odds values for future comparison, economic, social rights, and public services provision for comparison 1, future comparison, economic and public services presentation for comparison 2, and future comparison for comparison 3 have been found to be significant (Ari, 2016)

In their studies, Mutluer and Buyukkidik used the data of Program for International Student Assessment known as PISA survey and recurring every three years since 2000. Of students in mathematics literacy using the 2012 data in Turkey research, using the selected math classes eagerly waiting, interest in mathematics, mathematics self-detection / rapid learning, perseverance, giving up quickly, making reading about mathematics, enjoying mathematics, and parents' educational status as variables, they tried to determine what level of accuracy it can classify using logistic regression technique. As a result of the analysis of the data, it was observed that the mother education level, father education level, enjoying mathematics, math self-perception / fast learning, perseverance, and giving up quickly have significant effects on the classification of 23 mathematical literacy. Whether or not reading about mathematics and the level of mathematics interest does not have a significant effect on mathematical literacy. In the obtained logistic regression model, the correct

classification rates of the observations were reported to be 85.2%. According to the research findings, 94.9% of the students who succeeded in mathematics literacy, and 54.9% of the unsuccessful students were estimated correctly (Mutluer & Buyukkidik, 2017).

In his study, Gurel used logistic regression analysis to determine the factors affecting the work-life balance of 221 female academicians working in three major state universities. In the study, participants stated that 72.4% could not establish a work-life balance, 27.6% stated that they established work-life balance. As a result of the analysis; on the presence of work-life balance from high to low risk, respectively; the number of children, the presence of spouse support, marital status, the presence of a dependent person, family support, the presence of the assistant person and age were found to be statistically significant (Gurel, 2018).

Ugurlu et al. examined the variables that cause absenteeism in students in their studies and revealed the relationship between these variables and some demographic features through logistic regression analysis. The study group of the research consists of 548 students studying at Cumhuriyet University. In the research, the “Absenteeism Tendency Scale” (ATS) was used as a data collection tool. According to the findings, it was revealed that students' GPA, weekly course hours, weekly absenteeism levels were significant on their absenteeism tendencies. While the general averages of the students decrease their tendency to absenteeism, the increase in the grade averages shows that the students' interest in their lessons increases. A one-unit increase in one-week lesson hours also increases the tendency towards attendance. It has been observed that the increase of the weekly class hours of students at school decreases their absenteeism behaviors and enables the students to attend their classes more (Ugurlu et al., 2018).

Hanimoglu examined the effect of social skill level, gender, and class level on internet addiction level in primary school second-grade students. The research was carried out on 558 students in, Çukurova and Seyhan districts of Adana, attending 6th, 7th and 8th grade, 283 (50.3%) girls, and 275 (49.7%) boys. In the study, personal information form, Internet Addiction Scale, and Matson Evaluation of Social Skills in Children were used. Internet addiction level (internet dependent, not

internet dependent) dependent variable; the logistic regression analysis was applied by taking “positive social behavior level (sub-dimension)”, “negative social behavior level (sub-dimension)” and “gender” and “class level” as independent variables. As a result, according to the results obtained from this study, it was found that students with low social skills, boys according to girls and students in the higher class are in the risk group for Internet addiction (Hanimoglu, 2019).

3.3 Introduction to Logistic Regression Model

Before explaining the logistic regression model, information should be given about the linear regression model. The simple linear regression model has only one explanatory (independent) variable. Assuming a linear relationship between the independent variable and the dependent variable, we estimate the dependent variable using the independent variable.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i=1, \dots, n \quad (3.1)$$

In this model β_0 cut point β_1 the slope of the line ε error term. Since $E\{\varepsilon_i\} = 0$, the probability of the random variable. $E\{\varepsilon_i\} = 0$ can be written as $E\{Y_i\} = \beta_0 + \beta_1 X_i$. Since the Y_i random variable Bernoulli is distributed, the expected value of Y_i is modeled on the probability of success.

When Y_i is a Bernoulli random variable, then its the probability distribution can be written as follows; $P(Y_i = 1) = \pi_i$ for $Y_i = 1$, $P(Y_i = 0) = 1 - \pi_i$ for $Y_i = 0$

Expected value of a Bernoulli random variable is

$$E\{Y_i\} = 0(1 - \pi_i) + 1(\pi_i) = \pi_i \quad (3.2)$$

So $E\{Y_i\}$ is found as follows;

$$E\{Y_i\} = \beta_0 + \beta_1 X_i = \pi_i \quad (3.3)$$

As it is understood from the expression above, it is seen that the conditional expected value of the model is equal to the conditional probability of $E\{Y_i | x\}$. Since a probability value will not be less than 0 and not greater than 1, the conditional expected value of the model must be between 1 and 0.

$$0 \leq E(Y_i) = \pi_i \leq 1 \quad (3.4)$$

The change in the $E\{Y_i | x\}$ per unit change in x becomes smaller as the conditional mean gets closer to 0 or 1. In the logistic regression, the relationship between the independent (predictor) and response variable is not a linear function. This can be seen in Figure 3.1 and this shape of the curve is said to be S shaped and resembles the plot of a cumulative distribution function of a random variable.

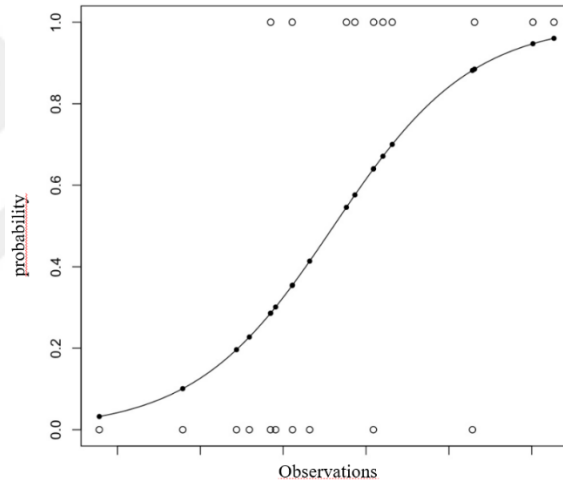


Figure 3.1 Logistic regression plot (S Curve) (Daishizuka, 2015)

In case of various hypothetical distortions, Logistic Regression Analysis is applied as an alternative to discriminant analysis and cross-tabulation applications. The most basic approach as a reason for use can be made in linear regression analysis; when the dependent variable is binary such as 0 and 1 or discrete variable containing more than two categories, normality assumption is broken and linear regression analysis can not be applied.

These assumptions are summarized below.

- Prediction errors are not normally distributed.
- Heteroscedasticity is a problem.
- The assumption that the probability value $E\{Y_i | x\}$ increases in parallel with the values of the X variable.
- The R^2 value, which is an indicator of goodness of fit, becomes small and loses its indicator quality.
- Failure to fulfill the condition $0 \leq E\{Y_i | x\} \leq 1$

In logistic regression, a complex formula should be used and transition from logistic equation to Ordinary Least Squares (OLS) type equation. The logistic formulas are stated in terms of the probability that $Y = 1$, which is referred to as P. The probability that Y is 0 is $1 - P$.

$$\pi(x_i) = E(Y | x) = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} \quad (3.5)$$

$$g(x_i) = \ln\left[\frac{\pi(x_i)}{1 - \pi(x_i)}\right] = \ln\left[\frac{E(Y = 1 | x_i)}{E(Y = 0 | x_i)}\right] \quad (3.6)$$

$$1 - \pi(x_i) = 1 - \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} = \frac{1}{1 + \exp(\beta_0 + \beta_1 X_i)} \quad (3.7)$$

$$g(x_i) = \ln\left[\frac{\frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}}{\frac{1}{1 + \exp(\beta_0 + \beta_1 X_i)}}\right] = \ln(\exp(\beta_0 + \beta_1 X_i)) = \beta_0 + \beta_1 X_i \quad (3.8)$$

Here, the ratio $\frac{\pi(x_i)}{1 - \pi(x_i)}$ in the logit transformation is called odds.

3.4 Fitting of Simple Logistic Regression Model

People who deal with observation, measurement and experimental problems and who want to reveal something new with the data they get from nature have a purpose. Observation is not measuring and experimenting, but interpreting the

results. Usually, after measurement, observation, and experimentation, thousands of results are produced that contain at least two interdependent variables. All of these results have an error rate. Various methods have been proposed to estimate the parameters in the logistic model. In general, the maximum likelihood method maximizes the probability of obtaining the observed dataset. In this way, the values of unknown parameters are found. In order to apply this method, the maximum likelihood function is created. The maximum likelihood estimators of the parameters are chosen to find the values that maximize the function. Thus, the estimators obtained are very close to the observed data.

Estimation methods in logistic regression are examined in three different concepts. These are given below,

1. The Maximum Likelihood Method.
2. The Least Squares Method
3. The Minimum Logit Chi-Square Method.

In this thesis, maximum likelihood of estimating the parameters (maximum likelihood) estimation method will be used.

3.4.1 Likelihood Function and Maximum Likelihood Estimation Method

If Y is coded as 0 and 1, the expression $\pi(x)$ gives the conditional probability that y equals 1 for the given value of x . This probability is represented by the symbol $\pi(x) = P(Y = 1/x)$. Hence, the expression $[1 - \pi(x)]$ also shows the conditional probability of taking the 0 value of Y . This probability is shown as $[1 - \pi(x)] = P(Y = 0/x)$. For the pair (x_i, y_i) the contribution to the likelihood function when $y_i = 1$ is $\pi(x_i)$, whereas when $y_i = 0$ the contribution to the likelihood function is $1 - \pi(x_i)$.

A reliable way of expressing the contribution of the pair (x_i, y_i) to the likelihood function is given below:

$$P(Y = y_i) = f_i(y_i) = \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i} \quad (3.9)$$

where $y_i = 0$ or $y_i = 1$ for $i=1,2,\dots,n$

The maximum likelihood estimation method is used to calculate the logit coefficients. This method returns values for unknown parameters that maximize the probability of obtaining the observed dataset(VUPA,2004).

Since observations are assumed to be independent from each other, the likelihood function is obtained by multiplying terms.

$$L(\beta_0, \beta_1) = g(y_1, y_2, \dots, y_i) = \prod_{i=1}^n f_i(y_i) = \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i} \quad (3.10)$$

In the basic principle of maximum likelihood, it is emphasized that the estimate of maximizes the expression. Since it will be easier to work with logarithms mathematically, the log-likelihood function is obtained as follows:

$$\begin{aligned} L(\beta_0, \beta_1) &= \ln \prod_{i=1}^n \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i} \\ &= \sum_{i=1}^n \{ \ln[\pi(x_i)^{y_i}] + \ln[(1 - \pi(x_i))^{1-y_i}] \} \\ &= \sum_{i=1}^n \{ \ln[\pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i}] \} \\ &= \sum_{i=1}^n \{ y_i \ln[\pi(x_i)] + (1 - y_i) \ln[(1 - \pi(x_i))] \} \\ &= \sum_{i=1}^n \left\{ y_i \ln \left[\frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} \right] + (1 - y_i) \ln \left[1 - \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} \right] \right\} \end{aligned}$$

Hence

$$L(\beta_0, \beta_1) = \sum_{i=1}^n \{ y_i (\beta_0 + \beta_1 X_i) - \ln(1 + \exp(\beta_0 + \beta_1 X_i)) \} \quad (3.11)$$

In order to find the values of that make $L(\beta_0, \beta_1)$ 'maximum, the derivative of $L(\beta_0, \beta_1)$ according to β_0 and β_1 is taken and equalized to zero.

$$\frac{\partial \ln L(\beta_0, \beta_1)}{\partial \beta_1} = \sum_{i=1}^n \left\{ y_i - \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} \right\} = 0 \quad (3.12)$$

The Maximum Likelihood function is an iterative algorithm and this procedure is complex.

3.4.2 Testing for the Significance of the Coefficients

The coefficients have been estimated, and then the importance of the variable needs to be determined in the model formed.

In order to determine whether the independent variable in the model is related to the response variable, some statistical hypotheses are determined and these hypotheses are tested. In order to test the significance of the coefficient of a variable in the model, it is checked whether the model containing this variable provides more information about the response variable than the model without this variable. Control is achieved by comparing the values predicted by each of the two models. Three tests are generally used for hypothesis testing in the logistic regression model.

These are likelihood ratio test, wald test and score test.

3.4.2.1 Likelihood Ratio Test

In logistic regression, the comparison of observed and expected values fulfills the log probability.

$$D = -2 \log \left[\frac{\text{likelihood of the null model}}{\text{likelihood of the given model}} \right] \quad (3.15)$$

The expression likelihood ratio given in parentheses in the equation is called “likelihood ratio”. Taking the $(-2\ln)$ floor is to obtain a value whose distribution is known as much as mathematical. This value is used for hypothesis testing. Such a test is called likelihood ratio test.

In order to decide the importance of an independent variable, D values in the equation with and without an independent variable are compared. The change in D that occurs due to covering the independent variable is as follows

$$G = D(\text{model without variable}) - D(\text{model with variable}) \quad (3.16)$$

This calculated statistic also plays the same role as the part of the F test used in linear regression. Since the likelihood of the saturated model is common for both of the D values to be taken to calculate G, the G statistic takes the following form:

$$G = -2\ln[\text{likelihood without the variable} / \text{likelihood with the variable}] \quad (3.17)$$

Under the hypothesis $\beta_1 = 0$ the G statistic shows the χ^2 distribution with 1 degree of freedom. The validity of the model, which is based on the idea that the criteria based on the difference of the likelihood ratio values for the model estimated by the model containing all the variables will be distributed, is tested (Elhan, 1997).

3.4.2.2 Wald Test

The Wald test (named after Abraham Wald) is an unbounded estimate and under-hypothetical value based on the weighted distance between the statistical parameters related to the determination constraints. The null hypothesis is weight, it is estimated to remain sensitive.

The highest probability estimator of the slope parameter is compared to the estimated value of the standard error. When $\beta_1 = 0$, the distribution of the test statistics complies with the standard normal distribution. Test statistics for this test,

$$W = \frac{\beta_1}{\text{se}(\beta_1)} \quad (3.18)$$

under the hypothesis that $\beta_1 = 0$, W is a standard normal distribution. This test also can be written in an alternative manner. So, the Wald test statistic is written as follows:

$$W^2 = \left(\frac{\beta_1}{\text{se}(\beta_1)} \right)^2 \quad (3.19)$$

Where $W^2 \sim \chi^2(1-\alpha, 1)$. According to this equation, the p value is decided in the form of "redeemed hypothesis" in which the given α is smaller than the given.

3.4.2.3 Score Test

Another test used in testing the significance of parameters is the score test. This test is constructed according to the conditional distributions of the derivatives of the likelihood equations. In this way, the following formula can be obtained:

$$ST = \frac{\sum_{i=1}^n x_i (y_i - \bar{y})}{\sqrt{\bar{y}(1-\bar{y}) \sum_{i=1}^n (x_i - \bar{x})^2}} \quad (3.20)$$

This statistic shows a standard normal distribution for the $\beta_1 = 0$ hypothesis, which shows the slope parameter.

3.5 Fitting of Multiple Logistic Regression Model

There are more than one independent variable in multivariate logistic regression. Here, the independent variable $\tilde{x} = [1, x_1, x_2, \dots, x_p]$ takes the value of 0 or 1 in the regression parameters $\tilde{\beta} = [1, \beta_1, \beta_2, \dots, \beta_p]$ and the dependent variable. Also, the expected value of errors is $E\{\varepsilon_i\} = 0$. Logistic response function for multivariate case;

$$\pi(x_i) = E(Y = 1 | x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \quad (3.21)$$

$$\ln \left[\frac{\pi(x_i)}{1 - \pi(x_i)} \right] = \ln \left[\frac{E(Y = 1 | x_i)}{E(Y = 0 | x_i)} \right] = \ln(e^{\beta_0 + \beta_1 x_i}) = \beta_0 + \beta_1 x_i \quad (3.22)$$

3.5.1 Design Variable

If some arguments have more than two levels as categorical variables, it is appropriate to construct the model differently from the general formula in the logit transform. If the number of variable categories equals k , the design variables $k-1$ should be obtained. The representation showing the design variables is not the same as the logistic regression model. That is, let's assume that the independent variable has levels. Design variables are needed. The coefficients of these design variables are $m = 1, 2, \dots, k, k_{j-1}$ (Vupa, 2004).

The logit for a model with p independent variables and the j th independent variable being discrete is expressed as:

$$g(\bar{x}) = \beta_0 + \beta_1 x_1 + \dots + \sum_{m=1}^{kj-1} \beta_{jm} + D_{jm} + \beta_p x_p \quad (3.23)$$

3.5.2 Testing for the Significance of the Coefficients

3.5.2.1 Likelihood Ratio Test

Parameters in the multitude are reconstructed by the MLE method because Y is still a Bernoulli random variable with the same probability distribution. The generation of maximum likelihood estimators remains the same with the expectation that more parameters are included. The log-likelihood equation takes the form as follows:

$$A = \sum_{i=1}^n \{y_i (\beta_0 + \beta_1 X_{li} + \dots + \beta_1 X_{pi})\} \quad B = \ln(1 + \exp(\beta_0 + \beta_1 X_{li} + \dots + \beta_1 X_{pi}))\}$$
$$\ln L(\beta_0, \beta_1, \dots, \beta_p) = A - B \quad (3.24)$$

The log-probability equation is derived for each parameter. Each derivative is set to zero. The resulting equations are solved to obtain the estimates. The likelihood ratio test is used for the overall significance of the p coefficients for the independent variables in the model. This test consists of G statistics. It is compared with the log-likelihood of the model without the variable(s) to determine the significance of the model (Hosmer & Lemeshow, 1989). The test statistic, G , is calculated as follows:

$$G = -2 \ln[\text{likelihood without the variable} / \text{likelihood with the variable}] \quad (3.25)$$

When the significance of the model is to be checked, null and alternative hypotheses should be established. These hypotheses are given below.

$$\begin{aligned} H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0 \\ H_1 : \text{At least one of the } \beta_p \neq 0 \end{aligned}$$

G statistic can be used for testing all coefficients, and the slope parameter has a chi-square distribution with (V_2, V_1) degree of freedom in the test of the null hypothesis. V_2 is equal to one more than the number of variables in the full model. The decision rule for this test is that the p value $p\{X^2_{1-\alpha, df(V_2, V_1)}\}$.

