DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

A DECISION SUPPORT SYSTEM SOFTWARE FOR SMOKING CESSATION PATIENTS

by
Yağmur KARAKOÇ

October, 2019
İZMİR

A DECISION SUPPORT SYSTEM SOFTWARE FOR SMOKING CESSATION PATIENTS

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

by
Yağmur KARAKOÇ

October, 2019
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"A DECISION SUPPORT SYSTEM SOFTWARE FOR SMOKING CESSATION PATIENTS"** completed by **YAĞMUR KARAKOÇ** under supervision of **PROF. DR. ALP KUT** 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.

Prof. Dr. Alp KUT

Supervisor

Assoc. Prof. Dr. Derya BİRANT

(Jury Member)

Doç. Dr. Tuğba ÖZACAR ÖZTÜRK

(Jury Member)

Prof.Dr. Kadriye ERTEKİN

Director
Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

I would like to thank my esteemed teacher Prof. Dr. Alp KUT for helping me with my work.

I would like to thank Prof. Dr. Vildan MEVSİM for sharing the data set of smoking cessation patients I have analyzed.

Yağmur KARAKOÇ

A DECISION SUPPORT SYSTEM SOFTWARE FOR SMOKING CESSATION PATIENTS

ABSTRACT

Cigarette is a substance that causes great harm to human and environment. Smoking and exposure to cigarette smoke is one of the factors that can cause many diseases. Despite these damages caused by smoking, smoking continues to increase without slowing down. If it continues to increase at this rate, there will be great harm to both humanity and the environment. It is now reported that approximately 5 million people have died due to smoking. It is estimated that approximately 8 million people will die in 2030 if they continue at this rate. Many studies have been carried out to prevent these health problems. With the developing information technologies, solutions are sought for the existing problems in the field of health. Data mining algorithms and decision support systems are one of the methods that can be used in this field. In this study, it is predicted that a decision support system that will be formed by using data mining algorithms will help health. In the study, a web application was made by using open source data mining library. Thus, with this application, it is aimed to make a decision support system that can be used by healthcare professionals in estimating the probability of patients quitting smoking.

Keywords: Cigarette addiction, data mining, decision support systems

SİGARA BIRAKMA HASTALARINA YÖNELİK BİR KARAR DESTEK SİSTEMİ YAZILIMI

ÖZ

Sigara, insana ve çevreye büyük ölçüde zarar veren bir maddedir. Sigara ve sigara dumanına maruz kalmak birçok hastalığa neden olabilecek faktörlerden birisidir. Sigaranın vereceği bu zararlara karşın sigara kullanımı hız kesmeden artmaya devam etmektedir. Bu hızla artmaya devam etmesi durumunda hem insanlığa hem de çevreye büyük zararlar meydana gelmesi söz konusudur. Şimdilerde sigara kullanımına bağlı olarak yaklaşık 5 milyon kişinin hayatını kaybettiği raporlanmıştır. Bu hızla devam etmesi durumunda 2030 yılında yaklaşık 8 milyon kişinin sigara nedeniyle hayatını kaybedeceği tahmin edilmektedir. Bu sağlık sorunlarının önüne geçmek amaçlı birçok çalışma yapılmaktadır. Gelişen bilişim teknolojileri ile sağlık alanındasorunlara çözümler aranmaktadır. Veri madenciliği algoritmaları ve karar destek sistemleri bu alanda kullanılabilecek yöntemlerden birisidir. Bu çalışma kapsamında veri madenciliği algoritmaları kullanılarak oluşturulacak bir karar destek sisteminin sağlık konusuna yardımcı olacağı öngörülmüştür. Çalışmada açık kaynak veri madenciliği kütüphanesi kullanılarak bir web uygulaması geliştirilmiştir. Böylece bu uygulama ile hastaların sigarayı bırakma olasılığının tahmini konusunda sağlık çalışanlarının kullanabileceği bir karar destek sistemi yapmak amaçlanmıştır.

Anahtar Kelimeler: Sigara bağımlılığı, veri madenciliği, karar destek sistemleri

CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER ONE
INTRODUCTION

1.1 General

Bad habits are habits that blind our decision making ability and prevent us from having healthy thoughts and actions. These habitual substances are generally used for try purposes. There are many varieties of these substances that become addictive over time, and they have many harmful effects on human health. The most known of these habits is smoking. Smoking is known as one of the most threatening addictions in the world and in our country.

It was first used in the USA in the 1500s for pleasure and continued to be consumed with the idea that it was good for his headache and thus spread throughout the world. The fact that cigarettes become such a strong addiction is due to the high level of nicotine. Today, it is known that approximately half of the adult population smokes and the smoking rate gradually increases. In addition, the age of use is gradually decreasing due to the desire to experiment with curiosity. Therefore, it has a great impact on human health at a young age. According to the reports, Turkey ranks 11th in the world with 23.8 percent smoking rate.

Cigarette contains too many chemicals. Therefore, both smoking and exposure to cigarette smoke; can cause many dangerous diseases such as cancer, heart diseases, COPD and can deaths on result. It is estimated that around 5 million people die every year from diseases related to smoking. According to the World Health Organization's report, if tobacco and cigarettes continue at this rate; It is predicted that in 2030, 8 million people will die from fatal diseases (WHO, 2009). In the field of health, data mining algorithms are widely used in many fields such as cancer, heart and kidney diseases. In the literature, no study was done with data mining about smoking cessation.

1.2 Purpose

In this study, data gathered from the smoking cessation unit of İzmir Dokuz Eylül Hospital were used. These data are the first and only data set collected in our country in the field of smoking cessation. Currently, the physicians of our university consider the actual data collected on the subject from the SPSS system. This requires both SPSS license and usage skill. In such cases, it is not easy to store patient data, perform analyzes, and predict patients release. In order to eliminate these disadvantages, it is aimed to develop a web application that will facilitate the work of health workers. Firstly, the data mining algorithm, J48, was used to analyze the data set. In the light of these analyzes, an application was developed in which health workers predicted the probability of smoking cessation.

The most important aim of the study is to help health personnel to estimate the probability of smoking cessation with an easy-to-use web application and open-source data mining library.

1.3 Content of the Thesis

The thesis consists of nine chapters. These sections are as follows.

In Chapter 2, data mining algorithms and decision support systems made studies related are described.

In Chapter 3, used technologies are described.

In Chapter 4, the proposed project is described.

In Chapter 5, developing the decision support system software with vısual is described.

In Chapter 6, conclusion and future works are described.

CHAPTER TWO
LITERATURE REVIEW

Within the scope of the study, a decision support system application was made with the help of data mining algorithm. In this context, studies on data mining algorithms and decision support systems in health will be explained.

2.1 Data Mining Algorithms

In the studies of Asha Rajkumar and G.Sophia Reena, data mining algorithms have been used to estimate the risk of heart disease. From these algorithms, Naive Bayes algorithm is have been give result better than other algorithms (Rajkumar & Reena, 1010).

The studies of Divya Tomar, Sonali Agarwal have been given information about the use of various data mining techniques such as classification, clustering, association, regression in health field. In addition, the challenges of data mining practices in health services are described and the recommendations for the selection of appropriate data mining techniques are given (Tomar & Agarwal, 2013).

Ali Koyuncugil and Nermin Özgülbaş, aim to provide a new perspective on the use of Data Mining in health and to provide healthcare professionals with examples of data mining in the healthcare sector (Koyuncugil & Özgülbaş, 2009).

Pınar Yıldırım, Mahmut Uludağ and Abdülkadir Görür conducted a study on data mining in the field of health. It has been emphasized that important results can be obtained by analyzing the hidden data in large data sets stored in hospital information systems. In addition, studies on data mining in the fields of health and medicine are explained. Thus, the importance of the discovery of confidential information is explained (Yıldırım, Uludağ & Görür, 2008).

2.2 Decision Support Systems

Dr. Tang PC and McDonald CJ have been a studied on computer-based patient recording systems and computer applications in the field of health. In this study, the differences between the storage of medical records as electronic or document are explained and the benefits of electronic storage as well as functionality such as ease of reliability are mentioned (Tang & Mcdonald, 2001).

Prof. Dr. Musa Özata and Şebnem Aslan have tried to introduce the clinical decision support systems that increase the effectiveness of the decision that will support the decision-making process of the patient and give information about the sample applications used in various countries of the world (Özata & Aslan, 2004). Some of these examples include:

MYCIN:

This system is developed at Stanford University in the 1970s. Diagnosis of certain blood infections and detection of treatment methods is the purpose of developing the system. It is intended to make estimates that doctors do roughly but are very important. The idea that an expert system could help in finding a more effective treatment method, and as a result, MYCIN was developed. This system looks for answers to questions such as what tests should be done, what method of treatment should be or how a treatment plan should be performed (Yıldırım, 2000).

DeDOMBAL'IN LEEDS ABDOMINAL PAIN SYSTEM:

In the late 1960s, De Dombal and his associates at the University of Leeds developed a computer-based decision-assistance system that explored the symptom process of abdominal diseases using Bayesian probability theory. Using the sensitivity and sensitivity properties of the system, the various signs of the disease, symptoms, findings and test results based on the results of the calculations of appendicitis, peptic ulcer, diverticulus, gallbladder pain, pancreas, small intestine problems and the cause of unexplained abdominal pain has revealed the diagnosis of the disease (Musen, Yuval & Shortliffe, 2003).

HELP:

HELP hospital information system in LDS hospital is a medical information system that have been active since 1975. There are sub-units such as Blood Ordering System, Antibiotic Use Detection System.

Blood Ordering System: Provides advice on whether or not to apply the blood product to the patient. It also aims to identify patients who need a blood transfusion and inform the relevant staff.

Antibiotic Usage Detection System: This system helps to collect clinical data on a report and to create information that requires decisions that require antibiotic treatment (Haug, Rocha & Scott, 2003).

DxPLAİN:

This application determines the diseases that may occur based on clinical findings. Thus, the probability of occurrence of diseases is estimated according to the findings (Bilgen, 2002).

ISABEL:

Isabel is a decision support system used by doctors in hospitals in the UK. It is a system that can help make possible diagnoses about diseases. The Isabel system was introduced for the pediatric patients. It is also intended to be used for the diagnosis of diseases of adult patients in the future.

POEMS (Post Operative Expert Medical System):

This system has been developed to assist inexperienced healthcare professionals after patients' operations. It is also used in the diagnosis of possible disease conditions by storing the patient's family history and disease history (Savar, Brennan &Cole, 1991).

CHAPTER THREE
USED TECHNOLOGIES

3.1 Data Mining

Data mining is the task of obtaining the necessary information from a large amount of data. Thus, it is possible to make predictions for the future by obtaining the relationships between the data. Today, data mining applications are used in almost every field. For example, marketing, banking, education, security, health and so on.

3.1.1 *Data Mining Process*

In order to make accurate predictions as a result of data mining applications, a certain process must be followed (Bharati & Ramageri, 2010).

This process consists of the following steps:

- Cleaning,
- Integration,
- Reduction,
- Conversion,
- Implement data mining algorithms,
- Results and forecasts.

Figure 3.1 shows schema that data mining process.

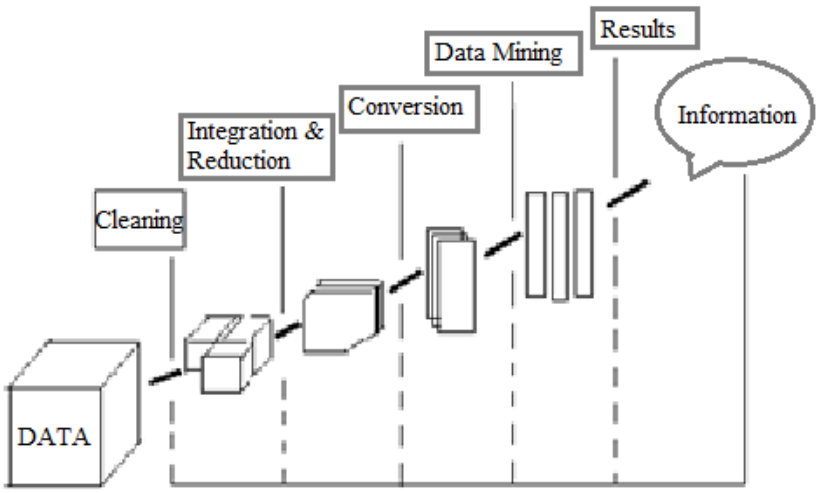

Figure 3.1 Data mining process

3.1.1.1 Data Cleaning

Sometimes the data to be researched consists of incomplete or inappropriate data. In such cases, the data is customized. If there is a missing value in the data set, these records are discarded from the data set or the data is edited according to the desired properties using some statistical methods.

3.1.1.2 Data Integration

It is the conversion of data obtained from different data sources into one type.

3.1.1.3 Data Reduction

It takes a long time to process with very large data sets. Therefore, if it is believed that the result will not change, the reduction process is performed by reducing the number of data or variables that are considered unnecessary in the data set.

3.1.1.4 Data Conversion

In some cases, the original form of the data is not suitable for the process. In order to make the data suitable for the study, some statistical methods are applied on these data. This makes the data useful.

3.1.1.5 Application of Data Mining Algorithms

Data mining algorithms are applied after applying the above mentioned items as required. These algorithms are based on methods called clustering, classification and association rules.

3.1.1.6 Results and Estimates

Following the implementation of all steps, the results are interpreted and predicted for the future.

3.1.2 Data Mining Methods

There are methods used for data mining and many algorithms developed in this direction. Most of these methods are statistics-based. Classification, clustering and association rules are the methods used in data mining.

3.1.2.1 Classification Method

It is used to reveal confidential information in the data set. A certain process is followed in order to apply the classification method. First, some amount of data in the data set is used as the training set. Then, with the help of learned information, estimations are made on the remaining data set and inferences are made. Thus, when a new record arrives, it is possible to make predictions about the result. Classification method is widely used in many areas such as voice recognition, credit application evaluation, character recognition and disease diagnosis.

3.1.2.1.1 J48 Algorithm (C4.5). J48 is an open source version of C4.5 in the WEKA program. It is a common decision tree learning algorithm based on the entropy used for classification problems.

Entropy: A measure of the uncertainty of a random variable. Calculated by the formula 3.1. The uncertainty of a variable with high entropy is also high.

$$H = -\sum p_i (\log_2 pi) \qquad (3.1)$$

where H is entropy value of calculated by each p_i probability values.

Knowledge Gain: A measure of how much uncertainty changes in the target variable when data is split using a predictive variable.

The model is obtained as a result of the classification process shows as a tree and is called the decision tree. Each attribute is represented by a node. Branches and leaves are elements of the tree structure. The last structure is called the leaf, the upper structure root and the structures that remain between them. The data should be categorical (Frank at al., 2010).

Steps of Algorithm:

- The target class is determined and the entropy is calculated,
- In view of this entropy value, earnings criteria are determined for each attribute,
- The predictive variable that provides the highest information gain is determined and the tree starts branching from this variable,
- Thus, the data will be distributed evenly under each branch,
- This process continues until all predictive variables are placed in the tree,
- After the first predictive variable is detected, the same process continues and the tree is created according to the most information acquisition (Yalçın, 2013).

Example:

Table 3.1 includes a set of education with nominal values. The entropy of the class variable is calculated. Then the entropy values of each attribute are calculated. The difference between the entropy value of the attribute and the gain gives the entropy value of the class variable. These operations are repeated for each attribute. The biggest gain is the first branch. These steps are repeated for the other branches and as a result the decision tree in Figure 3.2 is obtained.

Table 3.1 Education set

Attribute1	Attribute2	Attribute3	Class
a	<=	True	Class1
a	>	True	Class2
a	>	False	Class2
a	>	False	Class2
a	<=	False	Class1
b	>	True	Class1
b	<=	False	Class1
b	<=	True	Class1
b	<=	False	Class1
c	<=	True	Class2
c	<=	True	Class2
c	<=	False	Class1
c	<=	False	Class1
c	>	False	Class1

$H(ATTRIBUTE2_{<=}) = -[(7/9) * \log_2 (7/9) + (2/9) * \log_2 (2/9)]= 0.765$

$H(ATTRIBUTE2_{>}) = -[(2/5) * \log_2 (2/5) + (3/5) * \log_2 (3/5)]= 0.971$

$H(ATTRIBUTE2, CLASS) = 9/14 * H(ATTRIBUTE2_{<=}) +5/14* H(ATTRIBUTE_{>})$

$= 9/14 * 0.765 +5/14*0.971 = 0.836$

$GAIN(ATTRIBUTE2, CLASS) = 0.940 – 0.836 = 0.104$

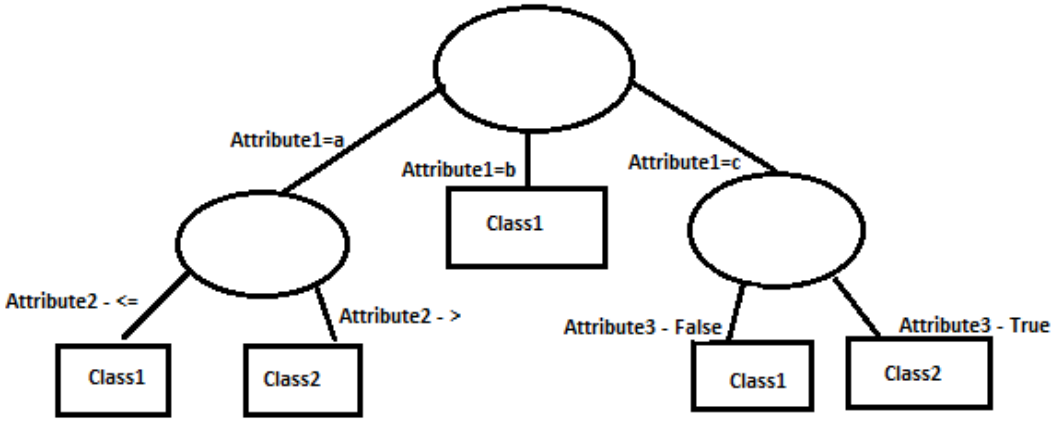

Figure 3.2 Decision tree

3.2 Weka Software

WEKA is one of the data processing packages used for data mining and data mining. It vas developed in JAVA language as on open source at University of Waikato. Distributed under GNU General Public Licence. The name comes from here and consist of the initial of the words Waikato Environment for Knowledge Analysis.

WEKA reads data from a simple file called. .ARFF format. It deals with numerical or nominal data. A library for data mining and statistics is available in WEKA. For example, data preprocessing, classification, clustering and visualization of results are possible with WEKA software (WEKA, 2019). When the WEKA software is opened, the input panel in Figure 3.3 is displayed.