If the p-value is less than a determined statistical significance (α), the null hypothesis (H_0) is rejected. Because the result obtained is statistically significant at the determined level.

3.5.2.2 Wald Test

For individual model coefficients, tests of hypotheses such as

$$\begin{aligned} H_0 : \beta_p = 0 \\ H_1 : \beta_p \neq 0 \end{aligned}$$

can be tested by looking at the deviation statistics.

There is also an approach based on probability estimation. For large samples, the distribution of the most likely estimator approaches the normal distribution, with or without a small side.

At the point or points that make the first derivatives of the function zero, the function has a maximum or a minimum. Second derivatives are checked to see if these points are the maximum.

A statistic similar to the t statistic can be used for This test statistic is known as the Wald (1943) test.

$$W = \frac{\beta_j}{se(\beta_j)} \quad (3.26)$$

This statistic is called the Wald test statistic. Standard normal distribution is used to interpret this statistic (Hosmer & Lemeshow, 2000). Some computer software programs take the square of the W^2 and compare it with the 1 degree of freedom chi-square distribution.

3.5.3 Interpretation of the Coefficients

The β_j regression coefficient estimated in the logistic response function is not as easy to interpret as in a linear regression model.

Odds ratio is a measure used in statistical analysis and especially in biostatistics studies to determine the effect of risk factors on outcome variables. Odds means probability as a concept.

$$\text{odds of \{Event\}} = \frac{\pi(x_i)}{1 - \pi(x_i)} \quad (3.27)$$

It is very difficult to measure the effect of a one-unit increase in the X variable in the logistic regression model created according to the starting point on the X axis. When interpreting the coefficient, it is multiplied by the probability estimate for a one-unit increase in X and the resulting logistic response function should be applied. The purpose of the logistics regression is to define the relationship between the result variables and the type of logit. The natural logarithm of the possibilities. Y is a binary result variable classified as "1" and "0". X is a continuous variable. If a distribution chart is drawn, the two parallel dashes corresponding to each result variable category are formed. This relationship does not have a linear relationship. For this reason, logistics regression is made of logit conversion on the result variable y. The simplest form of the logistics regression model can be written as:

$$LF(\pi(x_i)) = \ln \left[\frac{\pi(x_i)}{1 - \pi(x_i)} \right] = \ln \left[\frac{E(Y = 1|x_i)}{E(Y = 0|x_i)} \right] = \ln(e^{\beta_0 + \beta_1 x_i}) = \beta_0 + \beta_1 x_i \quad (3.28)$$

It is the ratio of the odds ratio of an event for one group to the odds ratio for another group; or an estimate based on a sample. These two groups may be, for example, male and female groups, experimental and control groups, or any other two-part classification. Let the probability of an event be p (A group) and q (B group) for these groups.

$$\text{odds ratio}[A \text{ vs. } B] = \frac{\text{odds}[A]}{\text{odds}[B]} = \frac{\frac{P_a}{1-P_a}}{\frac{P_b}{1-P_b}} \quad (3.29)$$

An odds ratio of 1 means that the event or situation in question is equally likely for both groups. If the situation is equal for both groups, it means that the odds ratio value is 1. If the odds ratio is greater than 1, it indicates that exposure to the factor is responsible for the occurrence of the cases. If the odds ratio is less than 1, it indicates a relationship between the occurrence of cases and exposure to the agent.

Interpreted as $OR=1$ indicates that it does not affect the probability of outcome, $OR>1$ indicates exposure associated with a higher probability of outcome, $OR<1$ indicates exposure associated with a lower probability of outcome (Park, 2013).

We can use the odds ratio with a 95% confidence interval to evaluate the contribution of each estimator (Katz, 1999) We can learn whether the 95% confidence interval indicates statistical significance, unlike the p value. If it does not overlap with the null value ($OR=1$), for the existence of statistical significance can be used as a proxy.

The 95% CI is used to estimate the precision of the odds ratio. If the confidence interval is large, it determines that the odds ratio has a low precision. When the CI is small, the precision of the odds ratio is higher.

An approximate confidence interval for the population log odds ratio is 95% confidence interval for the;

$$\ln(OR) = \ln(OR) \pm 1.96\{SE \ln(OR)\} \quad (3.30)$$

$\ln(\text{OR})$ is the sample log odds ratio. $\text{SE } \ln(\text{OR})$ is the standard error of the log probability ratio (Morris & Gardner, 1988). Antilog is taken. A 95% confidence interval is established:

$$95\% \text{CI for OR} = \exp(\ln(\text{OR}) \pm 1.96\{\text{SE } \ln(\text{OR})\}) \quad (3.31)$$

3.6 Comparison of Model Selection Criteria

The expected values of the observation values are found by the logistic regression model. Finding the expected values is done with the help of probabilities in the logistic regression model with an appropriate cut-off point. After the classification is made between the observed and expected values, the accuracy of this classification is obtained by looking at the classification criteria. These classification criteria are: sensitivity, specificity, precision, accuracy, F-score.

Table 3.1 Number of subjects classified by two classifications

		Predicted	
		0	1
Observed	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

- True Positives (TP): These are examples where the true value is 1 and the predicted value is 1.
- True Negatives (TN): These are instances where the true value is 0 and our predicted value is 0.
- False Positives (FP): These are instances where the true value is 0 but the predicted value is 1.
- False Negatives (FN): These are instances where the true value is 1 but our predicted value is 0.

There are some ratios calculated from the confusion matrix. These terminologies showing the relationship between TP, TN, FP and FN are given below.

3.6.1 Classification Criteria

Accuracy: It is a measure of how often the classifier guesses correctly

$$((n_{00} + n_{11}) / (n_{00} + n_{01} + n_{10} + n_{11})).$$

False Negative Rate: It is the rate of those predicted as 0 even though the true value is 1. It is also known as the loss rate. (Miss Rate)

False Positive Rate: It is the ratio of those predicted as 1 even though the true value is 0. Also known as By-Product. (Fall out)

Prevalence: It is a measure of how often a value of 1 is found at the end of the estimation.

Sensitivity (Positive Predictive Value) It is a measure of how accurately predicted from all classes. It should be as high as possible. Also known as Positive Predictive Value. (Positive Predictive Value) $((n_{11}) / (n_{10} + n_{11})).$

Specificity (True Negative Rate): It is a measure of how accurately the classifier has predicted the true negative value. $((n_{00}) / (n_{00} + n_{01})).$

F1-Scores: F1-Score is a measure combining both precision and recall. It is described as the harmonic mean of the two $2(\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$

Cohen's Kappa: A measure of how well the classifier actually performs.

CHAPTER 4

DECISION TREE

4.1 Decision Tree

The Decision Tree consists of classification, decision nodes and leaf nodes. It is a classification method that establishes a model in the form of a tree structure. Decision trees can be written in "if condition then result" text format (Gerritsen, 2020). Decision trees are widely used in classification problems. Decision trees are easier to shape, understand and interpret than other methods. In the decision tree method, a tree is created from the data we have in order to classify. The records in the dataset are applied to this tree. Classification of records takes place according to the result. In other words, when we have a data that we do not know which class belongs to the decision trees obtained from the database, the class is estimated according to the created rule set (Silahtaroglu, 2013). According to Vahaplar, decision trees are easy to set up and interpret. Decision trees can be easily obtained with database systems and can be preferred due to their high reliability (Vahaplar, 2003).

Decision tree has different definitions by many researchers. According to Frostman, the process of creating tree-shaped decision structures using the induction method from the education data set, whose classes we know, is called decision trees (Frostmacı, 2011). According to Kayri, decision trees are a method that helps to determine which factors are effective and to reveal the past relationships of these factors (Kayri, 2014).

The partition systems of decision trees have also been defined by different researchers. According to Berry & Linoff, the decision tree model divides large heterogeneous groups into homogeneous subgroups according to the target variable, that is, it creates a root-to-leaf rule set (Berry & Linoff, 2004). According to Frostman, in decision trees, trees form large-volume data sets by dividing smaller subgroups into decision-making rules. Thus, the members of the groups obtained by correct decomposition processes are more similar to each other.

Structure, steps and interpretation of decision trees have been explained by different researchers. The structure of decision trees is tree (root, branches, leaves). Decision trees begin with the root, which covers all observations in the data, and as it goes down, it is divided into branches that divide the data into subgroups. In this tree structure growing from this root to the branches, each knuckle is called the "knot" (Pehlivan, 2006). The first step in the classification of data using the decision tree method is the learning step. Here, a model is established using a classification algorithm in a previously known educational data. The learned model is represented as a decision tree.

In the classification step, the test data is used to determine the accuracy of the classification rules or decision tree.

It should be determined which fields in the training data will be used and in which order the tree will be created. The most commonly used measurement for this purpose is the Entropy measurement. The greater the entropy measure, the more uncertain and unstable the results obtained using that area.

Decision trees are used in many areas. These areas are: (1) Defining some elements that can be members of any class, (2) Separation of various cases into some risk groups, (3) Identification of associations specific to a particular group, (4) Making credit decisions using people's credit histories (Credit Scoring), (5) Analyzing production data to identify variables that cause product defects.

Many algorithms are used to create a decision tree. These; "CHAID (Chi-Squared Automatic Interaction Detector), Exhaustive CHAID, CART (Classification and Regression Trees), ID3, C4.5, MARS (Multivariate Adaptive Regression Splines), QUEST (Quick, Unbiased, Efficient Statistical Tree), C5.0, SLIQ (Supervised Learning in Quest)" can be listed as SPRINT (Scalable Parallelizable Induction of Decision Trees).

4.1.1 Structure of the Decision Tree

Decision tree in general; There are three sections: "node, branch and leaf". Each variable in the tree structure delivers to another node in the tree structure. In the decision tree, the upper part is called the root node and the lower part is called the leaf node. The division between root and leaf nodes is referred to as branches. The decision tree consists of paths starting from the root node and continuing according to the yes-no answers applied to the experimental units (Han & Kamber, 2000; Lewis, 2000; Pehlivan, 2006).

In decision trees, the depth in the mountain structure is as important as the nodes, branches and leaves. Each branch in the tree structure is a function that creates a rule. If the classification is not completed during the new branch creation, a new decision node is created. The number of decision nodes formed as a result of branching is expressed in depth. The number of depths depends on the size and homogeneity of the data set (Carvalho & Freitas, 2004). The figure between root node, leaf node and internal node is given below.

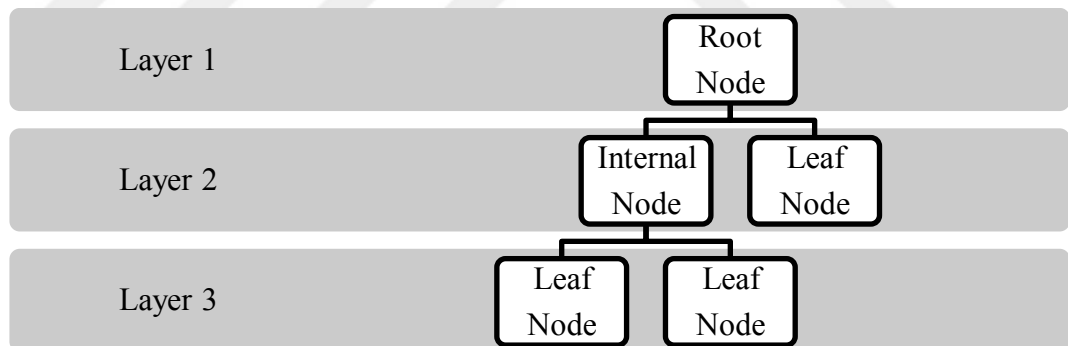


Figure 4.1 Root, leaf and terminal nodes

4.1.2 Create a Decision Tree

Since the shape of the tree created according to the algorithm used may change, the algorithm used when creating decision trees is important. The structure of the root knot, which is the first knot, affects the way and classification as it reaches the leaves at the tip. The main criterion for determining both the root node and each subsequent node is whether the remaining data is branched from similar size branches when the node is located. The response variable in the data set; if yes-no, the knot is divided into two equal parts; If yes, no and unstable, it has three categories, the node is divided into three parts (Lewis, 2000; Han & Kamber, 2000).

In the decision tree model created to classify the test data, starting from the root node, this test condition is applied to the data and appropriate branches are created. From these branches, the process is completed by reaching another branch node or leaf node. Thus, which class of this test data will be included is determined by looking at the leaf node where it ends. The most important step in the creation of decision trees is to determine which criterion or which variable will be made for the branching that will provide the classification of the variables in the data set. At this stage, the variable with the highest uncertainty is determined and used for testing at the root node of the tree. There are different approaches in the developed literature to determine this. The most important approaches are; Based on entropy, information gain and information gain rate (Quinlan, 1987; 1990; 2014) Twoing rule (Breiman et al., 1984), Gini criterion (Breiman et al., 1984) and Chi-square probability (Mingers, 1989) is table statistics (Han & Kamber 2000).

The Entropy result used in CART, ID3 and C4.5 algorithms takes values between 0 and 1 and indicates that uncertainty increases as you approach 1, and uncertainty decreases as you approach zero (0). If all the observations in the data set studied are for only one class, there is no uncertainty and the entropy result will be 0. If the probability of a condition is close to $p = 1$, there is a strong uncertainty in the realization of this condition. In decision tree algorithm, the aim is to try to reduce the entropy of the single node (homogeneous) leaf node to zero (Breiman et al., 1984; Quinlan, 1987; 1990).

D is the data set with class probability distribution ; P_i , i in the D data set. the class is probability and this value is i. The entropy of D is as in Equality (4.1) (Quinlan, 1987; 1990).

$$\text{Ent}(D) = -\sum p_i \cdot \log_2(p_i) \quad (4.1)$$

The following example (Dunham, 2003) can be given for simple Entropy calculation. D =a data set with elements $\{\square, \square, \Delta, \Delta, \Delta, \Delta, \Delta, \Delta\}$; There are 2 $\square$ and 6 Δ . If the probabilities are calculated. There are 0.25 (2/8) for $\square$ and 0.75 (6/8) for Δ . The entropy calculation of this example is as follows;

$$\text{Ent}(D) = -2/8 \log_2(2/8) - 6/8 \log_2(6/8) = 0.1887 \quad (4.2)$$

Since the entropy is close to 0, uncertainty is low in the realization of the event.

There is a relationship between decision trees and regression. Decision trees; The single response is similar to regression models in that it contains variable and multiple explanatory variables, and when the response variable is categorical, it discovers useful patterns that can be an alternative to logistic regression analysis (Han & Kamber, 2006).

So decision trees run on both discrete and continuous data. Continuous data in the decision tree can take place in more than one layer (in the lower node) and create branching according to different binary values. There is no specific limitation for this. Categorical variables also do not use the previously used feature to further divide any branch of the tree. However, the most important thing to note here is that the same feature can be used at a different level in a different branch of the tree (Stackexchange, 2018).

At the same time, decision trees are included in nonparametric methods. In decision trees, a model is created according to the input and a large data set is required for modeling (Baykal, 2006; Köktürk, 2012).

4.1.3 Accuracy of Classification Process (Train and Test Data)

The data obtained as a result of the data preparation stage and to be used in the analysis is needed at this stage. However, not all data is used in the training of the tree. The data is divided into trial and learning data. This separation is done to measure the accuracy of the generated tree. In some cases, the dataset may not be large enough to split, in such cases there are methods to test accuracy. If the accuracy at the end of the operations is satisfactory, a decision tree can be used to classify the data. Otherwise, the created decision tree rules cannot be used in the estimation process.

4.1.4 Decision Trees with Advantages and Disadvantages

There are some advantages and disadvantages encountered in the use of decision trees. Some advantages of decision tree based classification methods are listed as (Murthy, 1998; Mitra 2003; Oğuzlar 2004);

- It is more useful because it is easy to understand and interpret.
- It can be easily applied to very large datasets.
- Tree models are researchers as opposed to inferential models.
- Decision trees are non-parametric. As with other techniques, it does not deal with statistical assumptions.
- Hierarchical decomposition reveals the relationship direction and order of importance of its existing features.
- Due to the tree structure, mixed connections can be revealed.
- Can work with continuous and categorical qualifications.

Some disadvantages of decision trees are listed as (Myatt, 2007; Öğüdücü, 2008);

- Computing costs can be more costly than other classification methods, especially if large data sets have a large number of continuous variables.
- Finding the most suitable tree can sometimes be difficult. Complex and large trees can be produced while trying to create a useful decision tree.
- The tree may not catch enough information when the number of inputs is low.
- Some algorithms work only with categorical data and may not be successful in the classification of continuous variables.

- If the number of data and variables of the learning set is large or mixed, the time and tree complexities change.

4.2 Decision Tree Algorithms

Within the decision tree methods; Although there are different algorithms based on the logic of “classifying the answer variable according to the explanatory variables” in general terms, each of these methods serves different purposes. The first applications that form the basis of the tree-based methods were made with the algorithm of AID (automatic interaction detector; Morgan & Sonquist, 1963), and then additional algorithms were developed. These developed algorithms can be listed as follows.

- CART (classification and regression trees; Breiman et al., 1984)
- CHAID (Chi-squared automatic interaction detector; Kass, 1980)
- Exhaustive CHAID (Biggs et al., 1991)
- QUEST (quick, unbiased, efficient statistical tree; Loh & Shih, 1997)
- Random Forest (Random Forest; Breiman, 2001)
- C4.5 (successor of ID3) decision tree (Quinlan, 1993),
- C5 (successor of ID3) decision tree (Quinlan, 1993)
- ID3 tree (iterative dichotomiser 3; Quinlan, 1986)
- J48 tree (C4.5 implementation in Weka; Witten & Frank 2000)
- CAL5 decision tree algorithm (Michie et al., 1994)
- Rotation Tree (rotation forest; Rodriguez et al. ,2006)
- MARS (multivariate adaptive regression splines; Friedman, 1991)
- Increased trees (Friedman, 2002)
- SLIQ (supervised learning in QUEST, Shafer et al., 1996)
- SPRINT (scalable parallelizable induction of decision trees; Shafer et al., 1996)

4.2.1 CART Algorithm

In studies conducted in many fields, data are evaluated categorically. In the decision-making process, the dependent variable, which is categorical, is estimated by classification methods, taking into account different variables. There are some difficulties in dealing with statistical methods in such classification problems. These difficulties can be listed as follows.

- The large number of independent variables that make variable selection difficult,
- The dependent variable is not normally distributed,
- Complex interactions and patterns between variables
- It is difficult to use the results of traditional methods

Especially the second and fourth items are very important. In the fourth item you will find the result (value or probability). Sometimes this result cannot be interpreted or does not provide much information. However, finding the ratios between groups may give more meaningful results.

The CART algorithm was introduced by Breiman et al. in 1984. It is a continuation of Morgan and Sonquist's decision tree algorithm called AID (Automatic Interaction Detection). It uses both numeric and nominal data types. This algorithm is suitable as a solution to classification and regression problems.

4.2.1.1 Literature Review

American Express applied the decision tree algorithm in the process of deciding which loans to apply. (Michie, 1989).

Evans, B. & Fisher (1994) decided that the decision tree method was appropriate to determine the conditions in which banding occurs and does not occur in Gallatin.

Waheed, Bonnell, Prasher and Paulet (2006) applied the CART decision tree algorithm in agriculture. They investigated which causes have an effect on crop growth.

According to Abellan et al. (2013) applied decision trees according to different section criteria in their study examining the severity of injuries on rural roads in Spain. The next year, Kashani et al. (2014) applied the decision tree method to investigate the reasons that affect the severity of motorcycle accidents. Apart from these studies, decision trees have application areas in the following areas:

- Stock trading (Muh-Cherng Wu, Sheng-Yu Lin and Chia- Hsin Lin, 2006)
- Manufacturing (Bo-Suk Yang, Dong Soo Lin, Andy Chit Chiow Tan, 2005)
- Medical diagnosis (Kıran, 2010)
- Meteorology (David, Michael, Karen, Steve & Kevin, 2000)
- Geology, astronomy, physics (Kayri & Ceyhan, 2016).
- Economy (Tae Hee Moon and So Young Sohn, 2005)
- Education (Ceylan, 2014)
- E-Commerce (Tai-Chang Hsia, An-Jin Shie & Li-Chen Chen, 2008)
- Ecology (S. Caetanoa, J Aires-de-Sousab, M Daszykowskia, & Y.Vander Heydena, 2005).
- Analytical chemistry and drug discovery (S. Caetanoa, J Aires-de-Sousab, M Daszykowskia, & Y.Vander Heydena, 2005).
- Analysis of real estate mortgage data (David & Shulamith, 2005).

4.2.1.2 Basic Steps of CART Algorithm

CART analysis consists of four basic steps. These are ordered as (Lewis, 2000).

Tree Creation

The first step is to create a tree structure. It is processing the data in the dataset into this tree. The structure of decision trees consists of root nodes, nodes and branches. While creating the rules, questions are asked and actions are taken according to the answers. By combining the answers, new rules are created. After deciding on the first variable to ask the question, the root node, which is the first stage of tree structure, is created. The root node represents the dependent variable. As you go down the tree structure that starts with the root node, the data sets are divided into small pieces and branches and new nodes are created. Each node represents a question and is divided into as many branches as the number of answers

to be given to the question. In each of the branches, 'if-then' rules are formed and according to the selected decision, the next node is passed. The tree structure is processed in this way until a new question occurs. Finally, the last node representing a class is reached.

The most representative tree of root nodes and parent nodes should be created. The name of the technique used to determine this is the "qualitative selection measure". By using these selection criteria, we can arrive at the best qualifications. There are two popular techniques for attribute selection measure, which are:

- Information Gain
- Gini Index

There are many researchers studying on attribute selection measure. According to Breiman et al, the most important step in the creation of the CART decision tree is to determine which criterion or which variable will be made for the branching that will allow the classification of the variables in the data set. At this stage, the variable with the lowest uncertainty is processed and used for testing in the root node. There are different approaches to determine this, developed in the literature. However, "information gain" and "information gain rate", which involves the "entropy" process, are commonly used to determine which variable to branch out (Breiman et al., 1984).

Entropy has been used in the following studies:

- A Spatial Entropy-Based Decision Tree for Classification of Geographical Information, Xiang Li Naval & Christophe Claramunt, 2006.
- Du Ming, Wang Shu-mei & Gong Gu, Research on Decision Tree Algorithm Based on Information Entropy, 2011.
- Mai Shouman, Tim Turner & Rob Stocker, Using Decision Tree for Diagnosing Heart Disease Patients, 2011.

However, different researchers used different separation criteria. Akpınar used binary division based on the gini index in the CART algorithm. In this technique, which uses the gini index as the branching criterion; Each end node has two branches (Akpınar, 2000).

Information Gain: Information gain is measured for each attribute to determine the most distinctive feature. Entropy is used to measure Information Gain. The node is split according to the value of the information gain. Thus, the decision tree is created. The purpose of the decision tree is to try to maximize the information gain. Simply, the information gain formula is as follows:

$$\text{Ent}(D) = -\sum p_i \cdot \log_2(p_i) \quad (4.3)$$

Gini Index: This index is a measure of purity used when creating a decision tree with CART algorithm. The high quality index should be preferred according to the value that is high. In the gini index, binary compartments are created. A feature with a low Gini index should be preferred. The Gini index works either pass or fail for the categorical target variable. The Gini directory must be used for these binary compartments. Simply, the gini index is as follows:

$$\text{Gini}(D) = 1 - \sum (p_i)^2 \quad (4.4)$$

Consequently the evaluation criterion for selecting the attribute a_i is defined as:

$$\text{GiniGain}(a_i, S) = \text{Gini}(D) - \sum (p_i) \cdot \text{Gini}(D) \quad (4.5)$$

Gini index has been used in the following studies:

- Güner Z., Cart and Lojistic Regression Analysis in Data Mining: An Application on Pharmacy Provision System Data, 2014.
- Ghiasi M, Zendeboudi S & Mohsenipour A., Decision tree-based diagnosis of coronary artery disease: CART model, 2020.

Finally, the recursive split of the Nodes is used and consists of the formation of the decision tree from which a tree is built.

Stopping the Tree Creation

In the CART algorithm, it is controlled by the "stop rules" whether the tree growing process will end. In the following cases, stop rules are applied and node splitting is stopped (Breiman et al., 1984).

- When all observations in a node are compared with the outcome variable; when they have the same values,
- When the number of branches and nodes is less than the value specified by the user,
- The node is not split when the current tree depth reaches the highest limits of the value specified by the user.

The second step is to stop the tree creation process. Meanwhile, a "maximal" tree is constructed, possibly matching the information contained in the learning dataset.

Tree Pruning

Pruning removes redundant nodes from the tree to obtain the optimum decision tree. A very large tree does not provide us with real knowledge. A tree that is too small does not show all the required properties of the dataset. The pruning technique shows important features by reducing the size of the learning tree without reducing accuracy. There are two types of tree pruning methods. These are the "Cost Complexity Pruning" and "Reduced Defect Pruning" tree pruning technique.

The third step is "pruning" the more important nodes and creating a series of simpler trees. This process is created using the "pruning" technique.

Optimal Tree Selection

The termination process shows the sensitivity of the tree. In a decision tree with delayed termination, the number of branches increases. In addition, negative consequences can occur. Unobserved results can be critical relationships and classes. In a decision tree whose termination is done early, the possibility that full learning has not taken place comes to the fore. In the process of creating the decision tree or after the process is completed, the branches that do not affect the decision results and do not benefit the classification should be removed from the tree. In the classification

phase, the test data is used to query whether the decision tree is correct. The accuracy of a model applied to test data is the ratio of its correct classification to all classes in the test data. In each test sample, the known class is compared with the class predicted by the model. If the accuracy of the model is at an acceptable level, rules are used to classify new data.

4.2.1.3 Accuracy Tests of CART Algorithm

There are three accuracy tests in the CART for estimation. These are ordered as:

- replace,
- test sample prediction and
- cross validation.

It is the "n-fold cross validation" test, which is widely used for model validity testing. For this test, first of all, the data set is randomly divided into n (generally n=10) groups. One of these groups leaves and a model is established with the others. The classification performance of the model we built is tested on the excluded data. This process is repeated for the other groups in turn. And the average performance of k correct classification rates is evaluated (Breiman et al., 1984).

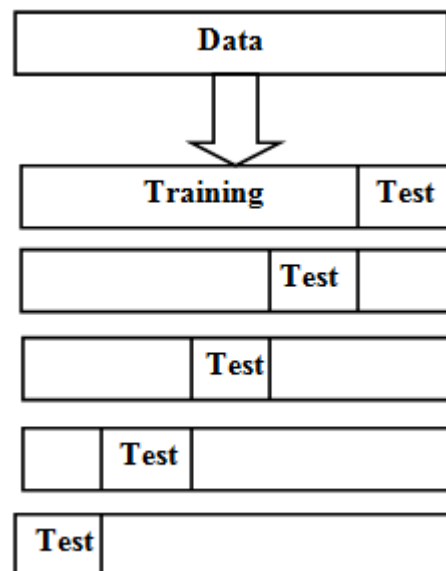


Figure 4.2 N-Fold cross validation (Priyadharshini.& Sundarambal, 2017)

The n-times cross-validation method is used to ensure that the results of the classification process on a data set of classifier models are consistent. Before applying the method, the parameter n must be determined.

4.2.1.4 Advantages and Disadvantages of CART Algorithm

The advantages of the CART algorithm are listed as follows:

- Performs variable scanning.
- It is quite simple to understand, interpret and visualize.
- Can use numerical and categorical data.
- It requires less effort from users to prepare the data. Nonlinear relationships within the parameters do not affect the reliability of the decision tree.

The disadvantages of the CART algorithm are listed as follows:

- Decision trees can form complex trees. This is called overfitting.
- The decision tree can be unstable. Minor changes in the data may result in the formation of a different tree.
- Greedy algorithms may not generate the best decision tree.
- If some classes become more dominant, it can create biased trees.

4.2.2 CHAID Algorithm

The CHAID algorithm, which is statistically based and intended for classification and regression, was developed by G. V. Kass in 1980. A regression equation to be created with CHAID is kept apart from known classical assumptions such as normal distribution, linearity and homogeneity. Because with a strong iteration algorithm, the whole population can be divided into stable sub-nodes. This situation can provide the assumptions of normality and homogeneity in the distribution of the data. For this reason, CHAID removes the distinction between parametric and nonparametric and has a statistically semi-parametric feature in the method algorithm(Firat, 2017).

The CHAID algorithm is a technique that finds the relationships, interactions and combinations of independent variables with each other. Therefore, the relationships

between the variables are also tested. With these tests, the dependencies between the variables are also examined.

In the CHAID analysis, we find the least significant category pair for each predictor variable. The method calculates p values based on the level of measurement of Y. If the dependent variable is continuous, the F test should be used. If the dependent variable is categorical, a two-way crosstab is prepared. The Pearson chi-square test or likelihood ratio test is used. If the target variable is ordered, an association model is determined and the likelihood ratio test is used. For the category pair of X with the largest p-value, compare the p-value with the predetermined alpha-level α . If the p value is greater than α unify, combine this pair under a single category. The new categorical set of X should start the process again. If the p value is less than α , the corrected p value is calculated using the Bonferroni correction.

In the CHAID analysis, the best split is found for each explanatory variable. The explanatory variables are then compared until the best is selected. All subsections are reanalyzed independently. It performs the possible divisions that each explanatory variable allows categories. Contingency tables are created according to the importance level in the chi-square test.

CHAID analysis uses chi-square statistics, Bonferroni's approach, and category association algorithm to enable the researcher to obtain the most important explanatory variables and interactions with the dependent variable by tree diagram. By applying the CHAID algorithm Chi-Square Test of Independence, starting with the variable with the strongest statistical relationship; branches the predictor variables from the target variable.

The test statistics to be used are determined according to the measurement level of the dependent variable. If the dependent variable is continuous, F statistics is used, and if categorical, chi-square statistics are used (Oğuzlar, 2004).

The reasons for using CHAID analysis can be listed as follows:

- Continuous and categorical data can be included in the model at the same time.

- The relationship between dependent and independent variables can be evaluated in more detail.
- It can show independent variables that affect dependent variables on a tree diagram.
- The tree diagram may be visually easier to interpret than other analysis results.
- The results obtained may be easier to understand.

4.2.2.1 Literature Review

According to Kayri & Boysan (2007), CHAID analysis should be used in advanced statistical methods, sensitive studies and applied educational science studies.

According to Avcı, Coşkuntuncel, & Inandi (2011), CHAID analysis should be used to determine the factors affecting students' positive or negative attitudes towards learning and to take measures according to the results obtained.

CHAID analysis was used in the detection studies of metabolic syndrome in young adults by Miller et al (2014).

Kağnicioğlu & Moğol (2014) tried to determine the most frequently used materials in room cleaning process with the CHAID algorithm and their relationships according to the data of a five-star hotel.

Isi et al. (2014) investigated the factors affecting the investment behavior of individuals with CHAID analysis.

Milanović & Stamenković (2016) identify the most dominant variables.

Önder & Uyar (2017) investigated the factors that affect the educational success of secondary school students.

Aksu & Keceoğlu (2019) investigated the variables that affect students' mathematics achievement.

4.2.2.2 Steps of the CHAID Algorithm

The steps of the CHAID algorithm are as follows:

Step 1: A crosstab is created with the independent and dependent variable categories in the data set. These include the pairs of categories that have the least importance, that is, the largest p-value.

- a) It is calculated with the F test for the continuous dependent variable.
- b) It is calculated with the chi-square test for the categorical dependent variable.

Step 2: The category pair with the largest p value is found. It is compared with the determined alpha level.

- a) When the p value is from alpha, this category pair is combined. It goes back to step 1 again.
- b) When the p value is less than alpha, go to step 3.

Step 3: Using Bonferroni correction, the adjusted p value is calculated for the dependent and independent variable categories.

Step 4: The most important argument is selected among the variables with the smallest corrected p value. This value is compared to the alpha level;

- a) When the p value is less than or equal to the alpha value, the node is divided by the specified argument.
- b) When the p value is greater than the alpha value, this node becomes the last node and no longer splits.

Step 5: This process continues until the tree stopping rules occur.

The chi-square test should be used to stop the loop. If homogeneity is denied in a node, separation continues. If homogeneity is achieved, this will be the final terminal node. As can be seen, the growth of the tree in the CHAID algorithm is determined by the stopping criterion of the chi-square test (Gülpınar, 2008).

4.2.3 QUEST Algorithm

QUEST is a decision tree algorithm for classification and data mining. A unique feature is that the feature selection method has a negligible bias (Loh & Shih, 1997).

It was developed by Loh and Shih experts in 1997. It allocates separate times to the variable that will ensure optimum division in the branching process. It determines the point at which optimum splitting will be achieved. In this respect, it differs from CART and CHAID algorithms. It provides faster results in classification models. In regression models, this algorithm cannot be used if the variable described in a decision tree application is continuous.

It stands for Fast, Unbiased, Efficient statistical tree. "Data classification and mining" is also implemented. When there are linear or univariate combinations it can be applied to separate them. The peculiarity of the QUEST algorithm is that it can neglect bias when selecting attributes. The bias in the feature selection method is negligible.

QUEST implements the binary tree structure. This is because the steps that make up the decision tree are used in binary structures. The CHAID and CART algorithms perform variable selection and splitting at the same time. The QUEST algorithm makes the feature selection general. This algorithm tries to reduce the cost. The argument must be nominal so we can use QUEST.

4.2.3.1 Literature Review

Low & Shih, (1997), "Split selection methods for classification trees, *Statistica Sinica*, 7, 815-840 (This is the definitive reference for QUEST.)."