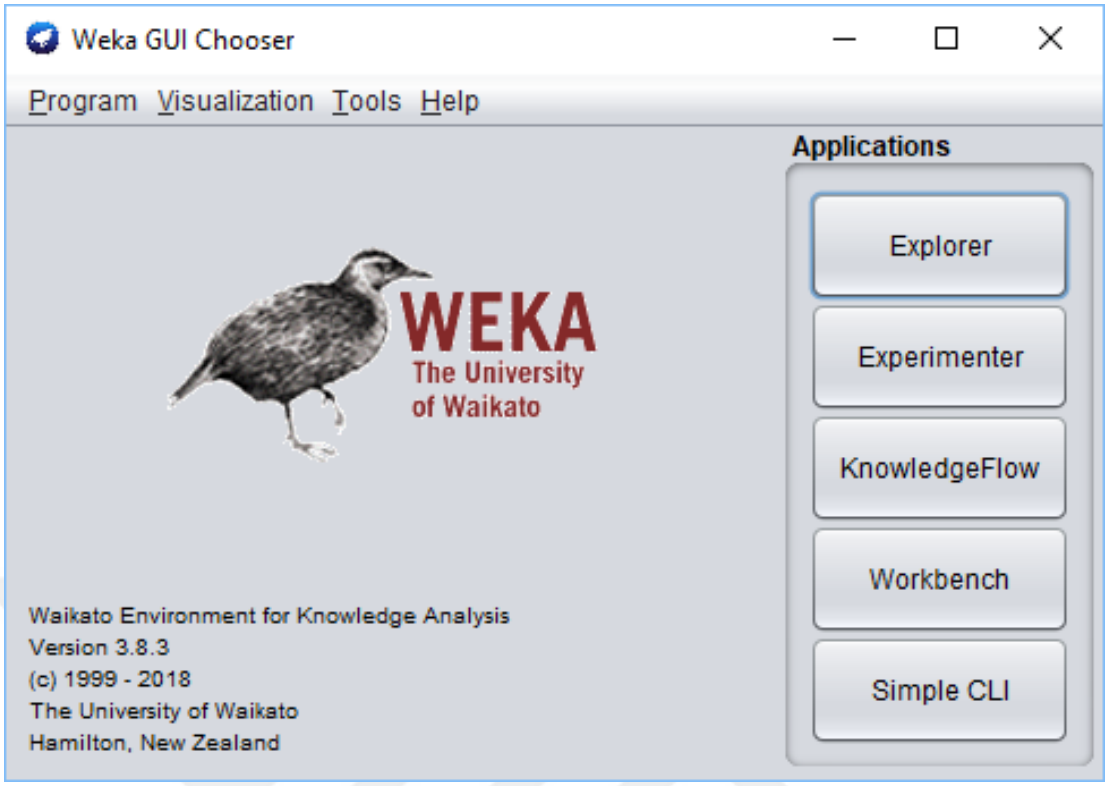

Figure 3.3 Weka input panel

WEKA includes preprocessing, classification - clustering - associate, select attributes and a visualize panel.

3.2.1 Weka Software Panels

3.2.1.1 Preprocess Panel

Preprocessing panel is the panel where data can be loaded and necessary processing can be performed. This panel provides information on the properties of the data, and makes adjustments to which attributes to use. Figure 3.4 shows the screen shot of this panel.

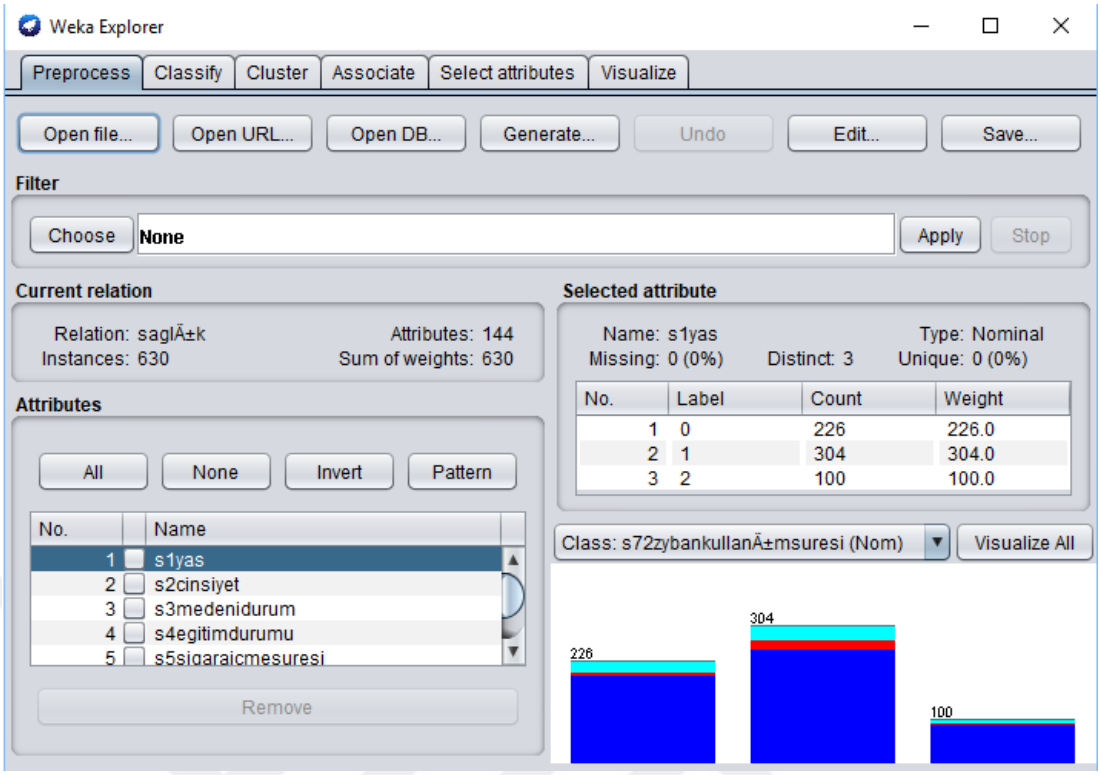

Figure 3.4 Classification of data

3.2.1.2 Classifier Panel

In this panel, is selected which classification method to use, test options, and which variable to test. In the test option section, it can be determined what percentage of the data set is used for training and what percentage is used for the test process. The results are shown in the classifier output window. Here, statistical results, accuracy rate and error rate can be accessed. Classification panel is shown in Figure 3.5.

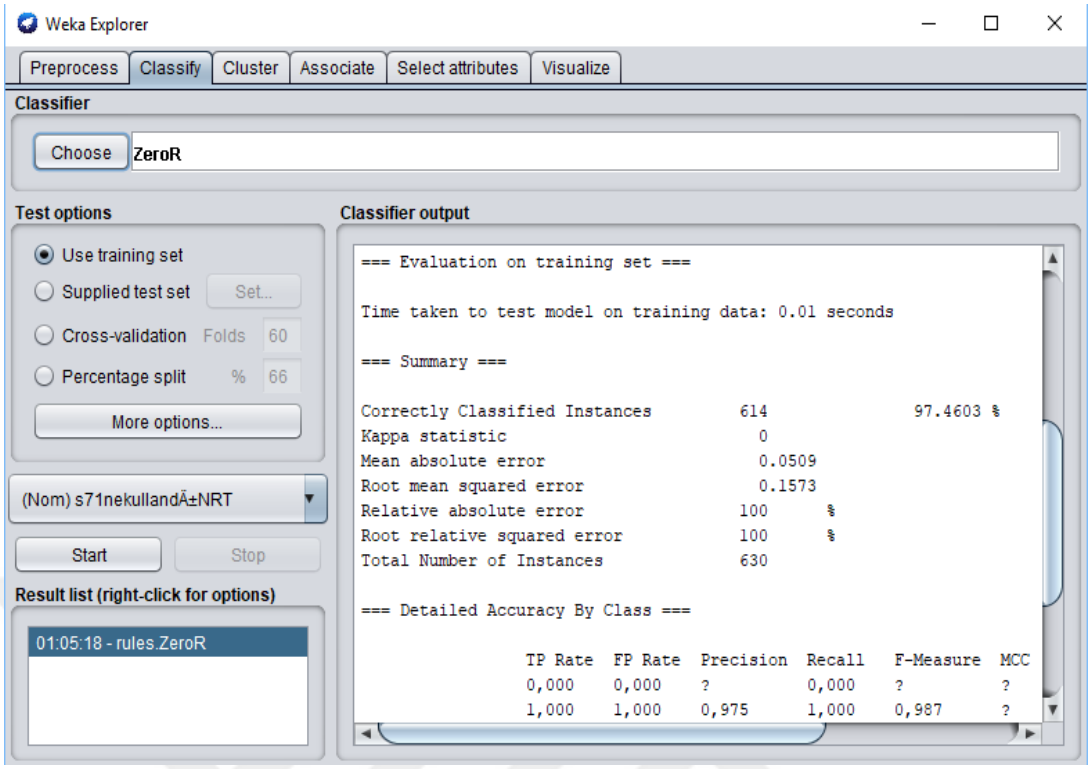

Figure 3.5 Classification panel

3.2.1.3 Cluster Panel

The clustering panel is showed which clustering algorithm to use, the clustering mode and the results obtained. In the conclusion section; is show that how many clusters are obtained, properties of clusters and accuracy rates. The example screenshot is shown in Figure 3.6.

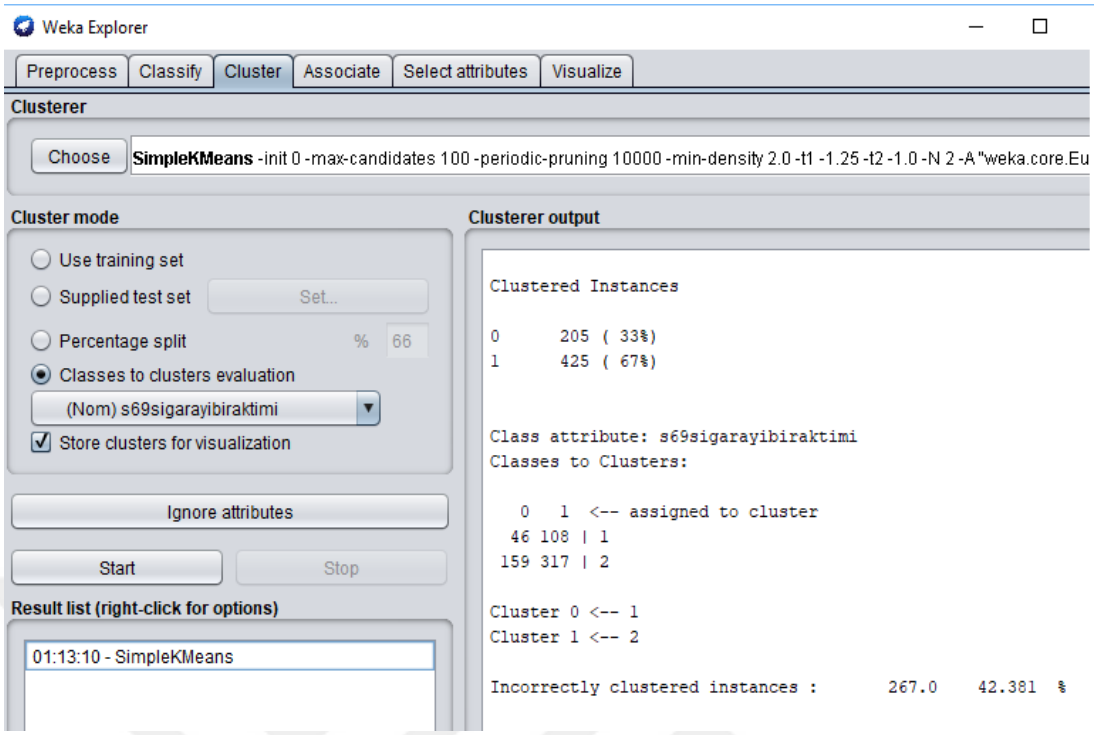

Figure 3.6 Clustering panel

3.2.1.4 Associate Panel

Associate panel; the top of the screen is selected which method to use, and the results are displayed on the screen.