Lim, Loh & Shih, (2000), "Comparison of prediction accuracy, complexity and training time of thirty-three old and new classification algorithms, *Machine Learning Journal*, 40, 203-228 (This article compares the performance of version 1.7 of QUEST to other methods.)."

Shih, (1999), “Families of separation criteria for classification trees, *Statistics and Computation*, 9, 309-315 (This article documents the extended class of splitting criteria in QUEST).”

Shih, (2004), “A note on split selection bias in classification trees, *Computational Statistics and Data Analysis*, 45, 457-466 (This article explains the existence of CART's split selection bias.)”

4.2.3.2 Features of the QUEST Algorithm

- It uses a neutral variable selection technique.
- Creates linear partitions using Fisher's LDA method.
- Applies proof to deal with missing values.
- It simply handles categorical predictive variables in many different situations.

The QUEST algorithm, developed by Loh and Shih experts in 1997, allocates separate times to decide on the variable that will provide the optimum division and the point at which the optimum division will be achieved in the branching process. In this respect, it differs from CART and CHAID algorithms and provides faster results in classification models. In regression models, if the variable explained in a decision tree application is continuous, this algorithm cannot be used. The QUEST estimator uses statistical tests when deciding whether to use the variables. The independent variable selection is made with the chi-square test if the variable is categorical, and with the F test if it is continuous.

This algorithm has been developed in order to make the biased selection more general and to reduce the computational cost during tree branching. Independent variables can be continuous or categorical, but the dependent variable should be categorical. In the algorithm, the target variable is categorical and unbiased tree is important, it is preferred when there is a large and complex data set and the tree is limited to binary division. For each separation of the algorithm, the association between each independent and dependent variable is found by calculating the Anova F test or Levene test or the Pearson Chi-Square value. For variable selection, the

variable with the minimum p value is selected. In the separation process, if the dependent variable has more than two categories, the two-means clustering algorithm is used to find two superclasses. Otherwise, Quadratic Discriminant Analysis (QDA) is used to find the best split of the independent variable. This process continues until any stopping rule is encountered. As with CART, the tree is pruned with the maximum-complexity measure. As with CHAID, the cost matrix does not directly participate in the tree growing process. However, for a symmetrical cost matrix, cost information can be included in the model using the preliminary probabilities in the cost matrix.

Cost information is only available in CART and QUEST algorithms. It is used to find branches that do not provide information about the correctness of the tree. The value parameter determines the magnitude of the two differences. It is the difference between the tree with the lowest risk and the pruned tree. For example, 2 is suitable for a tree whose risk estimate is greater than that of the full tree. Its default state is True, its value is 1. The maximum surrogates parameter can deal with missing values when they occur.

CHAPTER 5

APPLICATION

5.1 Purpose of Application

In practice, the life satisfaction data, which is a wide application area in data mining with using "classification" and "regression tree" methods of Turkey Statistics Institute (TURKSTAT) in 2017 were obtained from this database. The aim of this study is to determine the variables that are thought to be effective in life satisfaction. In the continuation of this study, logistic regression, which is one of the classification tree methods, was used first, while the decision tree methods CART, CHAID and QUEST algorithm were used in the other stages.

5.2 Life Satisfaction Definition and Literature Review

Data mining has many uses. It has a general evaluation of life in the investigation of socio-demographic characteristics in this area of use. In the literature, the general evaluation of life is examined as "Life Satisfaction" (Doğan & Sapmaz, 2012; TÜMÜ & Receptoğlu, 2011; Akın & Yalnız, 2015; Korkmaz et al., 2015; Kanbur & Özdemir, 2017). Life satisfaction can be affected by many situations because it is related to the emotional reactions of individuals to life and their own life and environment. Working life, income and health status, education level, environmental factors, social, physical and economic security can be counted among these situations (Kanbur & Özdemir, 2017). While positive feelings of individuals towards life increase life satisfaction, negative feelings decrease life satisfaction (Korkmaz et al., 2015). One of the largest and most important of the databasethe database of Turkey is the Turkey Statistical Institute (TURKSTAT, 2018).

The concept of life satisfaction has been studied in the literature with validity and accuracy studies, then descriptive statistics and basic hypothesis tests. While correlation and regression models are used for advanced analysis, it has been examined by modeling with classification and decision trees in the concept of machine learning in recent years.

Akın & Yalnız (2015) determined the validity and reliability of the life satisfaction scale by using factor analysis, Cronbach alpha coefficient and item analysis. Dağlı & Baysal (2016) adapted the life satisfaction scale of Diener et al. to Turkish using Cronbach's alpha and confirmatory factor analysis methods. Inferential statistics using statistical distributions (t, z, chi-square and F) related to life satisfaction have been used in many fields in recent years. Saygılı et al. (2017) determined whether life satisfaction differs in terms of personality traits of type A and B by using two independent samples t test. Bölükbaşı & Şentürk (2017) determined the relationship between life satisfaction and some demographic characteristics of employees in the insurance company with the chi-square relationship test. Demir & Murat (2017) and Taşlıyan et al. (2018) examined life satisfaction and some demographic characteristics with independent samples t-test and ANOVA. Tuncay & Fertelli (2018) examined the relationship between life activities and life satisfaction with chi-square.

Güler & Emeç (2006) examined life satisfaction in academic success using regression analysis and descriptive statistics. Demir (2011) studied relationships between identity functions and life satisfaction using correlation and regression analysis. Tümlü & Recepoğlu (2013) investigated psychological resilience and the relationship between life satisfaction. Korkmaz et al. (2015) examined life satisfaction with unit root, granger causality and regression analyzes.

By using different machine learning methods and TURKSTAT data, the life satisfaction of individuals with discriminant analysis was investigated by Gürsakal & Öngen (2008). Beşel (2015) studied with descriptive statistics on a provincial basis with TURKSTAT data. Berker (2015) examined its relationship with forced migration with regression analysis with TURKSTAT data. Arı & Yıldız (2016) investigated the sequential logistic regression analysis with TURKSTAT data. Finally Şehribanoğlu & Diler (2018) studied with CART and CHAID analysis, one of the decision tree algorithms with TURKSTAT data.

5.3 Data Definition

The data used in practice, Turkey Statistics Institute (TURKSTAT) in 2017 "life satisfaction" were obtained from the database. Life satisfaction scores of 8430 people in the "life satisfaction" database are given on a classifying scale. In this binary dependent variable (Life Satisfaction: LS) given in the classifier scale, S: Satisfied with Life (1); DS: Coded as I am Not Satisfied with Life (0). While the number of satisfied with life was 6408 (S = 76%), the number of those who were not satisfied with life was found to be 2022 (DS = 24%).

Life satisfaction is a binary variable and is considered as the dependent variable. There are 239 independent variables (risk factors) for the "satisfaction with life" scale. These variables are; gender, age, employment status, place of work, problem at work, marital status and satisfaction scales (health, marriage, income, social life, transportation, education, housing, district, work, earning, personal care, time spent in traffic, water, green field, relatives, friends and neighbors relations, public order, judicial situation, hospital, municipality, police, public services and country situation etc.).

5.4. Data Analysis

5.4.1 Descriptive Statistics

The first step in data analysis is the organization and definition of data. Since some of the variables thought to be related to satisfaction with life (LS) are similar and interrelated, the number of variables was reduced by looking at the correlation matrix. According to this matrix, the categorical variables of "gender", "marital status", "employment status", "sector", "school status", "age" and "health", "marriage", "income", "social life" and It was decided to take the scale variables for "transportation" satisfaction. The significance coefficient were taken as alpha 0.10 and 0.05. The data were analyzed in SPSS (24) and R(3.6.1, 2019) programs.

Since our variables do not come from normal distribution (they are categorical variable), Spearman Correlation coefficient was preferred. Since the statistical properties of the data set are important for data analysis and pre-processing, the chi-

square p values showing the frequencies of the selected categorical variables and their relationship with the satisfaction variable are given in Table 5.1.

Table 5.1 Chi-Square p-values for some variables

Variables	p-value
LS * Gender	0.001
LS * Marital Status	0.084
LS * Working Status	0.000
LS * Industry	0.000
LS * School	0.000
LS * Age	0.001
LS * Health	0.000
LS * Revenue	0.000
LS * Social Life	0.000
LS * Transportation	0.001

In Table 5.2, frequencies, percentages and chi-square p values of categorical independent variables are shown. In addition, relationship between independent variables and life satisfaction (LS) are also calculated in this table according to chi-square test.

Table 5.2 Frequencies, percentages and Chi-Square p-values for independent variables

	Satisfied(S)	Dissatisfied(DS)	Total	p-values	
	n(%)	n(%)	n(%)		
Gender					
Male	2983	954	3937(46.7)	0.034	
Woman	3492	1001	4493(53.3)		
Marital Status					
1: Not married	1299	404	1703(20.2)	0.009	
2: Married	4720	1456	6176(73.3)		
3: Widow	389	162	551(6.5)		
Working Status					
Working	2920	810	3686(43.7)	0.000	
Not working	3488	1212	4744(56.3)		
Sector					
Special	2373	719	3092(82.9)	0.000	
Public	547	93	640(17.1)		
School					
1: Not going	816	412	1228(14.6)	0.000	
2: Primary School	2047	765	2812(33.4)		
3: Secondary	838	300	1138(13.5)		
School					
4: High School	1326	324	1650(19.5)		
5: Associate	393	91	484(5.7)		
Degree					
6: Undergraduate	858	117	975(11.6)		
7: Msc.-Phd.	130	13	143(1.7)	0.004	
Age					
1: (-. 25)	1025	280	1305(15.5)		
2: (26.35)	1191	336	1527(18.1)		
3: (36.45)	1262	387	1649(19.6)		
4: (46.55)	1061	380	1441(17.1)		
5: (56.65)	852	310	1162(13.8)		
6: (66. -)	1017	329	1346(16.0)		

Statistical relationships between variables were analyzed using the chi-square test. The hypotheses for each relationship are given below for the Chi-square test.

H_0 = “There is no statistical relationship between Gender and Life Satisfaction variable.” H_1 = “There is statistical relationship between Gender and Life Satisfaction variable.” According to the chi-square relationship test, a statistical relationship was found between gender and life satisfaction with 95% confidence interval (p value = 0.034 $< \alpha = 0.05$).

H_0 = “There is no statistical relationship between Marital Status and Life Satisfaction variable.” H_1 = “There is statistical relationship between Marital Status and Life Satisfaction variable.” According to the chi-square relationship test, a statistical relationship was found between marital status and life satisfaction with 95% confidence interval (p value = 0.009 $< \alpha = 0.05$).

H_0 = “There is no statistical relationship between Working Status and Life Satisfaction variable.” H_1 = “There is statistical relationship between Working Status and Life Satisfaction variable.” According to the chi-square relationship test, a statistical relationship was found between working status and life satisfaction with 95% confidence interval (p value = 0.000 $< \alpha = 0.05$).

H_0 = “There is no statistical relationship between Working Sector and Life Satisfaction variable.” H_1 = “There is statistical relationship between Working Sector and Life Satisfaction variable.” According to the chi-square relationship test, a statistical relationship was found between working sector and life satisfaction with 95% confidence interval (p value = 0.000 $< \alpha = 0.05$).

H_0 = “There is no statistical relationship between School Level and Life Satisfaction variable.” H_1 = “There is statistical relationship between School Level and Life Satisfaction variable.” According to the chi-square relationship test, a statistical relationship was found between school level and life satisfaction with 95% confidence interval (p value = 0.000 $< \alpha = 0.05$).

H_0 = “There is no statistical relationship between Age and Life Satisfaction variable.” H_1 = “There is statistical relationship between Age and Life Satisfaction variable.” According to the chi-square relationship test, a statistical relationship was found between age and life satisfaction with 95% confidence interval (p value = 0.000 $< \alpha = 0.05$). According to the chi-square relationship test, a statistically

significant relationship was found between all variables and life satisfaction with 95% confidence interval.

In Table 5.3, frequencies and chi-square p values of satisfaction scales in likert scale are shown.

Table 5.3 Frequencies and Chi-Square p-values for satisfaction scales (S: Satisfied, DS: Dissatisfied)

Scales	H:Health (S-DS)	I:Income (S-DS)	SL: Social Life (S-DS)	T: Transportation (S-DS)
5: Very Satisfied	374-95	121-25	115-37	172-45
4: Satisfied	4363-1083	3172-642	3396-764	4513-1380
3: Neutral	1038-361	1391-301	1380-368	879-241
2: Dissatisfied	571-405	1514-841	1359-721	721-274
1: Very Dissatisfied	62-78	210-213	158-132	123-82
p-value	0.000	0.000	0.000	0.000

According to Table 5.3, it was haved that there was a statistical relationship between life satisfaction and health (H), income (I), social life (SL) and transportation (T) scales in 95% confidence interval (p values = 0.000 <0.05) (5: Very Satisfied (VS), 4: Satisfied (S), 3: Neutral (M), 2: Dissatisfied (DS), 1: Very Dissatisfied (VDS)).

5.4.2 Examination of Factors Affecting Life Satisfaction with Logistic Regression Model

For the regression step, which is within the supervised learning of machine learning for the life satisfaction data, variables that were found to be significant with the backward logistic regression model were established (Table 5.4).

Table 5.4 Logistic Reg. Model (80% Trainig Data): Wald Chi-Square p-values and OR(CI)

Variables	Coefficients	p-value	OR(CI)
Gender	0.17	0.009	1.18 (1.042-1.336)
Marital Status		0.005	
Married	0.26	0.000	1.31 (1.111-1.534)
Widow	0.19	0.000	1.21 (0.914-1.608)
School		0.000	
2: Primary School	0.30	0.000	1.34 (1.123-1.609)
3: Secondary School	0.35	0.000	1.41 (1.131-1.771)
4: High School	0.71	0.000	2.04 (1.637-2.547)
5: Associate Degree	0.78	0.000	2.19 (1.593-3.006)
6: Undergraduate	1.13	0.000	3.09 (2.353-4.071)
7: Msc.-Phd.	1.22	0.000	3.39 (1.802-6.408)
Health		0.000	
5: VS	0.87	0.000	2.39 (1.466-3.882)
4: S	1.01	0.000	2.75 (1.829-4.147)
3: N	0.77	0.000	2.17 (1.423-3.308)
2: DS	0.34	0.117	1.40 (0.918-2.146)
Income		0.000	
5: VS	1.03	0.000	2.79 (1.601-4.877)
4: S	1.11	0.000	3.03 (2.317-3.964)
3: N	1.15	0.000	3.15 (2.379-4.178)
2: DS	0.41	0.002	1.50 (1.158-1.943)
Social Life		0.000	
5: VS	0.13	0.627	1.14 (0.665-1.968)
4: S	0.67	0.000	1.96 (1.420-2.693)
3: N	0.70	0.000	2.01 (1.443-2.789)
2: DS	0.30	0.063	1.35 (0.984-1.851)
Transportation		0.000	
5: VS	1.04	0.000	2.82 (1.661-4.778)
4: S	0.73	0.000	2.07 (1.464-2.928)
3: N	0.96	0.000	2.61 (1.790-3.807)
2: DS	0.55	0.004	1.73 (1.193-2.521)
Constant		0.000	0.08
Reference: Male, Unmarried, and VDS on satisfaction scales			
Scales: 5: Very Satisfied (VS), 4: Satisfied (S), 3: Neutral(N), 2: Dissatisfied (DS), 1: Very Dissatisfied (VDS)			

In the logistic regression model, which was established in the form that 80% of the data would be training data and 20% test data, the chance of women being satisfied with life was 18% higher than men (OR = 1.18). In school situations, it was found that all categories starting from primary school (2) to Master & PhD (7) were more satisfied than those who did not attend school. It has been determined that those with a master's & doctorate degree are approximately 3.39 times more likely to be satisfied with life than those who do not attend school. Those who are married or divorced have a 31% and 21% more chance of being satisfied with life than those who are not married, respectively. In health satisfaction, the chances of being satisfied with life are approximately 2.75 times higher for those who are satisfied than those who are not satisfied. In income satisfaction, it has been found that the chance of having life satisfaction of "Dissatisfied" compared to "Very Dissatisfied" was found to be approximately 50% higher. In social service satisfaction, interpretation cannot be made since the confidence interval of the "Very Satisfied" and "Dissatisfied" odds ratio values includes 1. However, in social service satisfaction, it has been found that the chance of having life satisfaction of "Satisfied" compared to "Very Dissatisfied" was found to be approximately 96% higher. In transportation satisfaction, it has been determined that the chances of being "Very Satisfied" with life satisfaction are approximately 2.82 times higher than "Very Dissatisfied". The fact that all of these odds values do not contain 1 is also expressed in support of the comments.

Logit function for logistic regression is given as

Gender (Women)= $\pi(\hat{x})$

Marital Status Married= D_{21} Widow= D_{22}

School Primary School= D_{31} Secondary= D_{32} High School= D_{33}

Associate Degree= D_{34} Bachelor= D_{35} Msc.-Phd.= D_{36}

Health VS= D_{41} S= D_{42} N= D_{43} VD= D_{44}

Income VS= D_{51} S= D_{52} N= D_{53} VD= D_{54}

Social Life VS= D_{61} S= D_{62} N= D_{63} VD= D_{64}

Trasportation VS= D_{71} S= D_{72} N= D_{73} VD= D_{74}

$$g(\bar{x}) = \beta_0 + \beta_{11}D_{11} + \beta_{21}D_{21} + \beta_{21}D_{22} + \beta_{31}D_{31} + \beta_{31}D_{32} + \beta_{31}D_{33} + \beta_{31}D_{34} + \beta_{31}D_{35} + \beta_{31}D_{36} + \beta_{41}D_{41} + \beta_{41}D_{42} + \beta_{41}D_{43} + \beta_{41}D_{44} + \beta_{51}D_{51} + \beta_{51}D_{52} + \beta_{51}D_{53} + \beta_{51}D_{54} + \beta_{61}D_{61} + \beta_{61}D_{62} + \beta_{61}D_{63} + \beta_{61}D_{64} + \beta_{71}D_{71} + \beta_{71}D_{72} + \beta_{71}D_{73} + \beta_{71}D_{74}$$

$$g(\bar{x}) = -2.53 + 0.17D_{11} + 0.26D_{21} + 0.19D_{22} + 0.30D_{31} + 0.35D_{32} + 0.71D_{33} + 0.78D_{34} + 1.13D_{35} + 1.22D_{36} + 0.87D_{41} + 1.01D_{42} + 0.77D_{43} + 0.34D_{44} + 1.03D_{51} + 1.11D_{52} + 1.15D_{53} + 0.41D_{54} + 0.13D_{61} + 0.67D_{62} + 0.70D_{63} + 0.30D_{64} + 1.04D_{71} + 0.73D_{72} + 0.96D_{73} + 0.55D_{74}$$

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}$$

Here the coefficients are obtained from the logistic regression model. Logistic regression model is expressed with π .

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \quad g(x) \text{ including logit function,}$$

Prediction can be made with the established logistic regression model.

For example, a person whose gender is “female”, marital status is “married”, has “a master's degree” has given satisfaction to the health satisfaction survey, medium income satisfaction, moderate social life satisfaction and very satisfied answers to transportation satisfaction. Since the logistic regression probability of this person's satisfaction with life is greater than 0.5, it was found to be 1, that is, satisfied with life.

$$g(\bar{x}) = -2.53 + 0.17 + 0.26 + 1.22 + 1.01 + 1.15 + 0.70 + 1.04 = 3.02$$

$$\pi(\bar{x}) = \frac{e^{3.2}}{1 + e^{3.2}} = 0.954$$

For example, female gender, marital status, spouse died, a primary school graduate responded to the health satisfaction survey as not satisfied, not satisfied with income satisfaction, very satisfied with social life satisfaction and not satisfied

with transportation satisfaction. Since the logistic regression probability of this person's life satisfaction is less than 0.5, it was found to be 0, that is, not satisfied with life.

$$g(\bar{x}) = -2.53 + 0.19 + 0.30 + 0.34 + 0.41 + 0.13 + 0.55 = -0.61$$

$$\pi(\bar{x}) = \frac{e^{-0.61}}{1 + e^{-0.61}} = 0.352$$

Performance evaluation in the logistic regression model is given in the table below according to the accuracy, sensitivity, specificity, precision, F score and R^2 values of the model. In the logistic regression model, the cut-off point for probability is taken as 0.5.

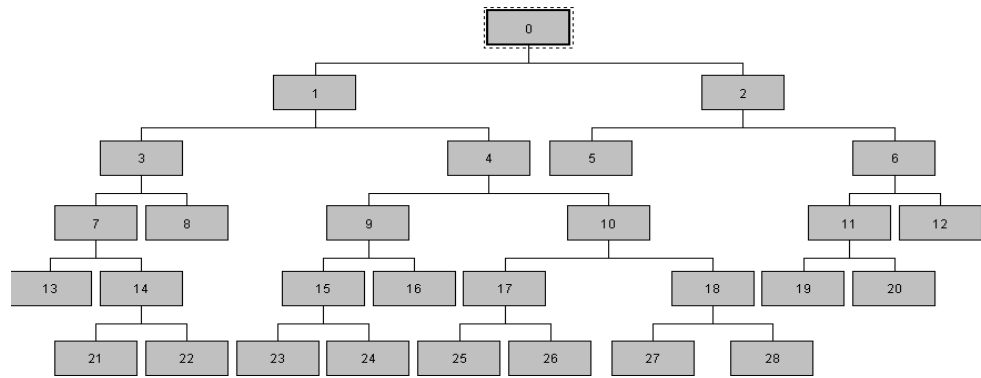
Table 5.5 Performance evaluation for logistic regression model

Algorithm	Accuracy	Sensitivity	Specificity	Precision	F-score	R^2
Log Reg(c=0.5)	0.779	0.796	0.552	0.961	0.870	0.671

According to this table, the accuracy value, sensitivity value, specificity value, precision value, F-Score value and R^2 value were given as 77.9%, 79.6%, 55.2%, 96.1%, 87% and 67.1%, respectively.

5.4.3 Examination of Factors Affecting Life Satisfaction with CART Algorithm

The maximum tree was created with CART, one of the classification and regression trees algorithms included in the supervised learning of life satisfaction data, machine learning, and the most appropriate tree structure was created with the pruning algorithm. For the logistic regression model, 80% of the data set is separated as training and 20% as test data, the data set we use for the CART algorithm is used with the Gini separation criterion. Figure 5.1 shows the first division of this tree. Nodes are numbered from 0 to 28. Terminal and sub-nodes of this tree, result variables, branches and comments of these nodes are given in Table 5.6. With the CART analysis using the Gini index, both the lower levels of the independent variables were determined and the relationships between the lower levels were obtained.



It was found by CART tree that the most important variable on life satisfaction is "income". The income variable is divided into two sub-nodes (1st sub-node = VDS, DS; 2nd sub-node = N, S, VS). It was determined that the proportion of those who were satisfied with life in the 1st sub node was 61.6%, and the ones who were not satisfied were 38.4%. In the income variable, the 1st sub-node is further divided into two sub-nodes (3rd sub-node = VDS, S, N; 4th sub-node = S, VS). It was determined that the proportion of those who were satisfied with life in the 4th sub node was 69.0% and the ones who were not satisfied were 31.0%. In the income variable, the 1st sub-node is divided into two sub-nodes (3rd sub-node = VDS, S, N; 4th sub-node = S, VS), and the 3rd sub-node is divided into two sub-nodes. In CART analysis, this process continued until the 28th sub node.

Table 5.6 CART analysis (80% Training Data)

Node	Variable	Classes	Satisfied(S)		Dissatisfied(DS)	
			N	%	n	%
0			5080	75.7	1631	24.3
1	Income	DS, VDS	1370	61.6	855	38.4
2	Income	VS, S, N	3710	82.7	776	17.3
3	Health	N, DS, VDS	444	50.3	438	49.7
4	Health	S, VS	926	69.0	417	31.0
5	School	P. sch., S. sch., nonsch.	2029	78.1	569	21.9
6	School	Associate, High Sch., Undergraduate, Msc.-Phd.	1681	89.0	207	11.0
7	Social life	DS, VDS	220	41.0	316	59.0
8	Social life	VS, S, N	224	64.7	122	35.3
9	Transportation	DS, VDS	152	57.8	111	42.2
10	Transportation	VS, S, N	774	71.7	306	28.3
11	Social life	S, N, DS, VDS	1631	89.6	190	10.4
12	Social life	VS	50	74.6	17	25.4
13	Income	DS, VDS	36	28.1	92	71.9
14	Income	VS, S, N	184	45.1	224	54.9
15	Gender	Male	70	49.3	72	50.7
16	Gender	Woman	82	67.8	39	32.2
17	Social life	DS, VDS	329	66.1	169	33.9
18	Social life	VS, S, N	445	76.5	137	23.5
19	School	Associate Degree, High Sch.	981	88.0	134	12.0
20	School	Undergraduate, Msc.-Phd.	650	92.1	56	7.9
21	Transportation	DS, VDS	28	34.1	54	65.9
22	Transportation	VS, S, N	156	47.9	170	52.1
23	School	High Sch., Undergraduate, nonsch.	23	38.3	37	61.7
24	School	Associate Degree, P. Sch., S.Sch., Msc.-Phd	47	57.3	35	42.7
25	Gender	Male	167	62.1	102	37.9
26	Gender	Woman	162	70.7	67	29.3
27	School	P. sch., S. sch., nonsch.	266	71.7	105	28.3
28	School	Associate, High Sch., Undergraduate, Msc.-Phd.	179	84.8	32	15.2

As a result of the CART analysis, it was found that the most effective variable on life satisfaction is "income" according to the decision tree. However, for the importance levels not seen in the CART tree, "normalized importance" values and order of importance are shown in Table 5.7. According to these values, it was observed that "income" variable is followed by "social life (86%)" and "health status (73.4%)", respectively.

Table 5.6 CART analysis normalized significance values

Independent variables	Importance Values	Normalized Significance Values (%)
Income	0.023	100.0
Social life	0.020	86.0
Health	0.017	73.4
Education	0.007	28.1
Transportation	0.002	9.8
Gender	0.001	4.0
Marital status	0.001	3.1

In addition, result variables, sub-nodes, comments and conclusion of each child node are tabulated in Table 5.8. According to conclusion, It has been determined whether the node is terminal or not.

Table 5.7 Nodes and results for CART algorithm

Node	Result Vrb. (Child Nodes)	Branches(Results)	Comments
1	B3 and B4 (Health)	B3=Health:VDS,DS,N Income:VDS,DS	Non-Terminal Node
		B4=Health:S,VS	Non-Terminal Node
3	B7 and B8 (Social Life)	B7=Social Life:VDS,DS; Health:VDS,DS,N Income:VDS,DS	Non-Terminal Node
		B8=Social Life:N,S,VS	Terminal Node 35.30% –NSWL; 64.70% –SWL
7	B13 and B14 (Income)	B13=Income:VDS,DS Social Life:VDS,DS Health:VDS,DS,N	Terminal Node 71.90%--NSWL; 28.10% –SWL
		B14=Income:N,S,VS	Terminal Node 71.90: %--NSWL; 28.10% –SWL.
14	B21 and B22 (Transportation)	B21=Transportation:VDS,DS Income:N,S,VS Social Life:VDS,DS Health:VDS,DS,N	Terminal Node 65.90% –NSWL; 34.10% –SWL
		B22=Transportation:N,S,VS	Terminal Node 52.10% –NSWL; 47.90% –SWL
4	B9 and B10 (Transportation)	B9=Transportation:VDS,DS Social Life:VDS,DS Health:S,VS Income:VDS,DS	Non-Terminal Node
		B10=Transportation:N,S,VS	Non-Terminal Node
9	B15 and B16 (Gender)	B15=Gender:Male Transportation:VDS,DS Social Life:VDS,DS Health:S,VS Income:VDS,DS	Non-Terminal Node
		B16=Gender:Woman	Terminal Node 32.20% –NSWL; 67.80% –SWL
2	B5 and B6 (Education)	B5=Education:P. Sch., S. Sch., nonsch. Income:N,S,VS	Terminal Node 21.90%–NSWL; 78.10%–SWL
		B6=Education:Associate, High Sch., Undergraduate, Msc.&Phd.	Non-Terminal Node

Table 5.8 continues

6	B11 and B12 (Social Life)	B11=Social Life:VDS,DS,N,S Education: Associate, High School, Undergraduate, Msc.&Phd. Income:N,S,VS	Non-Terminal Node
		B12=Social Life:VS	Terminal Node 25.40%–NSWL; 74.60% –SWL
11	B19 and B20 (Education)	B19=Education: Associate Degree, High School Social Life:VDS,DS,N	Terminal Node 12.00% –NSWL; %88.00 –SWL
		B20=Education: License, Msc.&Phd.	Terminal Node 7.90% –NSWL; 92.10%–SWL
15	B23 and B24 (Education)	B23=Education: High Sch., Undergraduate, nonsch. Social Life:VDS,DS	Terminal Node 61.70% –NSWL; 8.30% –SWL
		B24=Education:Associate Degree, P. Sch., S. Sch., Msc.&Phd.	Terminal Node 42.70% –NSWL; 57.30% –SWL
10	B17 and B18 (Social Life)	B17=Social Life:VDS,DS Health:S,VS Income:VDS,DS	Non-Terminal Node
		B18=Social Life:N,S,VS	Non-Terminal Node
17	B25 and B26 (Gender)	B25=Gender: Male Social Life:VDS,DS Health:S,VS Income:VDS,DS	Terminal Node 37.90% –NSWL; 62.10% –SWL
		B26=Gender: Woman	Terminal Node 29.30% –NSWL; 70.70% –SWL
18	B27 and B28 (Education)	B27=Education: P. Sch., S. Sch., nonsch. Social Life:N,S,VS Transportation:N,S,VS	Terminal Node 28.30% –NSWL; 1.70% –SWL
		B28=Education: P. Sch., S. Sch., nonsch.	Terminal Node 15.20% –NSWL; 84.80% –SWL

SWL: Satisfied with Life, NSWL: Not Satisfied with Life

With the established CART model, the percentage of satisfaction with life can be given. Using Figure 5.1 and Table 5.8, a person who is satisfied with his / her income

is a graduate graduate, and if he / she is satisfied with his social life, this person (D10) is 92.1% satisfied with life and 7.9% is not satisfied with life.

Performance evaluation in the CART model is given in the table below according to the accuracy, sensitivity, specificity, precision, F score and R^2 values of the model. Each was repeated 30 times using different 80% training and 20% test data. The results of these repetitions were averaged and the accuracy, sensitivity, specificity, precision, F-score and R values were calculated.

Table 5.8 Performance evaluation for CART algorithm

Algorithm	Accuracy	Sensitivity	Specificity	Precision	F-score	R^2
Decision Tree(CART)	0.780	0.790	0.578	0.964	0.875	0.543

According to this table, the accuracy value, sensitivity value, specificity value, precision value, F-Score value and R^2 value were given as 78%, 79%, 57.8%, 96.4%, 87.5% and 54.3%, respectively.

The purpose of separating the data set as training and test set is to avoid possible overfitting and to understand how the model performs on the data set that it has not seen before. However, there may be some errors due to the distribution of the model. In this study, "k-fold cross validation" method was used to minimize the errors in TURKSTAT data. For this, the training data was randomly divided into 10 pieces and 9 pieces were used for the training in 1 piece for the test data. This process was repeated 10 times. The performances from our classifier were evaluated with the average of the values obtained from each repetition. In each round, 834 observations were allocated to the test set, while the other parts were reserved for training. At the end of each round, it is to evaluate the performances from our classifier. To evaluate the classification success for each level, the accuracy, sensitivity, specificity, precision and F-score values were calculated on the test data set. As shown in Table 5.10, the proposed k-fold cross-over verification method achieves an average accuracy rate of about 77% on the test data set.

Table 5.9 k-fold cross validation

Multiple	Accuracy	Sensitivity	Specificity	Precision	F-Score
1	0.765	0.527	0.962	0.784	0.869
2	0.778	0.575	0.965	0.786	0.872
3	0.767	0.545	0.966	0.765	0.875
4	0.790	0.567	0.976	0.786	0.876
5	0.767	0.570	0.972	0.786	0.865
6	0.786	0.576	0.975	0.776	0.865
7	0.783	0.565	0.972	0.774	0.867
8	0.768	0.581	0.973	0.787	0.865
9	0.754	0.574	0.974	0.756	0.867
10	0.786	0.576	0.970	0.767	0.873
Mean	0.774	0.566	0.970	0.777	0.869

5.4.4 Examination of Factors Affecting Life Satisfaction with CHAID Algorithm

Tree structure was created with CHAID algorithm, which is one of the classification and regression trees algorithms within the supervised learning of machine learning.

The data set we use for the logistic regression and CART algorithm, of which 80% of the data set is separated as training and 20% as test data, is also used in the CHAID algorithm. Figure 5.2 shows the first division of this tree. In Table 5.12, terminal and sub-nodes of this tree, result variables, branches and comments for these nodes are given. Nodes are numbered from 0 to 22. With the CHAID analysis, both sub-levels of independent variables are determined and relationships between sub-levels are obtained. It was found with the CHAID tree that the most important variable on life satisfaction is "income". The income variable is divided into three sub-nodes (1st sub-node = VDS; 2nd sub-node = DS, VDS; 3rd sub-node = VS, S, N, DS). In the 1st sub node, the proportion of those who were satisfied with life was 47.5% and those who were dissatisfied was 52.5%. In the CHAID analysis, this process continued until the 22nd node.