3.2.1.5 Select Attributes Panel

It is used to set the selection and processing properties on the data set. If one of the selection schemes converts the data, the transformed data can be viewed on the visualization screen.

3.2.1.6 Visualize Panel

This panel shows the data set. The size of the cells and the dots, the number of cells on the matrix can be adjusted. In addition, when working with very large data sets, it is also possible to use only sub-sample space for ease of operation.

3.2.2 ARFF Format

The .ARFF format is a data format used for scientific purposes. The most important advantage is that it can be used with Weka software. The .ARFF file format is converted from the .CSV friend format.The .CSV file format separates data with a comma, provide them to be in a specific order. The .CSV file format is easily obtained with the help of Excel. Data stored in Excel is saved in .CSV format and converted to this format. When the saved .CSV format is opened with the help of a text editor, it provides the data property of the .ARFF format.

The .ARFF file consists of @relationship, @ attribute, and @data. Relation is the area where we need to give our data community a name. Attribute is the data type used.

NUMERIC: They are numerical values.

REAL: Contains all real numeric values.

STRING: Textual values (often used in Text Mining).

NOMINAL: They are cluster values.

DATE: The values are date.

DATA: A group of existing data.

After this process, the information that should be defined in .ARFF format is defined by the @ sign. Name of the data set, properties and class information, if any, are written. An example format is shown below.

Example:

@RELATION data

@ATTRIBUTE dt DATE “yyyy-mm-dd”

@ATTRIBUTE age NUMERIC

@ATTRIBUTE length NUMERIC

@ATTRIBUTE weight NUMERIC

@ATTRIBUTE gender{0, 1}

@ATTRIBUTE class {Class1, Class2}

@DATA

2000-04-15, 26, 165, 55, 0, Class1

2001-09-23, 20, 185, 80, 1, Class1

2001-12-12, 55, 160, 70, 0, Class2

2001-01-01, 33, 170, 65, 0, Class1

3.3 SPSS

Statistical Package for the Social Sciences is abbreviation for SPSS. It is a computer program which is used frequently in social sciences, educational sciences, health sciences and natural sciences. The program runs on Windows and Mac computers. It has a similar appearance to Microsoft Excel.

In academic studies, SPSS is mostly used in survey-related studies, demographic characteristics of participants, frequency and percentage calculations, reliability, correlation, regression and difference analysis (Landau & Everit, 2014).

Example:

- Finding the reliability coefficient of the questionnaire used in the research (such as Cronbach alpha),
- Factor analysis,

- Analysis of participants according to variables such as gender, graduated school, marital status,
- Displaying these with tables, bell curve, graph,
- Correlation between the two data, i.e. the determination of the relationship (such as Pearson r),
- Finding the regression equation,
- Determining whether any data differs statistically between men and women or according to the school completed (tests such as t-test, ANOVA, Mann Whitney U, Kruskal Wallis, Chi-square).

3.4 Visual Studio

3.4.1 What is Visual Studio?

Visual Studio is an integrated development program using many programming languages. Visual Studio is used in many areas such as web application, web site, web service, mobile application development.

Microsoft software development platforms such as Visual Studio, Windows API, Windows Forms, Windows Presentation Foundation, Windows Store and Microsoft Silverlight are used. It can generate both native code and managed code.

Visual Studio includes IntelliSense (code completion component) and a code editor that supports code reorganization. The integrated debugger works as both a resource level debugger and a machine level debugger. Other built-in tools include a code profile builder, form designer, web designer, class designer, and database schema designer for building GUI applications. Add-ons that increase functionality at almost every level are accepted.

Different libraries are also available in Visual Studio. For example, when the library of Weka software is added, it is possible to perform all the functions of the

Weka software in the Visual Studio environment. This makes it possible to carry out more professional studies in terms of software (ASP.NET MVC Pattern, n.d.).

3.4.2 Languages Supported by Visual Studio

Visual Studio supports different programming languages. C, C ++ and C ++ / CLI (with Visual C ++), VB.NET (with Visual Basic .NET), C # (with Visual C #), F # (from Visual Studio 2010)) and TypeScript (as of Visual Studio 2013 Update 2).

In addition to supporting Python, Ruby, Node.js, and M, other languages can be used by installing the required services. It also supports XML / XSLT, HTML / XHTML, JavaScript and CSS.

CHAPTER FOUR
THE PROPOSED PROJECT

The J48 algorithm is an algorithm of the classification method. This algorithm makes more accurate analyses with nominal data types. Therefore, numerical data types have been converted to nominal data type using some statistical methods.

4.1 Data Collection, Regulation and Create Decision Tree

In İzmir Dokuz Eylül Hospital Family Medicine Quit Smoking Unit; A total of 145 data were stored on the SPSS, including personal information such as age, gender, marital status, cause of onset, family history, smoking cessation, hospital records, laboratory results, the duration of smoking cessation medication, and whether or not to quit smoking. This data set consists of 22 numeric and 123 nominal data types. Figure 4.1 shows some of these data.

	s3yas	s5cinsiyet	s6medenidurum	s7egitimdurumu	s9sigaraicmesuresi	s10gundeiçilensigaramiktari	s11sigaratuketimipaketyili	s12sigarabaslamayasi	s13sigarabaslamanedenimerak	s13sigarabaslamanedeniozenti	s13sigarabaslamanedenipsikolojikbaski	s13sigarabaslamanedeniysagatepki
1	48	1	1	4	24	20	24,00	21	1	2	2	2
2	37	2	1	5	14	10	7,00	14	1	1	2	2
3	34	2	1	6	17	10	15,00	14	1	2	2	2
4	53	2	1	4	35	20	35,00	15	1	1	2	2
5	46	2	1	6	20	30	30,00	15	2	1	2	2
6	45	1	2	3	24	60	72,00	18	2	1	2	2
7	69	2	1	6	52	10	26,00	13	1	1	2	2
8	39	2	1	5	9	10	4,50	18	1	2	2	2
9	58	1	2	6	40	20	40,00	15	1	1	1	1
10	48	1	1	3	25	30	25,00	15	2	2	1	2
11	30	2	2	6	10	20	10,00	17	1	2	2	2
12	44	2	1	6	20	20	20,00	18	2	1	2	2
13	40	2	2	6	21	20	21,00	16	2	1	2	2
14	56	2	1	6	34	40	68,00	18	2	1	2	2
15	25	2	2	6	6	30	9,00	15	1	1	2	2
16	38	1	2	6	12	20	12,00	19	2	1	2	2
17	26	1	1	5	10	10	5,00	17	1	1	2	2
18	44	1	1	4	20	15	15,00	21	2	2	2	2
19	42	2	1	6	23	20	23,00	18	2	1	2	2
20	56	1	3	4	22	20	27,00	28	2	2	2	2
21	30	2	1	6	11	20	11,00	18	1	2	2	2
22	29	2	2	5	15	40	30,00	13	1	1	2	2
23	53	2	1	3	22	25	24,00	21	2	1	2	2

Figure 4.1 Sample data

The range value of the data with numerical data hascalculated using by SPSS. Figure 4.2 and Figure 4.3 show SPSS result. This is divided into 3 categories according to the range value. The laboratory results of the patients were categorized according to their actual values. Thus, the data are converted to nominal data type.

The missing data in the data set is completed with the mode values after being converted to the nominal data type. Table 4.1 and Table 4.2 show the conversion of the data. Figure 4.4 shows the final state of the converted data.