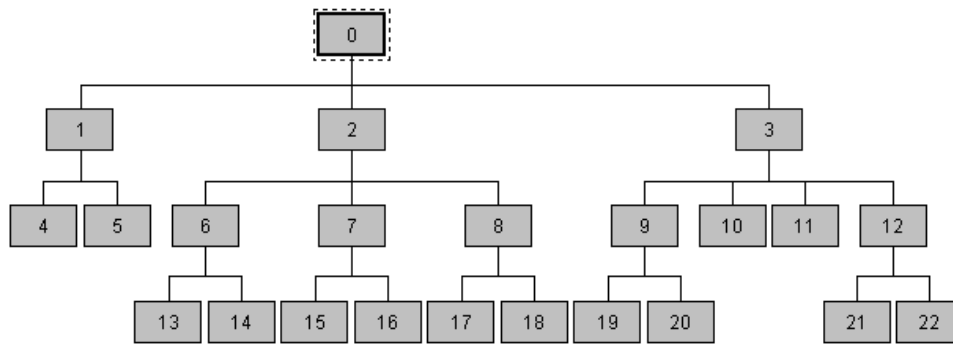


Figure 5.2 CHAID analysis (80% Training Data) decision tree numbering

Tree structure was created with the CHAID algorithm, which is one of the classification and regression tree algorithms, whose life satisfaction data is within the supervised learning of machine learning. The data set we use for logistic regression and CART algorithm, of which 80% of the data set is separated as training and 20% as test data, is also used in the CHAID algorithm. Figure 5.4 shows the first division of this tree. In Table 5.11, terminal and sub-nodes of this tree, result variables, branches and comments for these nodes are given. Nodes are numbered from 0 to 22. With the CHAID analysis, both the lower levels of the independent variables are determined and the relationships between the lower levels are obtained. It was found by CART tree that the most important variable on life satisfaction is "income". The income variable is divided into two sub-nodes (1st sub-node = VDS; 2nd sub-node = DS, VDS; 3rd sub-node = VS, S, N, DS). In the 1st sub node, the proportion of those who are satisfied with life is 47.5% and those who are not satisfied are 52.5%. In the CHAID analysis, this process continued until the 22nd node.

Table 5.10 CHAID analysis (80% Training Data)

			Satisfied(S)		Dissatisfied(DS)	
Node	Variable	Classes	n	%	N	%
0			5080	75.70	1631	24.30
1	Income	VDS	160	47.50	177	52.50
2	Income	DS, VDS	1210	64.10	678	35.90
3	Income	VS, S, N, DS	3710	82.70	776	17.30
4	Health	N, DS, VDS	61	34.90	114	65.10
5	Health	S, VS	99	61.10	63	38.90
6	Health	DS, VDS	191	48.70	201	51.30
7	Health	N, DS	192	61.00	123	39.00
8	Health	VS, S	827	70.00	354	30.00
9	Education	Associate Degree, High Sch.	1012	87.30	147	12.70
10	Education	P. Sch., S. Sch.	1585	79.50	409	20.50
11	Education	Undergraduate, Msc. & Phd.	669	91.80	60	8.20
12	Education	Nonsch.	444	73.50	160	26.50
13	Social Life	DS, VDS	99	41.20	141	58.80
14	Social Life	VS, S, N	92	60.50	60	39.50
15	Social Life	DS, VDS	85	50.60	83	49.40
16	Social Life	VS, S, N	107	72.80	40	27.20
17	Gender	Sale	403	65.00	217	35.00
18	Gender	Woman	424	75.60	137	24.40
19	Health	N, DS, VDS	935	88.40	123	11.60
20	Health	VS	77	76.20	24	23.80
21	Health	DS, VDS	81	61.40	51	38.60
22	Health	VS, S, N	363	76.90	109	23.10

In addition, result variables, sub-nodes, comments and conclusion of each child node are tabulated in Table 5.12. According to conclusion, It has been determined whether the node is terminal or not.

Table 5.11 Nodes and results for CHAID algorithm

Node	Result Vrb. (Childs N.)	Branches(Results)	Comments
1	B4 and B5 (Health)	B4=Health:VDS,DS,N; Income:VDS,DS B5=Health:S,VS	Terminal Node 65.10% –NSWL; 34.90% –SWL Terminal Node. 38.90% –NSWL; 61.10% –SWL
2	B6 B7 and B8 (Health)	B6=Health:VDS,DS; Income:VDS,DS B7=Health:DS,N B8=Health:S,VS	Non-Terminal Node Non-Terminal Node Non-Terminal Node
6	B13 and B14 (Social Life)	B13=Social Life:VDS,DS; Health:VDS,DS Income:VDS,DS B14=Social Life:N,S,VS	Terminal Node. 58.80%–NSWL; 41.20%–SWL Terminal Node. 39.50%–NSWL; 60.50% –SWL
7	B16 and B16 (Social Life)	B15=Social Life:VDS,DS; Health:VDS,DS Income:VDS,DS B16=Social Life:N,S,VS	Terminal Node 49.40% –NSWL; 0.60% –SWL Terminal Node 27.20% –NSWL; 72.80% –SWL
8	B17 and B18 (Gender)	B17=Gender:Male; Health:S,VS Income:VDS,DS B18=Gender:Women	Terminal Node 35.00%–NSWL; 65.00%–SWL Terminal Node 24.40%–NSWL; 75.60%–SWL
3	B9 and B10 (Education)	B9=Education: Associate, High Sch. Income:DS,N,S,VS B10=Education:P. Sch., S. Sch. B11=Education:Undergraduated, Msc.&Phd. B12=Education: Nonsch.	Non-Terminal Node Terminal Node 20.50%–NSWL; 79.50%–SWL Terminal Node 8.20%–NSWL; 91.80%–SWL Non-Terminal Node
9	B19 and B20 (Health)	B19=Health:VDS,DS,N Education: 2_3 years college, High Sch. Income: DS B20=Health:VS	Terminal Node 11.60%–NSWL; 88.40%–SWL Terminal Node 23.80%–NSWL; 76.20%–SWL
12	B21 and B22 (Health)	B21=Health:VDS,DS; Education: Nonsch. Income: Dissatisfied B22=Health:N,S,VS	Terminal Node 38.60%–NSWL; 61.40%–SWL Terminal Node 23.10%–NSWL; 76.90%–SWL
SWL: Satisfied with Life, NSWL: Not Satisfied with Life			

With the established CHAID model, the percentage of satisfaction with life can be given.

The example used in the CART model is used in the CHAID model. Using Figure 5.3 and Table 5.12, a person who is satisfied with his / her income is a graduate graduate, if he is satisfied with his social life, 91.80% of this person is satisfied with life with the (B11) node, and 8.20% is not satisfied with life.

Performance evaluation in the CHAID model is given in the table below according to the accuracy, sensitivity, specificity, precision, F score and R^2 values of the model. Each was repeated 30 times using different 80% training and 20% test data. The results of these repetitions were averaged and the accuracy, sensitivity, specificity, precision, F-score and R^2 values were calculated.

Table 5.12 Performance evaluation for CHAID algorithm

Algorithm	Accuracy	Sensitivity	Specificity	Precision	F-score	R^2
Decision Tree(CHAID)	0.771	0.781	0.614	0.963	0.873	0.544

According to this table, the accuracy value, sensitivity value, specificity value, precision value, F-Score value and R^2 value were given as 77.1%, 78.1%, 61.4%, 96.3%, 87.3% and 54.4%, respectively.

5.4.5 Examination of Factors Affecting Life Satisfaction with QUEST

The tree structure was created with the QUEST algorithm from the classification and regression tree algorithms, whose life satisfaction data are within the supervised learning of machine learning. Logistic regression with 80% of the data set separated as training and 20% as test data. The dataset we use for the CART and CHAID algorithm is also used in the QUEST algorithm. Figure 5.3 shows the first division of this tree. Table 5.14 shows the terminal and sub-nodes of this tree. result variables. branches and comments for these nodes are given.

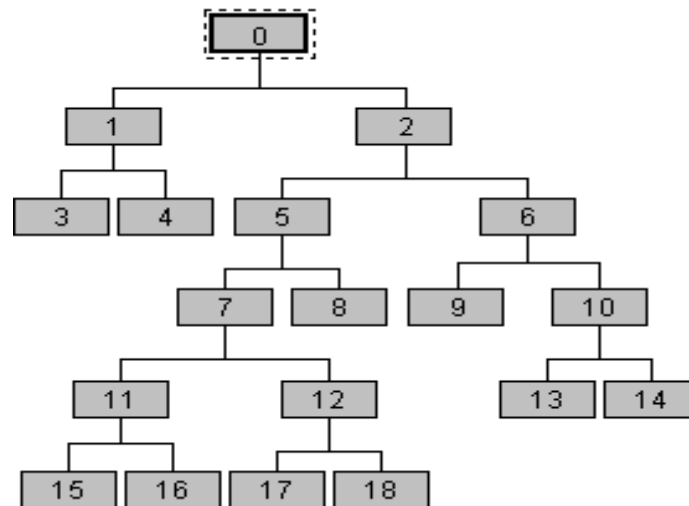


Figure 5.3 QUEST analysis (80% Training Data) decision tree numbering

Table 5.13 QUEST analysis (80% Training Data)

Node	Variable	Classes	Satisfied(S)		Dissatisfied(DS)	
			n	%	N	%
0			5080	75.7	1631	24.3
1	Income	DS, VDS	160	47.5	177	52.5
2	Income	VS, S, N	4920	77.2	1454	22.8
3	Health	N, DS, VDS	61	34.9	114	65.1
4	Health	S, VS	99	61.1	63	38.9
5	Income	N, DS, VDS	2319	71.6	920	28.4
6	Income	VS, S	2601	83.0	534	17.0
7	Income	DS, VDS	1210	64.1	678	35.9
8	Income	VS, S, N	1109	82.1	242	17.9
9	Education	P. Sch., S. Sch., nonsch.	1373	78.1	386	21.9
10	Education	Associate, High Sch., Undergraduate, Msc. & Phd.	1228	89.2	148	10.8
11	Health	DS, VDS	191	48.7	201	51.3
12	Health	VS, S, N	1019	68.1	477	31.9
13	Transportation	N, DS, VDS	365	88.8	46	11.2
14	Transportation	VS, S	863	89.4	102	10.6
15	Social Life	VS, S	99	41.2	141	58.8
16	Social Life	VS, S, N	92	60.5	60	39.5
17	Social Life	DS, VDS	433	61.5	271	38.5
18	Social Life	VS, S, N	586	74.0	206	26.0

In addition, result variables, sub-nodes, comments and conclusion of each child node are tabulated in Table 5.15. According to conclusion, It has been determined whether the node is terminal or not.

Table 5.14 Nodes and results for QUEST algorithm

Node	Result Vrb. (Child Nodes)	Branches(Results)	Comments
1	B3 and B4 (Health)	B3=Health:VDS,DS,N Income:VDS,DS B4=Health:S,VS	<i>Terminal Node</i> 65.10% –NSWL; 34.90% –SWL <i>Terminal Node</i> 38.90% –NSWL; 61.10% –SWL
2	B5 and B6 (Income)	B5=Income:VDS,DS,N B6=Income:S,VS	Non-Terminal Node Non-Terminal Node
5	B7 and B8 (Income)	B7=Income:VDS,DS B8=Income:N,S,VS	Terminal is not a node. <i>Terminal is node.</i> 17.00% -NSWL; 83.00 %–SWL
6	B9 and B10 (Education)	B9=P. Sch.; S. Sch. primary education; Nonsch. Income:S,VS B10= Associate; High Sch.; Undergraduate; Msc.&Phd.	<i>Terminal Node</i> 21.90% –NSWL; 78.10% –SWL Non-Terminal Node
10	B13 and B14 (Education)	B13= Education: Associate; High Sch. Income:S,VS B14=Education: Undergraduate; Msc.&Phd.	<i>Terminal Node</i> 12.90% –NSWL; 87.10%–SWL <i>Terminal Node</i> 7.50% –NSWL; 92.50 %–SWL
7	B11 and B12 (Health)	B11=Health:VDS,DS Income:VDS,DS B12=Health:N,S,VS	Non-Terminal Node Non-Terminal Node
11	B15 and B16 (Social Life)	B15=Social Life:VDS,DS Health:VDS,DS B16=Social Life:N,S,VS	<i>Terminal Node</i> 58.80%–NSWL; 41.20%–SWL <i>Terminal Node</i> 39.50%–NSWL; 60.50%–SWL
12	B17 and B18 (Social Life)	B17=Social Life:VDS,DS Health:N,S,VS B18=Social Life:N,S,VS	<i>Terminal Node</i> 38.50%–NSWL; 61.50%–SWL <i>Terminal Node</i> 26.00%–NSWL; 74.00%–SWL
SWL: Satisfied with Life, NSWL: Not Satisfied with Life			

As a result of the QUEST analysis, it was found that the most effective variable on life satisfaction is "income" according to the decision tree.

With the established CHAID model, the percentage of satisfaction with life can be given.

The example used in the CART model is used in the QUEST model. Using Figure 5.3 and Table 5.15, a person who is satisfied with his / her income is a graduate graduate, if he is satisfied with his social life, this person is 88.80% satisfied with life with (B13) node, and 11.20% is not satisfied with life.

Each was repeated 30 times using different 80% training and 20% test data. The results of these repetitions were averaged and the accuracy, sensitivity, specificity, precision, F-score and R^2 values were calculated.

Table 5.15 Performance evaluation for QUEST algorithm

Algorithm	Accuracy	Sensitivity	Specificity	Precision	F-score	R^2
Decision Tree(QUEST)	0.767	0.783	0.615	0.969	0.872	0.545

According to this table, the accuracy value, sensitivity value, specificity value, precision value, F-Score value and R^2 value were given as 77.1%, 78.1%, 61.5%, 96.8%, 86.2% and 54.4%, respectively.

Node numbers, terminal node numbers and tree width numbers for CART, CHAID and QUEST decision tree algorithms are given in Table 5.17.

Table 5.16 Node numbers for CART CHAID and QUEST algorithms

Decision Algorithm	Number of Nodes	Number of Terminal Nodes	Depth
CART	29	15	5
CHAID	23	14	3
QUEST	19	10	5

5.5 Performance Comparisons

The best way to summarize the results of an established logistic regression model is to create a classification table. This table is obtained by cross classification of a binary variable produced by the logistic probabilities estimated by the levels of the final value Y. To obtain this derived binary argument, a cutoff point is determined and each estimated probability value is compared with c. If the estimated probability

exceeds the value of c , the derived binary variable will equal 1. Otherwise it is equal to 0. The most commonly used value of c is 0.5. Accordingly, performance evaluations for the new c values changed in logistic regression, CART, CHAID.

Table 5.18 Performance evaluation according to algorithms of models

Algorithm	Accuracy	Sensitivity	Specificity	Precision	F-Score	R ²
Logistic Regression (c=0.5)	0.779	0.796	0.552	0.961	0.870	0.671
Decision Tree (CART)	0.780	0.790	0.552	0.961	0.870	0.671
Decision Tree (CHAID)	0.771	0.781	0.614	0.963	0.873	0.544
Decision Tree (QUEST)	0.767	0.783	0.615	0.969	0.872	0.545
k-Fold Cross Validation	0.774	0.772	0.534	0.943	0.864	0.544
Logistic Regression (c=0.6)	0.761	0.808	0.459	0.906	0.854	0.643
Logistic Regression (c=0.4)	0.782	0.781	0.579	0.988	0.872	0.675

When the accuracy values for the Life Satisfaction TUIK data set are examined, while the c value is 0.4, the logistic regression model has the highest value. When compared as sensitivity values, the logistic regression model with a c value of 0.6 gave the highest result. When looking at the specificity values, the QUEST algorithm gave the highest result. Considering the precision values. The logistic regression model with a c value of 0.4 gave the highest result. The algorithm with the highest results for the F-score is the decision tree. The performance values obtained as a result of running these algorithms are shown comparatively (Figure 5.6).

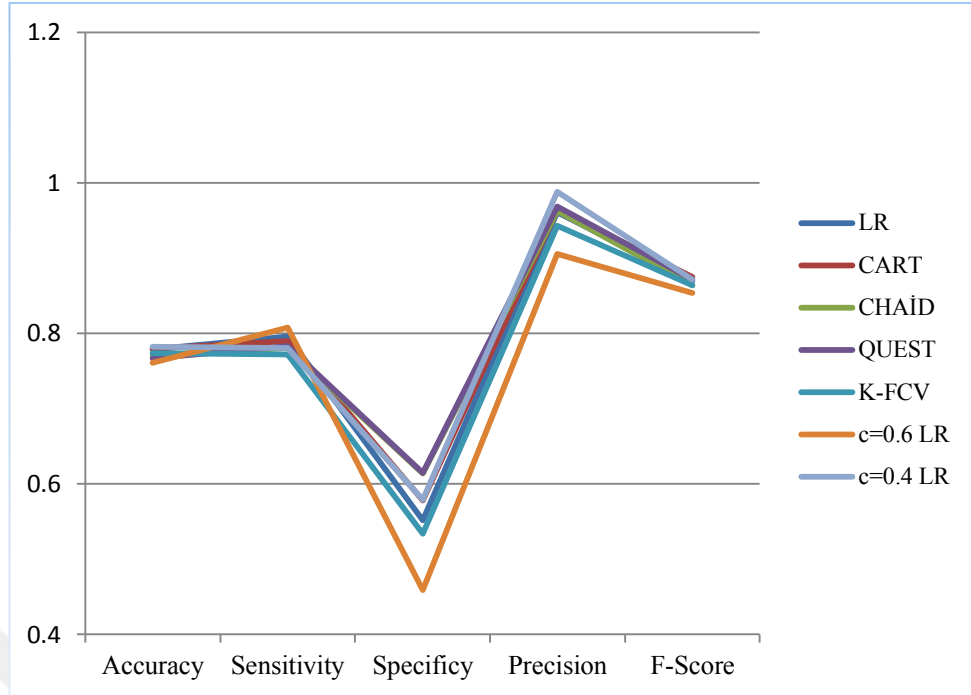


Figure 5.4 Performance of the models

5.6 Model Selection

A tree structure was created for 100% data set with the QUEST algorithm, which is one of the classification and regression tree algorithms, whose life satisfaction data is within the supervised learning of machine learning. The first division of this tree is shown in Figure 5.5. In Table 5.19, terminal and sub-nodes of this tree, result variables, branches and comments for these nodes are given.

Nodes are numbered from 0 to 18. With QUEST analysis, both sub-levels of independent variables are determined and relationships between sub-levels are obtained. It was found by the QUEST tree that the most important variable on life satisfaction is "income".

The SPSS output of the 1, 2, 3 nodes of the decision tree structure shown in Table 20 and Figure 5.5 is given in Figure 5.6. In addition, result variables, sub-nodes, comments and result of each node are tabulated.

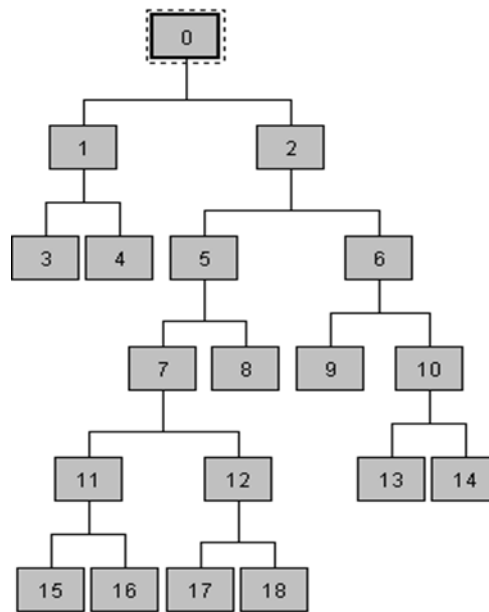


Figure 5.5 QUEST analysis (100% Training Data) decision tree numbering

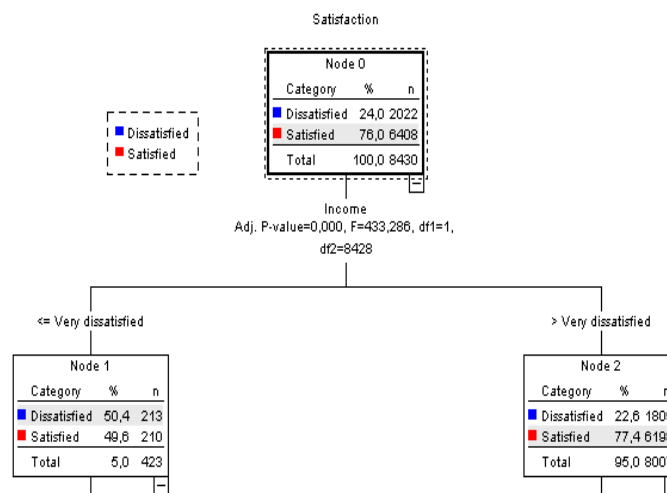


Figure 5.6 SPSS output for QUEST

Table 5.19 QUEST analysis (100% Training Data)

Node	Variable	Classes	Satisfied(S)		Dissatisfied(DS)	
			n	%	n	%
0			6408	76.0	2022	24.0
1	Income	DS, VDS	210	49.6	213	50.4
2	Income	VS, S, N	6198	77.4	1809	22.6
3	Health	N, DS, VDS	78	36.6	135	63.4
4	Health	S, VS	132	62.9	78	37.1
5	Income	N, DS, VDS	2905	71.8	1142	28.2
6	Income	VS, S	3293	83.2	667	16.8
7	Income	DS, VDS	1514	64.3	814	35.7
8	Income	VS, S, N	1391	82.2	301	17.8
9	Education	P. Sch., S. Sch., Nonsch.	1739	78.4	478	21.6
10	Eduacation	Associate, High Sch., Undergraduate, Msc.&Phd.	1554	89.2	189	10.8
11	Health	DS, VDS	243	48.2	261	51.8
12	Health	VS, S, N	1271	68.7	580	31.3
13	Transportation	N, DS, VDS	928	87.1	138	12.9
14	Transportation	VS, S	626	92.5	51	7.5
15	Social Life	VS, S	130	41.8	181	58.2
16	Social Life	VS, S, N	113	58.5	80	41.5
17	Social Life	DS, VDS	536	62.0	329	38.0
18	Social Life	VS, S, N	735	74.5	251	25.5

Table 5.20 Nodes and results for QUEST algorithm

Node	Result Vrb. (Child Nodes)	Branches(Results)	Comments
1	B3 and B4 (Health)	B3=Health:VDS,DS,N Income:VDS,DS B4=Health:S,VS	<i>Terminal Node</i> 63.40% –NSWL; 36.60% –SWL <i>Terminal Node</i> 37.10% –NSWL; 62.90% –SWL
2	B5 and B6 (Income)	B5=Income:VDS,DS,N B6=Income:S,VS	Non-Terminal Node Non-Terminal Node
5	B7 and B8 (Income)	B7=Income:VDS,DS B8=Income:N,S,VS	Terminal is not a node. <i>Terminal is node.</i> 17.80% –NSWL; 82.20 %–SWL
6	B9 and B10 (Education)	B9=Education: P. Sch.; S. Sch. Primary education; Nonsch. Income:S,VS B10=Education: Associate; High Sch.; Undergraduate; Msc.&Phd.	<i>Terminal Node</i> 21.60% –NSWL; 78.40% –SWL Non-Terminal Node
10	B13 and B14 (Education)	B13= Education: Associate; High Sch. Income:S,VS B14=Education: Undergraduate; Msc.&Phd.	<i>Terminal Node</i> 12.90% –NSWL; 87.10%–SWL <i>Terminal Node</i> 7.50% –NSWL; 92.50 %–SWL
7	B11 and B12 (Health)	B11=Health:VDS,DS Income:VDS,DS B12=Health:N,S,VS	Non-Terminal Node Non-Terminal Node
11	B15 and B16 (Social Life)	B15=Social Life:VDS,DS Health:VDS,DS B16=Social Life:N,S,VS	<i>Terminal Node</i> 58.20%–NSWL; 41.80%–SWL <i>Terminal Node</i> 41.50%–NSWL; 58.50%–SWL
12	B17 and B18 (Social Life)	B17=Social Life:VDS,DS Health:N,S,VS B18=Social Life:N,S,VS	<i>Terminal Node</i> 38.00%–NSWL; 62.00%–SWL <i>Terminal Node</i> 25.50%–NSWL; 74.50%–SWL
SWL: Satisfied with Life, NSWL: Not Satisfied with Life			

With the QUEST model established, the percentage of satisfaction with life can be given.

Using Figure 5.5 and Table 5.20, a graduate who was satisfied with his income level found that 92.50% was satisfied with life with the D14 node, while 7.50% was not satisfied with life.

CHAPTER 6

CONCLUSION

Information has become a very important concept in every field and for everyone, especially in recent years. Especially with computers coming into our lives in every field; information can be stored, shared and analyzed more. Stronger, more accurate and faster statistical data analysis processes have also been developed depending on the aforementioned developments in computer technology and information, and resources that increase day by day in response to increasing needs. Therefore, in this process, data mining methods that facilitate the analysis, inspection and regulation of data volumes that grow depending on information technology have emerged. However, having a lot of data and converting these data to usable information can only be possible with good management. Raw data that is not transformed into information does not contribute to the area where it is located. Along with technological developments, the use, analysis and interpretation of existing data is important for companies or individuals in the field. However, the fact that the data preparation phase is the most time-consuming part of the process in data mining studies shows the importance of data quality and integrity. Disruption in the structure of data can be shown as different or incomplete data entry of different users, inconsistent or non-functional data structures, and factors that disrupt data quality and integrity. It can be said that all of these are due to the deficiencies related to the data and the transformation process of data into information.

Each field has a unique data structure. The variables of interest in many areas show themselves in a categorical structure. Until today, multivariate statistical methods such as clustering, discriminant and logistic regression analysis have been used in the classification and analysis of categorical data. However, with the increase in data size in the presence of technological developments, decision tree algorithms, which are newer and more popular than these methods, have started to be used.

Life satisfaction means the level of positive feeling that people reach, based on their evaluation of their personal life quality as a whole. However, life satisfaction

includes pleasure and tells how people enjoy the life they live. It is seen that factors such as having a life with a goal and meaningfulness, being enjoyed in the living environments, and having many areas of occupation in life are directly related to life satisfaction.

The data set subject to the application is the people who participated in the 2018 TURKSTAT "Life satisfaction" survey. For this purpose, when TURKSTAT and life satisfaction were examined together in studies conducted in the literature, Gürsakal & Öngen (2008) found the variables of gender, rural-urban status, health, housing, income, relatives and neighbors relationships using discriminant analysis in their study where they examined life satisfaction in 2007. Beşel (2015) analyzed life satisfaction with descriptive statistics using 2013 TURKSTAT data and compared the happy & unhappy provinces socially and politically. Arı & Yıldız (2016) examined the TUIK life satisfaction for 2014 with the sequential logistic regression method. According to the partial proportional ratio model, they found the variables of gender, marital status, employment status, health, income, and friends to be significant. Finally, Şehribanoğlu & Diler (2018), using the decision tree algorithms CART, CHAID and analysis, examined the life satisfaction of TURKSTAT for 2013. In this study, they found that income with CHAID analysis, income with CART analysis, hope, health, marriage and social security variables were significant. However, they did not compare the performance of the model and stated that the CHAID algorithm was better than the number of variables entering the model.

In this study, the life satisfaction of 8430 people who participated in the 2018 TURKSTAT "Life satisfaction" survey was obtained on a classifying scale. For this scale, 7 variables were included in the models with the operations carried out during the data preparation phase. Firstly, logistic regression analysis, which is interpreted with the help of odds ratios estimated for the risk factors in the model, was applied. The data were repeated 30 times using 80% training and 20% test data. In the logistic regression analysis made according to the $c = 0.5$ cut-off point, it was determined that the variables of gender, marital status, school status, income, social life, health and transportation were significant. According to this model, it was concluded that women, married or widowed, readers, and most of the other categories other than

"not satisfied at all" in the scale are satisfied with life. The accuracy value, sensitivity value, specificity value, precision value, F-Score value and R^2 value were found to be 77.9%, 79.6%, 55.2%, 96.1%, 87% and 67.1%, respectively.

According to the results obtained by CART analysis; It has been found that the most influential variable on life satisfaction is income according to the decision tree. It is observed that the depth of this decision tree is 5, the number of nodes is 29 and the number of terminal nodes is 15. In addition, according to the normalized significance values in the CART analysis, it was found that the variables found important after income were social life and health status. Each was repeated 30 times using different 80% training and 20% test data. The results of these repetitions were averaged and the accuracy, sensitivity, specificity, precision, F-score and R values were calculated. In the calculated algorithm, the accuracy value, sensitivity value, specificity value, precision value, F-Score value and R^2 value were found to be 78%, 79%, 57.8%, 96.4%, 87.5% and 54.3%, respectively.

According to the results obtained by CHAID analysis; It has been found that the most influential variable on life satisfaction is income according to the decision tree. It is observed that the depth of this decision tree is 3, the number of nodes is 23 and the number of terminal nodes is 14. Each was repeated 30 times using different 80% training and 20% test data. The results of these repetitions were averaged and the accuracy, sensitivity, specificity, precision, F-score and R values were calculated. In the calculated algorithm, the accuracy value, sensitivity value, specificity value, precision value, F-Score value and R^2 value were found to be 77.1%, 78.1%, 61.4%, 96.3%, 87.3% and 54.4%, respectively.

According to the results obtained by QUEST analysis; It has been found that the most influential variable on life satisfaction is income according to the decision tree. It is given that the depth of this decision tree is 5, the number of nodes is 19, and the number of terminal nodes is 10. Each was repeated 30 times using different 80% training and 20% test data. The results of these repetitions were averaged and the accuracy, sensitivity, specificity, precision, F-score, and R values were calculated. In the calculated algorithm, the accuracy value, sensitivity value, specificity value,

precision value, F-Score value and R^2 value were found to be 77.1%, 78.1%, 61.5%, 96.8%, 86.2% and 54.4%, respectively.

In order to minimize the errors in TURKSTAT data, the k – folded cross validation method was also used. In the performance evaluation for logistic regression, CART, CHAID, QUEST and k-fold cross validation methods included in classification and decision trees, a logistic model with $c = 0.4$ cutoff point should be preferred. Compared to the sensitivity values, the logistic regression model with a c value of 0.6 yielded the highest result. Looking at the specificity values, the QUEST algorithm gave the highest result. Considering the precision values. The logistic regression model with a value of 0.4 c gave the highest result. The algorithm with the highest results for the F score is the QUEST decision tree.

This study demonstrates the use of algorithms in machine learning for analyzing and interpreting demographic and scale data in TURKSTAT data. Performance comparisons that can be made using the classification and decision tree methods used in the study and data with the same characteristics were determined.

Income differences are of great importance in determining the life satisfaction levels of individuals. Differences in income directly affect people's living standards and quality of life. Corrections and improvements should be made on situations that negatively affect life satisfaction. The happy and contented individual will live more productively and efficiently. Social development and advances will come with them when countries make individual happiness and better life reach to wider masses as one of their main goals. Comparative analysis of determinants of life satisfaction by income groups can be expanded by adding other economic variables with future studies. In addition, by obtaining data sets prepared for different years, the changes in the effects of income differences on life satisfaction over time can be examined.

As a result, machine learning methods are statistically based and can be applied to many different fields and data types. In this study, it has been applied to a social data set that has a flowing data structure with human factors and is indispensable in every period. However, machine learning has started to be used in vaccination studies and construction data such as asphalt dump, which is a growing sector very recently. Cost

calculation gains importance especially in these data types. For example, if it is known which factor is important in determining profit and loss in asphalt casting data, companies can determine their policies and ensure their sustainability. In the future, it is planned to use machine learning methods with such data to evaluate performance.



REFERENCES

- Abdallh, M. M. A., Bilal, K. H.& Babiker, A. (2016). Machine learning (Pattern Recognition) Review. *International Journal of Engineering, Applied and Management Sciences Paradigms*, 36 (1), 17-27.
- Abellan, J., Lopez, G., & De Ona, J. (2013). Analysis of traffic accident severity using decision rules via decision trees. *Expert Systems with Applications*, 40 (15), 6047–6054.
- Akbulut, R., & Rencber, O.F. (2015). Evaluation of efficiencies based on financial performance with data envelopment and logistic regression analysis in the cement businesses. *International Journal of Alanya Faculty of Business*, 7 (3), 123-135.
- Akpınar, H. (2000). Veri tabanlarında bilgi keşfi ve veri madenciliği. *Istanbul University Journal of the School of Business Administration*, 29 (1), 1-22.
- Aksu, G., Reyhanlioglu, G., & Keceoglu, C. (2019). Comparison of results obtained from logistic regression, CHAID analysis and decision tree methods. *Eurasian Journal of Educational Research*, 84, 115-134.
- Alpar, R. (2013). *Uygulamalı çok değişkenli istatistiksel yöntemler*. Ankara: Detay Yayıncılık.
- Alpaydm, E. (2010). *Introduction to machine learning* (2nd ed.). *Massachusetts Institute of Technology*. London: The MIT Press.
- Altunoz, U. (2015). Prediction of financial failure of banks with logistics regression and diskriminant analysis: Case of Turkish banks. *Kahramanmaras Sutcu Imam University Faculty of Economics and Administrative Sciences Journal*, 5 (1), 45-54.
- Arı, E. (2016). Investigating the relationship between individuals' level of hope and their thoughts about future by multinominal logistic regression analysis. *Dumlupınar University Journal of Social Sciences*, 47, 184-194.

- Artinger, F., Petersen, M., Gigerenzer, G., & Weibler, J. (2015). Heuristics as adaptive decision strategies in management. *Journal of Organizational Behavior*, 36 (S1), S33–S52.
- Avcı, E., Tuncel, O.Ç., & İnandı, Y. (2011). Ortaöğretim 12. sınıf öğrencilerinin matematiğe yönelik tutumları. *Journal of Mersin University Faculty of Education*, 7 (1), 50-58.
- Awad, M., & Khanna, R. (2015). *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*. New York: Springer.
- Baykal, A. (2006). Veri madenciliği uygulama alanları. *Journal of Dicle University Ziya Gökalp Faculty of Education*, (7), 95-107.
- Berry, M., & Linoff, G. (2004). *Data mining techniques for marketing, sales and customer support* (2nd ed.). New York: Wiley.
- Biggs, D., Ville, B., & Suen, E. (1991). A method of choosing multiway partitions for classification and decision trees. *Journal of Applied Statistics*, 18 (1), 49-62.
- Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984) *Classification and regression trees*. New York: Chapman and Hall.
- Breiman, L. (2001). Random forests. *Journal of Machine Learning*, 45 (1), 5-32.
- Brownlee J. (2013). *How to prepare data for machine learning*. Retrieved October 2020, from <https://machinelearningmastery.com/how-to-prepare-data-for-machine-learning/>,2013.
- Buse, A. (1982). The likelihood ratio, wald and lagrange multiplier tests: An expository note. *The American Statistician*, 36 (3), 153-157.
- Caetano S., Aires-de-Sousa, J., Daszykowska, M., & Vander Heyden, Y. (2005). Prediction of enantioselectivity using chirality codes and classification and regression trees. *Analytica Chimica Acta*, 544, 315–326.