Statistics

		yas	sigaraicmesuresi	sigaramiktari	sigaratuketimipaketyil	sigarabaslamayasi	nekadarsurebiraktinizkacay	evdesigaraicmeyipdumanamaruzkalankackisivar	isyerindegundekacsigaraiciyor
N	Valid	630	630	630	630	630	211	509	440
	Missing	40	40	40	40	40	459	161	230
Mode		35	20	20	30,00	18	1	0	20
Range		61	61	98	199,00	39	119	9	50
Minimum		15	1	2	1,00	7	1	0	0
Maximum		76	62	100	200,00	46	120	9	50

Figure 4.2 Statistical values

Table 4.1 Conversion of data

Attributes	Min	Max	Meaning of values
Age	15	76	0 = between 15 and 35 1 = between 36 and 55 2 = between 56 and 76
Age of smoking	7	46	0 = between 7 and 20 1 = between 21 and 34 2 = between 35 and 46
Smoking duration	1	62	0 = between 1 and 20 1 = between 21 and 41 2 = between 42 and 62
Year of cigarette consumption package	1	200	0 = between 1 and 66 1 = between 67 and 133 2 = between 134 and 200
How many months quit smoking	1	120	0 = between 1 and 39 1 = between 40 and 79 2 = between 80 and 120
Number of smokers at work	0	50	0 = between 1 and 10 1 = between 11 and 20 2 = between 21 and 50

Statistics

		FNBTtotalskor	hemogram	hematokrit	lokosit	trombosit	kolesterol	trigliserid	HDL	LDL	AST	ALT
N	Valid	630	532	532	532	532	516	509	513	510	526	535
	Missing	40	138	138	138	138	154	161	157	160	144	135
Mode		7	14,9	46,0	8000	240000	161[a]	79	46[a]	122	18	18
Range		10	10,2	29,4	15310	373000	313	745	80	252	161	116
Minimum		0	8,6	25,3	3290	90000	86	25	15	32	8	9
Maximum		10	18,8	54,7	18600	463000	399	770	95	284	169	125

Figure 4.3 Statistical values of health data

Table 4.2 Conversion of health data

Attributes	**Meaning of values**		**Missing value**
WBC	0→normal 1→abnormal	(between 4500-11000) (other)	0
Platelet	0→normal 1→abnormal	(between 150000-400000 (other)	0
FBNT	0→very little dependent 1→little dependent 2→moderately dependent 3→highly dependent 4→very high dependent	(between 0-2) (between 3-4) (value = 5) (between 6-7) (between 8-10)	---
Cholesterol	0→ normal 1→at the border 2→too high	(<200) (between 201-240) (>241)	0
Triglyceride	0→ normal 1→at the border 2→ high 3→ very high	(<150) (between 150-199) (between 200-499) (>500)	0
LDL	0→normal 1→high 2→very high	(<129) (between 130-159) (>160)	0
AST	0→normal 1→abnormal	(<40) (>41)	0
ALT	0→normal 1→abnormal	(<40) (>41)	0
	Woman	**Man**	
Hemogram	0→normal (between 12-16) 1→abnormal (other)	0→normal (between 13.5-17.3) 0→abnormal (other)	0
HTC	0→normal (between 38-46) 1→abnormal (other)	0→normal (between 42-54) 1→abnormal (other)	0
HDL	0→normal (between 50-60) 1→abnormal (other)	0→normal (between 40-60) 1→abnormal (other)	1

	s1yas	s2cinsiyet	s3medenidurum	s4egitimdurumu	s5sigaraicmesuresi	s6gunTuk	s7yılPakTuk	s8basYas	s9merak	s9ozenti	s9pskBaskı	s9Tepki	s9kendinispat
1	1	1	1	4	1	0	0	1	1	2	2	2	2
2	1	2	1	5	0	0	0	0	1	1	2	2	2
3	0	2	1	6	0	0	0	0	1	2	2	2	2
4	1	2	1	4	1	0	0	0	1	1	2	2	2
5	1	2	1	6	0	1	0	0	2	1	2	2	2
6	1	1	2	3	1	2	1	0	2	1	2	2	2
7	2	2	1	6	2	0	0	0	1	1	2	2	2
8	1	2	1	5	0	0	0	0	1	2	2	2	2
9	2	1	2	6	1	0	0	0	1	1	1	1	1
10	1	1	1	3	1	1	0	0	2	2	1	2	1
11	0	2	2	6	0	0	0	0	1	2	2	2	2
12	1	2	1	6	0	0	0	0	2	1	2	2	2
13	1	2	2	6	1	0	0	0	2	1	2	2	2
14	2	2	1	6	1	1	1	0	2	1	2	2	2
15	0	2	2	6	0	1	0	0	1	1	2	2	1
16	1	1	2	6	0	0	0	0	2	1	2	2	2
17	0	1	1	5	0	0	0	0	1	1	2	2	2
18	1	1	1	4	0	0	0	1	2	2	2	2	2
19	1	2	1	6	1	0	0	0	2	1	2	2	1
20	2	1	3	4	1	0	0	1	2	2	2	2	2

Figure 4.4 Data converted to nominal data type

The set values have been converted to .ARFF format that WEKA software can use. Figure 4.5 shows the .ARFF format file.

```
@relation saglık

@ATTRIBUTE s1yas {0,1,2}
@ATTRIBUTE s2cinsiyet {1,2}
@ATTRIBUTE s3medenidurum {1,2,3}
@ATTRIBUTE s4egitimdurumu {1,2,3,4,5,6}
@ATTRIBUTE s5sigaraicmesuresi {0,1,2}
@ATTRIBUTE s6gunTuk {0,1,2}
@ATTRIBUTE s7yılPakTuk {0,1,2}
@ATTRIBUTE s8basYas {0,1,2}
@ATTRIBUTE s9merak {1,2}
@ATTRIBUTE s9ozenti {1,2}
@ATTRIBUTE s9pskBaskı {1,2}
@ATTRIBUTE s9Tepki {1,2}
@ATTRIBUTE s9kendinispat {1,2}
@ATTRIBUTE s9uzuntu {1,2}
@ATTRIBUTE s9diger {1,2}
@ATTRIBUTE s10degisiklik {1,2,3,4}
@ATTRIBUTE s11hastalık {1,2}
@ATTRIBUTE s11topBaskısı {1,2}
@ATTRIBUTE s11hastKorkusu {1,2}
@ATTRIBUTE s12birakmayidusundunuzmu {1,2}
@ATTRIBUTE s13birakmayidenedinizmi {1,2}
@ATTRIBUTE s14kacAybıraktınız {0,1,2}
@ATTRIBUTE s15MaxBırakGunu {0,1,2,3,4}
@ATTRIBUTE s16Sinir {1,2}
@ATTRIBUTE s16kabızlık {1,2}
@ATTRIBUTE s16konsantrasyonbzk {1,2}
@ATTRIBUTE s16basAgrısı {1,2}
@ATTRIBUTE s16uyusma {1,2}
@ATTRIBUTE s16zorlukYok {1,2}
@ATTRIBUTE s16uykubzk {1,2}
@ATTRIBUTE s16istahArtısı {1,2}
@ATTRIBUTE s16dengesizlik {1,2}
@ATTRIBUTE s16agızyarası {1,2}
@ATTRIBUTE s16asırıistek {1,2}
@ATTRIBUTE s16diger {1,2}
@ATTRIBUTE s17profdestekaldinizmi {1,2}
@ATTRIBUTE s18yontemfarmakolojik {1,2}
@ATTRIBUTE s18yontempsikoterapi {1,2}
@ATTRIBUTE s18yontemdiger {1,2}
@ATTRIBUTE s19evdebaskasigaraicenvarmi {1,2}
@ATTRIBUTE s20isteicenvarmı {1,2}

@DATA

1,1,1,4,1,0,0,1,1,2,2,2,2,2,2,1,2,2,2,2,1,1,2,2,2,2,2,2,1,1,0,0,2,2,2,2,1,1,2,1,2,2,2,2,2,2,2,2,1,2,3,0,0,1,1,1,1,1,1,4,1,1,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2
,2,2,2,1,0,0,0,1,0,1,0,0,0,1,1,1,0,0,1,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1,1,0,1,0,0,0,0,0,1,1,1,0,0,1,0,1,0,0,0,1,1,2,0,2,2,1,1

1,2,1,5,0,0,0,0,1,1,2,2,2,2,2,1,2,2,1,1,1,1,1,2,2,1,2,2,1,1,0,0,1,1,1,2,2,2,2,1,2,2,1,2,1,2,2,2,2,2,1,0,0,2,1,1,1,1,2,1,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2
,2,2,2,0,0,0,1,1,0,0,0,0,1,0,1,0,1,1,1,0,1,1,0,1,1,0,1,0,0,0,0,0,1,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,2,0,0,0,0,1,1,2,0,2,2,2,1

0,2,1,6,0,0,0,0,1,2,2,2,2,2,2,1,2,2,1,2,2,1,2,2,2,2,2,2,1,1,0,0,1,2,2,2,2,2,1,1,2,2,1,2,2,1,2,2,2,1,2,0,0,1,1,1,2,1,2,1,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2
,2,2,2,0,0,0,0,0,1,0,1,0,0,0,1,0,1,0,1,0,1,1,0,1,1,0,0,0,0,0,1,0,0,0,0,1,0,1,0,0,0,0,0,0,1,0,0,0,1,0,0,1,0,0,0,1,1,2,0,2,1,2,1

1,2,1,4,1,0,0,0,1,1,2,2,2,2,2,3,2,2,2,2,2,2,2,2,2,2,2,1,1,1,0,0,1,2,2,2,2,2,1,1,2,2,1,2,2,2,2,2,1,1,0,1,0,1,1,1,1,1,2,4,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2
,2,2,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,1,0,1,0,1,1,2,0,2,2,2,1

1,2,1,6,0,1,0,0,2,1,2,2,2,2,2,3,2,2,2,2,2,1,2,2,2,2,2,1,1,1,0,0,1,2,1,1,1,2,2,2,2,2,1,2,2,2,2,2,2,1,3,2,1,1,2,1,2,1,2,4,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2
,2,2,2,1,0,0,0,0,0,0,0,0,0,0,1,0,1,0,0,0,1,0,0,0,1,0,0,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,1,0,0,1,0,1,0,1,1,1,2,0,2,2,2,1
.
.
.
```

Figure 4.5 Data in .ARFF format

The generated .ARFF file is an input to the J48 algorithm in the library of the WEKA software.