- Carvalho, D.R., & Freitas, A.A. (2004). A hybrid decision tree/genetic algorithm method for data mining. *Journal of Information Science and Engineering*, 163 (1-3), 13-35.
- Ceylan, G. (2014). *Üniversite öğrencilerinin yansıtıcı düşünme düzeyleri ve araştırmaya yönelik kaygılarının çeşitli değişkenler açısından CART analizi ile incelenmesi*, Master's Thesis, Yüzüncü Yıl University, Van.
- Chang, T. S. (2011). A comparative study of artificial neural networks, and decision trees for digital game content stocks price prediction. *Expert Systems with Applications*, 38, 14846-14851.
- Changnon, D., Ritsche, M., Elyea, K., Shelton, S., & Schramm, K. (2000). Integrating climate forecasts and natural gas supply information into a natural gas purchasing decision. *Meteorological Applications*, 7, 211-216.
- Coussement, K., Van den Bossche, F. A. M., & De Bock, K. W. (2014). Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees. *Journal of Business Research*, 67 (1), 2751–2758.
- Coussement, P., Maertens, J., Beauprez, J., Van Bellegem W., De Mey M , (2014). One step DNA assembly for combinatorial metabolic engineering. *Journal of Metabolic Engineering*, 23, 0–77.
- Cox, M. & Ellsworth, D. (1997). Visualization proceedings. *IEEE Application-Controlled Demand Paging for Out-of-core*, 235-244.
- Daizaburo S. (2015). *Logistic regression in R*. Retrieved October 2020, from <https://sites.google.com/site/daishizuka/toolkits/plotting-logistic-regression-in-r>.
- Das, S., Dey, A., Pal, A., & Roy, N. (2015). Applications of artificial intelligence in machine learning: review and prospect. *International Journal of Computer Applications*, 115 (9), 31-41.

- Demirel, D. (2010). *Veri madenciliğinde CHAID algoritmasının sosyal güvenlik kurumu veri tabanına uygulanması*. PhD Thesis, Gazi University, Ankara.
- Deo, R.C. (2018). Machine learning in medicine. *PMC*, 132 (20), 1920–1930.
- Diab, D. L., Pui, S., Yankelevich, M., & Highhouse, S. (2011). Lay perceptions of selection decision aids and US and non-US samples. *International Journal of Selection and Assessment*, 19 (2), 209–216.
- Doğan, N. & Özdamar, K. (2003). CHAID analizi ve aile planlaması ile ilgili bir uygulama. *Journal of Basic Clinical Medical Sciences*, 23, 392-397.
- Doğan, Ş., & Türkoğlu, İ. (2007). Hypothyroidi and hyperthyroidi detection from thyroid hormone parameters by using decision trees. *Journal of Eastern Anatolia Region Studies*, 5 (2), 163-169.
- Elite Data Science. (2019), *Modern machine learning algorithms: strengths and weaknesses*. Retrieved November 25 2020, from <https://elitedatascience.com/machine-learning-algorithms>.
- Elhan, A.H. (1997). *Review of logistic regression analysis and implementation in medicine*. PhD Thesis, Ankara University, Ankara.
- Evans, B., & Fisher, D. (1994). Overcoming process delays with decision tree induction. *IEEE Expert*, 9, 60-66.
- Fırat, M. (2017). Understanding Turkish students' preferences for distance education depending on financial circumstances: A large-scale CHAID analysis. *International Review of Education*, 63 (2), 197-212.
- Foote, K. (2019). *A brief history of machine learning*. Retrieved October 15 2020, from <https://www.dataversity.net/a-briefhistory-of-machine-learning>.
- Freund, R. J., & Wilson, W.J. (2006). *Regression analysis: Statistical modeling of a response variable (2nd ed.)*. Burlington: Academic Press.
- Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 13 (2), 50-67.

- Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38 (4), 367-378.
- Gahi, Y., Guennoun, M. & Mouftah, H. T. (2016). Big data analytics: Security and privacy challenges. *IEEE Symposium on Computers and Communication*, 952-957.
- Gerritsen, R. (1999). Loan risks: A data mining case study. *IT Professional*, 1 (6), 16-21.
- Geron, A. (2019). *Hands on machine learning with scikit learn, keras, and tensorflow: Concepts, tools, and techniques to build intelligent systems (2nd ed.)*. Massachusetts: O'Reilly Media.
- Guido, J.S., Winters, P.C., & Rains, A.B. (2006). *Logistic regression basics*. New York: University Press.
- Gurel, P.A. (2018). Determination of factors that provide work-life balance for women academics: Logistic regression analysis. *Istanbul University Journal of Women's Studies*, 16, 31-44.
- Gülpınar, V. (2008). *Avrupa birliği ülkeleri ile Türkiye'nin ekonomik göstergelerinin karar ağacı yöntemi ile karşılaştırılması*. Master Thesis, Marmara University, İstanbul.
- Güner, Z.B. (2014). Veri madenciliğinde CART ve lojistik regresyon analizinin yeri: İlaç provizyon sistemi verileri üzerinde örnek bir uygulama. *Social Security Journal*, 6, 53-99.
- Hand, D.J. (1998). Data mining: Statistics and more?. *Journal of the American Statistical Association*, 52, 112-118.
- Han, D.J., & Kamber, M. (2000). *Data mining: concepts and techniques*, San Francisco: Morgan Kaufmann.
- Han, D.J., & Kamber, M. (2006). *Data mining: concepts and techniques (2nd ed.)*. San Francisco: Morgan Kaufmann.

- Hanımoglu, E. (2019). The analysis of the relation among internet addiction and social skills level with logistic regression analysis in primary education second stage students. *Kastamonu Education Journal*, 27 (2), 467-474.
- Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, 1 (3), 333–342.
- Hoare, R. (2004). Using CHAID for classification problems. In: *New Zealand Statistical Association Conference*. New Zealand: Wellington.
- Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: John Wiley & Sons Inc.
- Hsia, T. C., Shie, A.J., & Chen, L.C. (2008). Course planning of extension education to meet market demand by using data mining techniques – an example of Chinkuo technology university in Taiwan. *Expert Systems with Applications*, 34, 596-602.
- Hütter, M., & Ache, F. (2016). Seeking advice: A sampling approach to advice taking. *Judgment and Decision Making*, 11, 401-415
- Isi, A., Ozaydin, O., & Cemrek, F. (2014). Determination of factors affecting individuals investment behaviour using CHAID analysis, *Journal of Economics Finance and Accounting*, 1 (4), 273-284.
- Jacobs, P., (1999). Data Mining: What general managers need to know. *Harvard Management*, 4 (10),8-9.
- Kağnicioğlu, H. C., & Moğol, M. (2014). Implementation of CHAID algorithm: A hotel case. *International Journal of Research in Business and Social Science*, 3 (4), 42-52.
- Kashani, A. G., Olsen, M. J., Parrish, C. E., & Wilson, N. (2015). A review of LIDAR radiometric processing: From ad hoc intensity correction to rigorous radiometric calibration. *Sensors*, 15 (11), 28099-28128.

- Kass, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of Applied Statistics*, 20 (2),119-127.
- Katz, M.H. (1999). *Multivariable analysis: A practical guide for clinicians*. New york: Cambridge University Press.
- Kaygın, C. Y., Tazegül, A., & Yazarkan, H. (2016). Estimation capability of financial failures and successes of enterprises using data mining and logistic regression analysis. *Ege Academic Review*, 16 (1), 147-159.
- Kayri, M., & Boysan, M. (2007). Araştırmalarda Chaid analizinin kullanımı ve baş etme stratejileri ile ilgili bir uygulama. *Journal of Ankara University Faculty of Educational Sciences*, 40 (2), 133-149.
- Kayri, M., (2014). *Karar ağaçları*. Decision Trees Workshop, Alparslan University. Muş.
- Kayri, M., & Ceyhan, G. (2016). Investigating the effect of certain socio-demographic factors on university students' level of reflective thinking by using regression tree. *Anthropologist*, 24 (3), 694-701.
- Kıran Z. (2010). *An application on pharmacy provision system data of social security institution by logistic regression and CART analysis technics*. Master Thesis, Gazi University, Ankara.
- Korkmaz, M., Germir, H.N. & Gürkan, A. (2015). Yaşam memnuniyeti üzerine etkili olan sosyodemografik bileşenler üzerine bir analiz. *International Journal of Psychiatry and Psychology Research*, 3 (2), 78-111.
- Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13, 8–17.
- Köktürk, F. (2012). *Comparing classification success of knearest neighbor, artificial neural network and decision trees*. Master Thesis, Bülent Ecevit University, Zonguldak.

- Lewis, R.J. (2000). An introduction to classification and regression tree (CART) analysis. In *Annual Meeting of the Society for Academic Emergency Medicine*. California: San Francisco.
- Lim, T. S., Loh, W. Y., & Shih, Y.S. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning Journal*, 40, 203-228.
- Liu, Y., Lin, D., Xiao, T., Ma, Y., Hu, Z., Zheng, H., ... & Gao, Y. (2011). An immunohistochemical analysis-based decision tree model for estimating the risk of lymphatic metastasis in pN0 squamous cell carcinomas of the lung. *Histopathology*, 59 (5), 882–891.
- Loh, W.Y., & Shih, Y.S. (1997). Split selection methods for classification trees. *Statistica Sinica*, 7, 815-840.
- Loh, W.Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1 (1), 14–23.
- Marks, K. (2018). *Comparing the accuracy of decision trees and logistic regression in personnel selection*, Master Thesis, Middle Tennessee State University, ABD.
- Marr, B. (2016). *Key business analytics. The 60+ business analysis tools every manager needs to know*. United Kingdom: Pearson Education Limited.
- Michie, D., Spiegelhalter, D.J., & Taylor, C.C. (1997). *Machine learning, neural and statistical classification*. (131-174). New York: Prentice Hall.
- Milanović, M., & Stamenković, M. (2016). CHAID decision tree: Methodological frame and application. *Economic Themes*, 54 (4), 563-586.
- Miller, T. R., Wiek, A., Sarewitz, D., Robinson, J., Olsson, L., Kriebel, D., & Loorbach, D. (2014). The future of sustainability science: A solutions-oriented research agenda. *Sustainability Science*, 9 (2), 239–246.
- Mitra, S., & Acharya, T. (2003). *Data mining multimedia, Soft Computing and Bioinformatics*. Canada: New Jersey.

- Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of machine learning* (2nd ed.). London: The MIT Press.
- Moon, T.H., & Sohn, S.Y. (2005). Intelligent approach for effective management of governmental funds for small and medium enterprises, *Expert Systems with Applications*, 29, 566-572.
- Morgan, J.N., & Sonquist, J.A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58, 415–434.
- Morris, J. A., & Gardner, M. J. (1988). Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. *British Medical Journal (Clinical Research Edition)*, 296 (6632), 1313-1316.
- Murphy, K.P. (2012). *Machine learning: A probabilistic perspective. Adaptive computation and machine learning series*. London: The MIT Press.
- Murthy, S.K. (1998) Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery*, 24, 345-389.
- Mutluer, C., Buyukkıdık, S. (2017). Estimation on the mathematics literacy with logistic regression according to PISA 2012 data. *Marmara University Atatürk Education Faculty Journal of Educational Sciences*, 46, 97-112.
- Myatt, G. J. (2007). *Making sense of data a practical guide to exploratory data analysis and data mining*. Canada: John Wiley & Sons.
- Oluk, E. (2019). *Machine learning concept and usage areas*. Retrieved December 22 2020, from <https://www.gazetebilkent.com/2019/04/13/makine-ogrenimi-kavrami-vekullanım-alanlari-nursen-erginsoy/>, 2019.
- Oğuzlar, A. (2004). CART Analizi ile hane halkı işgücü anketi sonuçlarının özetlenmesi. *Ataturk University Journal of Economics and Administrative Sciences*, 18 (3-4), 79-90.

- Önder, E., & Uyar, S. (2017). CHAID analysis to determine socioeconomic variables that explain students' academic success. *Universal Journal of Educational Research*, 5 (4), 608–619.
- Öğüdücü, G. Ş., (2008). *Data mining basic classification methods*. Retrieved December 12 2020, from <http://ninova.itu.edu.tr/tr/dersler/bilisim-enstitusu/195/bbl-606/ekKaynaklar?g29172>.
- Palechor, M., Fabio, E., Colpas, A., Paola, P., Ojeda, S., Jorge, A., Manotas, A., & Melo, P.M. (2016). Fertility analysis method based on supervised and unsupervised data mining techniques. *International Journal of Applied Engineering Research*, 11 (21), 10374-10379.
- Park, H. (2013). An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain. *Journal of Korean Academy of Nursing*, 43 (2), 154-164.
- Pehlivan, G. (2006). *Chaid analysis and an application*. Master Thesis, Yıldız Teknik University, İstanbul.
- Priyadharshini, A, & Sundarambal, M. M. (2017). Spectrum sensing using hybrid gravitational search genetic algorithm based predictor model in cognitive radio networks. *SSRN Electronic Journal*, 330-346.
- Provost, F., & Danyluk, A. (1999). Problem definition, data cleaning and evaluation: A classifier learning case study. *Informatica*, 23, 123-136.
- Quinlan J.R. (1986). Induction of decision trees. *Machine Learning*, 1 (1), 81-106.
- Quinlan, J.R. (1987). Simplifying decision trees. *International Journal of Man-Machine Studies*, 27 (3), 221-234.
- Quinlan, J.R. (1990). Decision trees and decision-making. *IEEE Transactions on Systems*, 20 (2), 339-346.
- Quinlan, J.R. (1993). C4.5: Programs for machine learning. *Machine Learning*, 16, 235-240.

- Rodriguez, J., Kuncheva, L., & Alonso, C. (2006). Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28 (10), 1619–1630.
- Sajda, P. (2006). Machine learning for detection and diagnosis of disease. *Annual Review Biomedical Engineering*, 8, 537-565.
- Schapire, R. (2008), *Theoretical machine learning*. Retrieved September 25 2020, from http://www.cs.princeton.edu/courses/archive/spr03/cs511/scribe_notes/0204.pdf, 2008.
- Shafer, J .C., Agrawal R., & Mehta, M. (1996). A scalable parallel classifier for data mining. *In: Proc. 22nd International Conference on very large databases*. India: Mumbai.
- Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. New York: Cambridge University Press.
- Shutz (2017). *Machine learning: the power and promise of computers that learn by example The Royal Society*. Retrieved December 20 2020, from <https://royalsociety.org/~media/policy/projects/machinelearning/publications/machine-learning-report.pdf>.
- Shih, Y.S. (1999). Families of splitting criteria for classification trees. *Statistics and Computing*, 9, 309-315.
- Shih, Y.S. (2004). A note on split selection bias in classification trees. *Computational Statistics and Data Analysis*, 45, 457-466.
- Silahtaroglu, G. (2013). *Veri madenciliği kavram ve algoritmaları*. İstanbul: Papatya Yayıncılık.
- Sinha, A. P., & May, J. H. (2004). Evaluating and tuning predictive data mining models using receiver operating characteristic curves. *Journal of Management Information Systems*, 21 (3), 249–280.

- Smola, A., & Vishwanathan, S.V.N. (2008). *Introduction to machine learning*. New York: Cambridge University Press.
- Stauffer, J. M., Ree, M. J., & Carretta, T. R. (1996). Cognitive-components tests are not much more than g: An extension of Kyllonen's analyses. *The Journal of General Psychology*, 123, 193-205.
- Stewart, R. W., Tuerk, P. W., Metzger, I. W., Davidson, T. M., & Young, J. (2016). A decision-tree approach to the assessment of posttraumatic stress disorder: Engineering empirically rigorous and ecologically valid assessment measures. *Psychological Services*, 13 (1), 1–9.
- Şehribanoğlu, S., & Diler, S. (2018). 2013 yılı yaşam memnuniyet araştırmasının CART ve CHAID algoritmaları ile incelenmesi. *The Journal of Academic Social Science*, 6 (67), 132-141.
- Thomas, L.C. (2000). A survey of credit and behavioral scoring: Forecasting financial risk of lending to consumer. *International Journal of Forecasting*, 16 (2), 149–172.
- Thomas, E., H., & Galambos, N. (2004). “What satisfies students?” Mining student – Opinion data with regression and decision tree analysis. *Research in Higher Education*, 45 (3), 251-269.
- Ugurlu, O., Yildiz, S., Loughney, S., & Wang, J. (2018). Modified human factor analysis and classification system for passenger vessel accidents (HFACS-PV). *Ocean Engineering*, 161, 47-61.
- Ünlüsavuran, M. (2019). *Performance evaluation of automated machine learning algorithmsin omics data*. Master Thesis, Erciyes University, Kayseri.
- Vahaplar, A. (2003). *A geographic data mining application*. Master Thesis, Ege University, İzmir.
- Vupa, Ö. (2004). *Model building of logistic regression*. Phd Thesis, Dokuz Eylül University, İzmir.

Waheed, T., Bonnell, R.B., Prasher, S.O., & Paulet, E. (2006). Measuring performance in precision agriculture: CART- A decision tree approach. *Agricultural Water Management*, 84 (1-2), 173-185.

Witten, I.H. & Frank, E., (2000). *Data mining, Practical Machine Learning Tools and Techniques with Java Implementations*. San Fransisco: Morgan Kaufman Publishers.

Wu, M.C., Lin, S.Y., & Lin, C.H. (2006). An effective application of decision tree to stock trading. *Expert Systems with Applications*, 31, 270-274.

Yang, B.S., Lin, D.S., Chit, A., & Tan, C. (2005). VIBEX: An expert system for vibration fault diagnosis of rotating machinery using decision tree and decision table. *Expert Systems with Applications*, 28, 735-742.