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.03 seconds

=== Summary ===

Correctly Classified Instances         616               97.7778 %
Kappa statistic                          0.9388
Mean absolute error                      0.0401
Root mean squared error                  0.1406
Relative absolute error                 10.8464 %
Root relative squared error             32.725  %
Total Number of Instances              630

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
                 0,929    0,006    0,979      0,929    0,953      0,939    0,988     0,966
                 0,994    0,071    0,977      0,994    0,985      0,939    0,988     0,994
Weighted Avg.    0,978    0,056    0,978      0,978    0,978      0,939    0,988     0,987

=== Confusion Matrix ===

   a   b   <-- classified as
 143  11 |   a = 1
   3 473 |   b = 2
```

Figure 4.6 J48 output

The J48 algorithm accurately predicted 616 of a total of 630 data with a 97.77 percent success rate.According to the confusion matrix, although 154 people stopped smoking, the algorithm incorrectly estimated 11 people as not to quit smoking. Although 481 people have quit smoking, the algorithm incorrectly estimates 3 people as quitting.In addition, statistical results and error values are shown in Figure 4.6.

As a result of the study, the decision tree in Figure 4.7 was obtained.

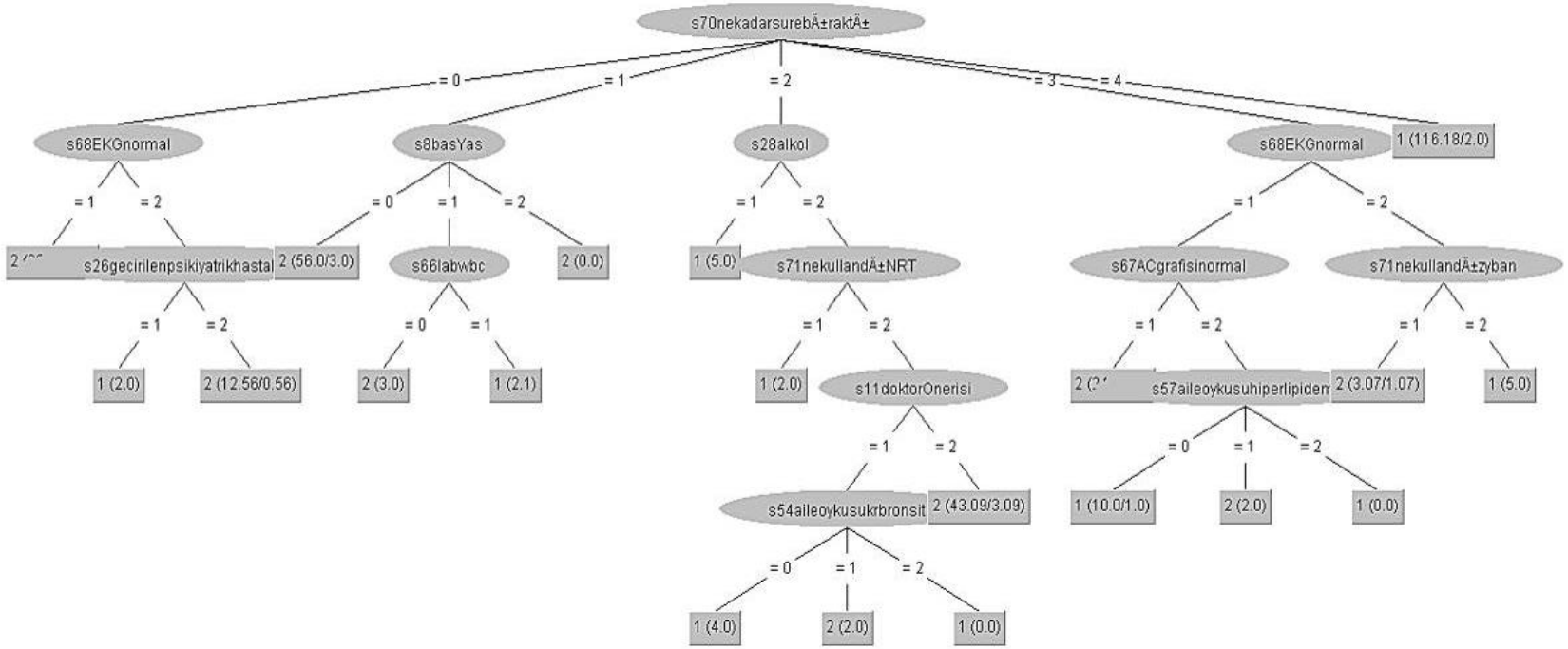

Figure 4.7 Decision tree derived from J48 algorithm

4.2 Comparison of Algorithms

In addition to the J48 algorithm, the data set for smoking cessation patients was also run in the classification algorithms Naive Bayes, Random Forest and K-Neasrest Neighbors (KNN) algorithms. The Naive Bayes algorithm estimated 95.7143 percent. In total, it correctly estimated 603 of 630 data. In total, it correctly estimated 603 of 630 data. This algorithm, incorrect estimated 15 of the remaining 27 data as he did not quit smoking and 12 as quit smoking. Random Forest and KNN algorithms correctly predicted all of 630 patients with ratio 100 percent. Some statistical results of these algorithms can be seen in Table 4.3. When these three algorithms are compared, it is seen that Random Forest and KNN algorithms makes more accurate estimates than the others.

Table 4.3 Comparison of algorithms

	J48	**Naive Bayes**	**Random Forest**	**KNN**
Correctly classified instances	97.7778 %	95.7143 %	100 %	100 %
Kappa statistic	0.9388	0.8832	1	1
Mean absolute error	0.0401	0.0658	0.091	0.0016
Avg. TP Rate	0.978	0.957	1	1
Avg. FP Rate	0.56	0.080	1	1
Avg. Precision	0.978	0.957	1	1
Avg. Recall	0.978	0.957	1	1
Avg. F measure	0.978	0.957	1	1
Avg. MCC	0.939	0.883	1	1
Avg. ROC Area	0.988	0.986	1	1
Avg. PRC Area	0.987	0.988	1	1
Confusion matrix	a b 143 11 | a=1 3 473 | b=2	a b 139 15 | a=1 12 464 | a=2	a b 154 0 | a=1 0 476 | b=2	a b 154 0 | a=1 0 476 | b=2

CHAPTER FIVE
DEVELOPING THE DECISION SUPPORT SYSTEM SOFTWARE WITH VISUAL

This section describes decision support software and its functionality. In the study, WEKA library was added to visual studio where data mining algorithms can be used with the help of this library J48 classification data mining algorithm was used. The same decision tree obtained in WEKA software was obtained.

Using the obtained decision tree model, the factors affecting smoking cessation are determined and these factors are listed below.

- How long did the patient stop smoking,
- ECG status,
- Age of starting smoking,
- Alcohol use,
- Psychological illness,
- Laboratory WBC value,
- NRT, which drug used,
- ZYBAN, which drug used,
- Status of the AC radiograph,
- Doctor's recommendation,
- Family history of bronchitis,
- Family history Peptikalcus.

Of the 150 characteristics of each patient collected in the hospital for smoking cessation, 12 important features were distinguished that would help predict this issue. Thus, it is not necessary to record 150 information to calculate the probability of smoking cessation of each new patient. For this reason, only 12 features were analyzed.

Accurate interpretation of the characteristics obtained in the decision tree obtained from the analyzes is important. For the correct interpretation of this tree, it is

important to know what the values of the properties mean. Descriptions of the characteristics obtained from the decision tree are given below. Thus, the interpretation of the decision tree creates a true result.

BırakmaSüresi (how long did he stop smoking): 0→<30 day, 1 → 31-90 day, 2 → 91-180 day, 3→ 181-360 day, 4 → >361 day

EKGNormal (EKG status): 1 → normal, 2 → abnormal

BasYası (age of onset of smoking): 0 → young, 1 → middle aged, 2 → old

WBC (blood value): 0 → normal, 1 → abnormal

Alcohol (alcohol use): 1 → yes, 2 → no

DrOnerisi (doctor recommendation): 1 → yes, 2 → no

PskHastalık (psychiatric illness): 1 → yes, 2 → no

NRT (NRT drug use): 1 → yes, 2 → no

AileOykusuBronşit (bronchitis in family history): 0 → no, 1 → 1st degree, 2nd degree

ACGrafisi (lung X-ray): 1 → normal, 2 → abnormal

ZYBAN (ZYBAN drug use): 1 → yes, 2 → no

AileOykusuHiperlipidemi (hiperlipidemi in family history): 0 → no, 1 → 1st degree, 2 →2nd degree

5.1 Storing Data in SQL Server

For the permanent storage of patient records, Microsoft SQL Server - SQL Server Manage Studio was used. Family history, medication used, the reason for quitting smoking and patients were formed in Figure 5.1 different tables. The figure is shown.

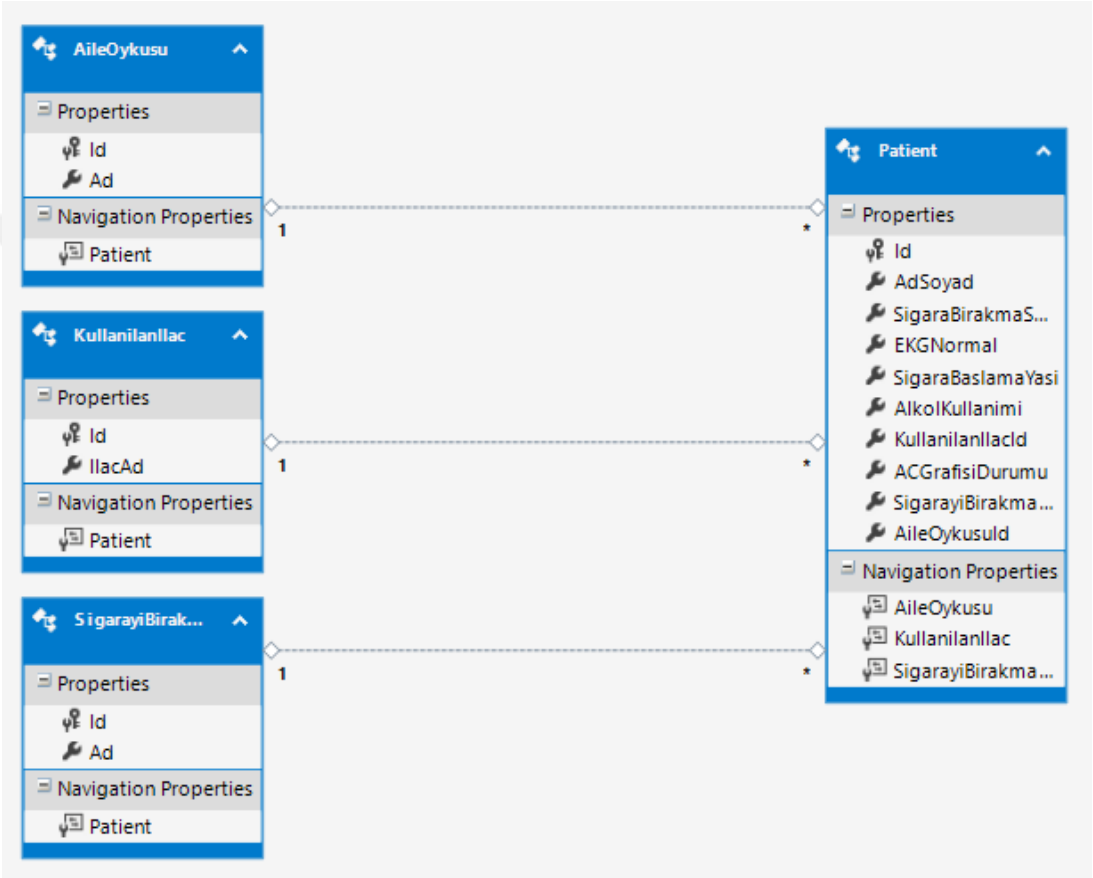

Figure 5.1 SQL tables

5.2 Web Application

In order to make patient records in the web application, first of all, the family history, medications and reasons for smoking cessation should be completed. Figure 5.2 shows the buttons for the data to be completed. Thus, patient data will be created.

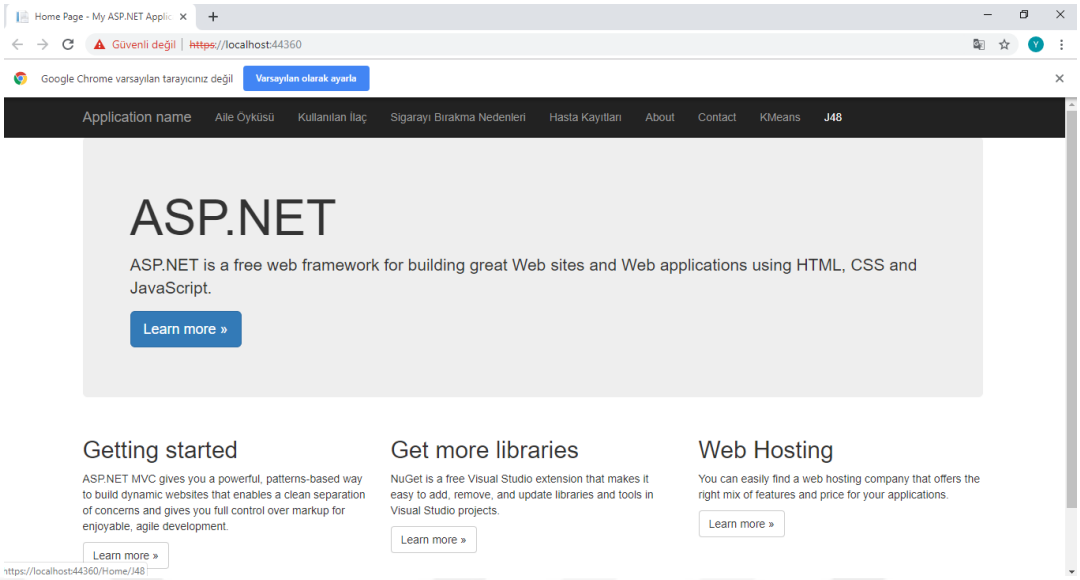

Figure 5.2 Home page

5.2.1 Family Story

A database is created with the diseases that occurred in the family history of the patients. Adding and deleting new diseases for family history can also be done. Family historysituations that may be in the way are easily recorded in the database with the application. Figure 5.3 shows the page that will be opened when the family history button is clicked.

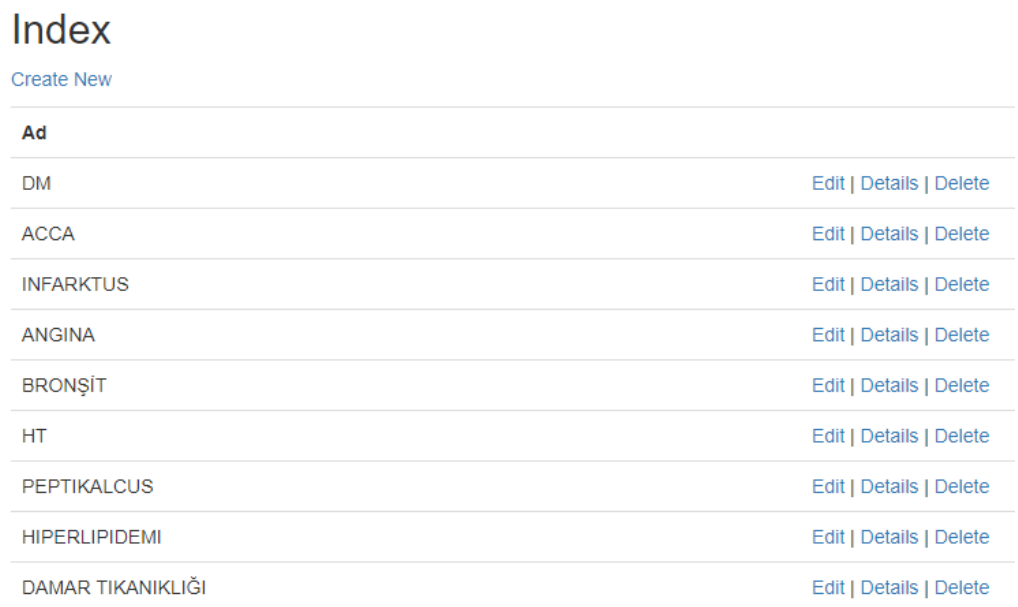

Figure 5.3 Family story

5.2.2 Reasons to Stop Smoking

The reasons for quitting smoking to assist in patient records are included in this section. These reasons were recorded in the database to be edited when necessary. Figure 5.4 shows possible reasons for quitting.

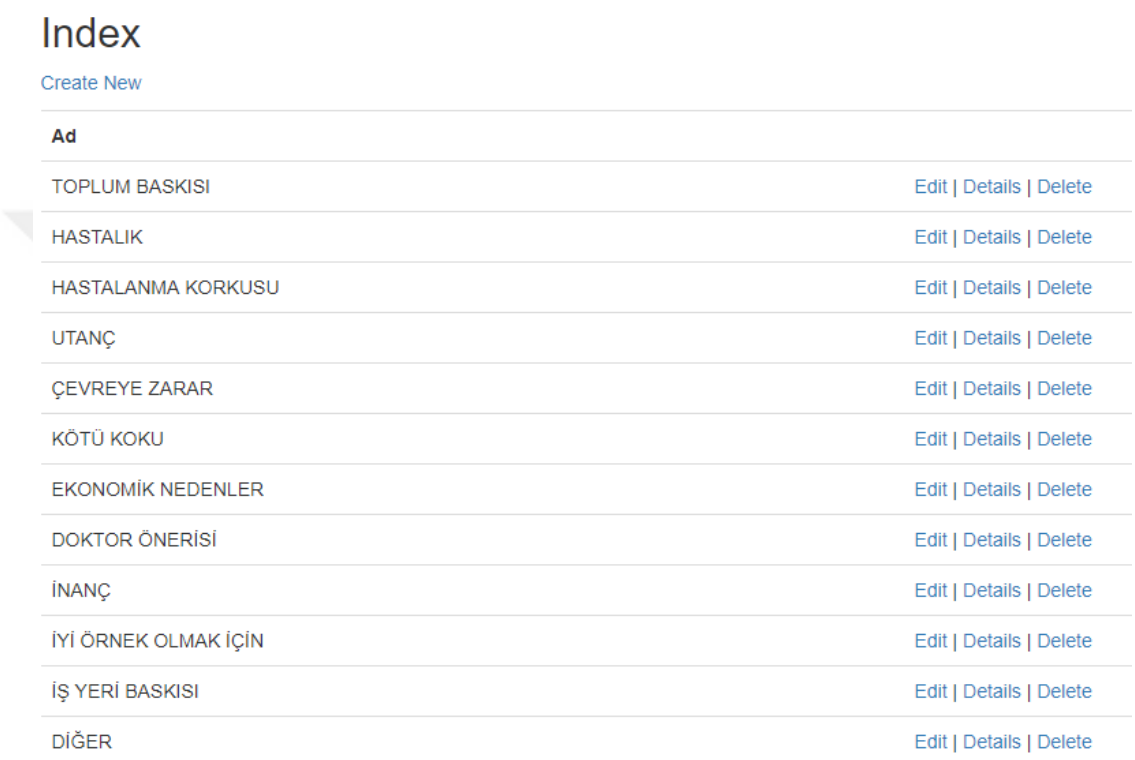

Figure 5.4 Reasons to stop smoking

5.2.3 Drug Used

The names of drugs that doctors can recommend to quit smoking have been added to the database. Records are shown in Figure 5.5.

Figure 5.5 Drug used

5.2.4 Patient Records

This information should be filled in to estimate the probability of smoking cessation of each new patient coming to the doctor. Necessary information are showed in Figure 5.6. Thus, the database of patients is created. As shown in Figure 5.7, each new patient record is made here.

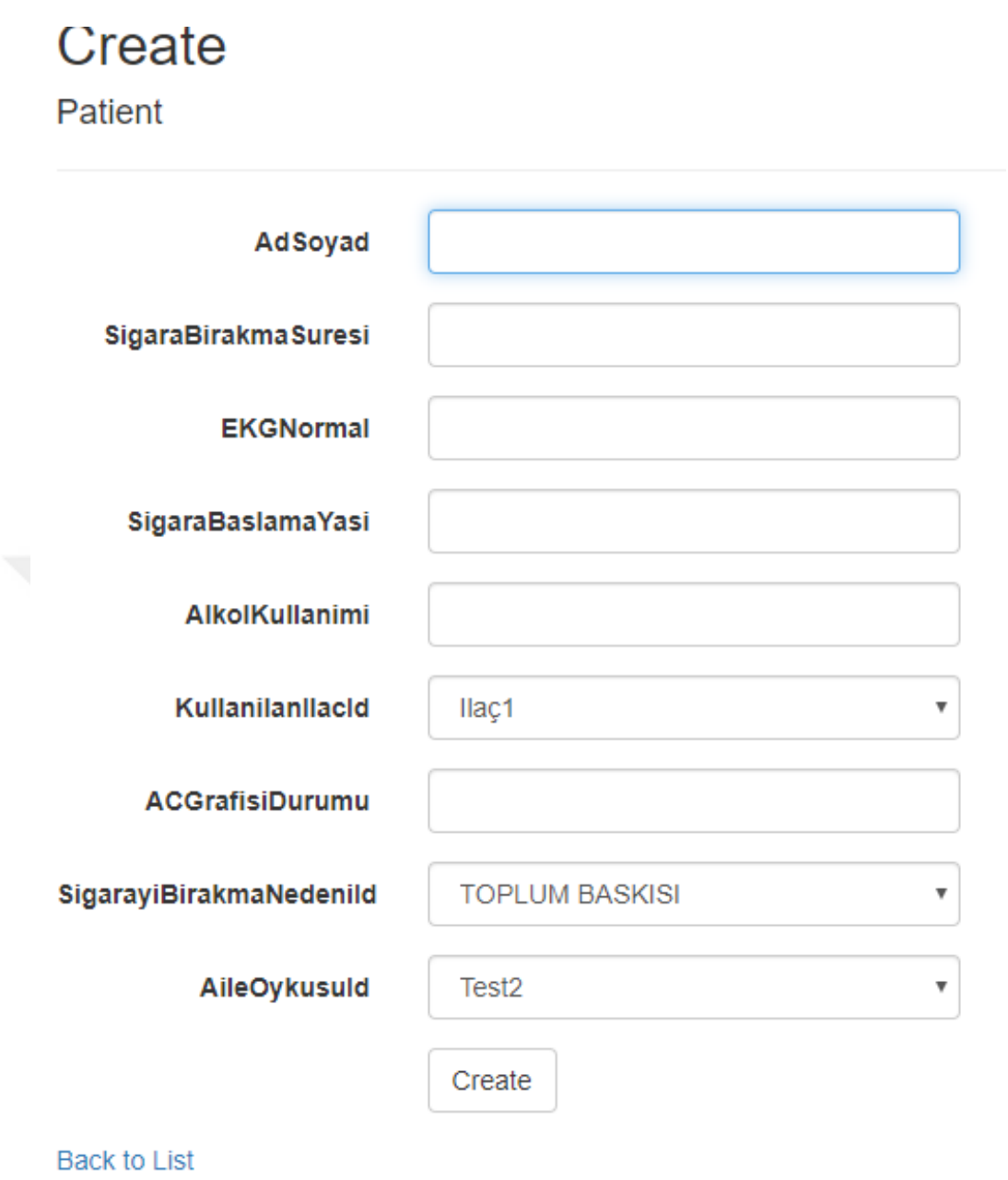

Figure 5.6 Patient record

Figure 5.7 List of registered patients

5.2.5 Estimation of the Possibility of Smoking Cessation

The results of the analysis of 630 patients who came to the hospital for smoking cessation are shown in Figure 5.8. The probability of quitting smoking is associated with the duration of smoking cessationIn graph; blue circles are indicated that patients have stopped smoking, orange circles are indicated that patients do not stop smoking.

It is possible to estimate the probability of smoking cessation only with the information in the chart. If the characteristics of the patient follow the blue-striped states, it may be concluded that smoking is more likely to stop smoking. In other cases, we can comment that the probability of quitting smoking is low.

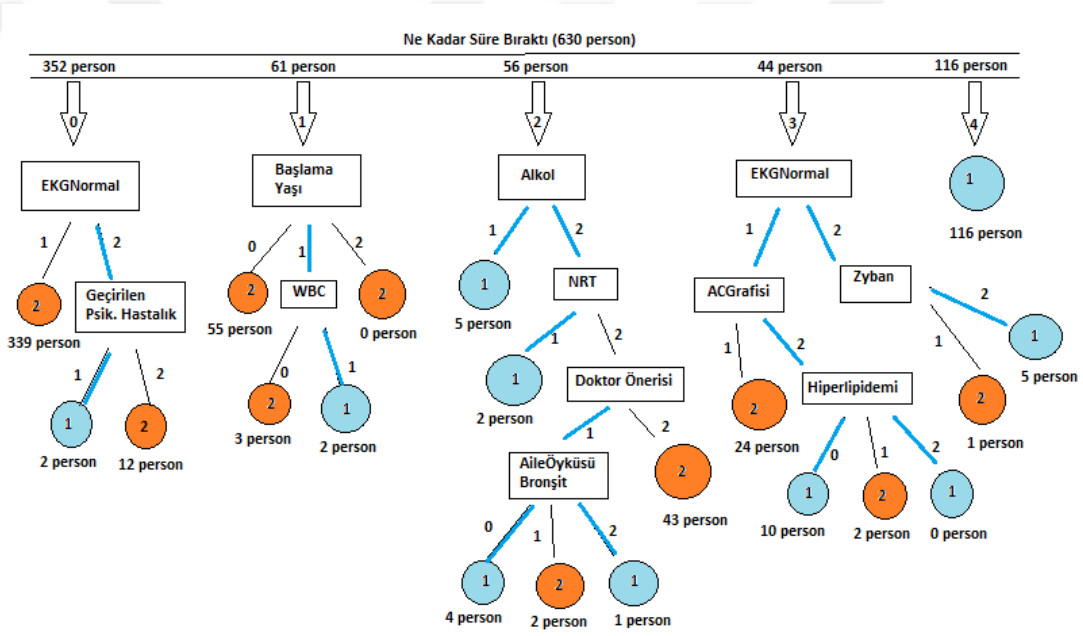

Figure 5.8 Probability prediction graph

CHAPTER SIX
CONCLUSION AND FUTURE WORK

6.1 Conclusion

Analysis is performed on 150 attributes of 630 patients stored in the Smoking Cessation Policlinic of Dokuz Eylül Hospital. These data are analyzed using J48 algorithm which is one of the data mining algorithms. As a result of this analysis, a decision tree was formed. Finally, 12 data features were identified in this decision tree to help predict smoking cessation. In other words, ease of operation is provided by accessing the valuable data among too many data.

The following results were reached in the decision tree obtained.

- The smoking cessation status have been associated with the cessation time.
- If the ECG results were abnormal and the patient had a psychiatric disorder, the likelihood of smoking cessation has been seen increased.
- It has been observed that the probability of quitting smoking is abnormal if the WBC outcome is abnormal.
- It was observed that the drug NRT was effective in quitting smoking, and those who did not use this medication had a family history of bronchitis.
- The effect of Zyban drug on smoking cessation have been seen lower than NRT.
- It has been seen that the higher the smoking cessation time, the higher the probability of smoking cessation.
- In the estimations made as a result of data mining, 616 out of 630 patients with the mean absolute error rate of 0.04 were correctly predicted to quit smoking.

The analyzes were evaluated using the decision support system. Estimates were made about smoking cessation using only 12 features of the patients. Thus, a decision support system was developed to assist health workers in predicting patients smoking cessation.

6.2 Future Work

Within the scope of the study, analyzes were made on the data obtained from the Smoking Cessation Policlinic of İzmir Dokuz Eylül University. As a result, a decision tree was obtained that we can comment on the situations in which patients quit smoking. A small-scale web application has been developed in which the data obtained will be stored and used when necessary.

We are plan to make our web application more useful by developing our web applications in the future. In addition, it is aimed to look at and interpret the subject from a different perspective by conducting more detailed analyses with some features of the patients. In addition, it is aimed to look at and interpret the subject from a different perspective by conducting more detailed analyzes with some features of the patients.

REFERENCES

ASP.NET MVC Pattern (n.d.). Retrieved April 22, 2019, from https://dotnet.microsoft.com/apps/aspnet/mvc

Bharati M., & Ramageri M. (2010). Data mining techniques and applications. *Indian Journal of Computer Science and Engineering*, *1*(4), 301-305.

Bilgen S. (1998). *Tuena Sağlık Bilgi Sistemleri çalışma belgesi*. TUBİTAK Retrieved May 15, 2002, from www.tuena.tubitak.gov.tr

Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H., & Trigg, L. (2010). Weka-a machine learning workbench for data mining. Maimon O, Rokach L, (Ed.). *Data Mining and Knowledge Discovery Handbook*(2st ed.). Berlin: Springer.

Haug, P. J., Rocha, B. H., & Scott, R. (2003). Decision support in medicine: Lessons from the help system. *International Journal of Medical Informatic*, *69*, 273-284.

Koyuncugil, A. S., & Özgülbaş, N. (2003). Veri madenciliği: Tıp ve sağlık hizmetlerinde kullanımı ve uygulamaları. *Bilişim Teknolojileri Dergisi*, *2*(2), 21-32.

Landau, S., & Everit, B. (2014*). A Handbook of Statistical Analyses using SPSS.* United Kingdom: Chapman & Hall/CRC Press LLC.

Musen, M. A., Yuval, S., & Shortliffe, E. H. (2003). *Clinical decision-support systems.* Retrieved August 8, 2003, from https://www.ie.bgu.ac.il/mdss/ch16.final.pdf

Özata, M., & Aslan, Ş. (2004). Klinik karar destek sistemleri ve örnek uygulamalar. *Kocatepe Tıp Dergisi*, *5*, 11–17.

Rajkumar, A,. & Reena, G. S. (2003). Diagnosis of heart disease using data mining algorithm. *Global Journal of Computer Science and Technology*,*10*(1), 38.

Sawar, M. J., Brennan, T. G., & Cole, A. J. (1991). Representing knowledge in medical decision support systems, Poems. *Proceedings of IJCA191 One Day Workshop*. 745-749.

Tang, P. C., & Mcdonald, C. J. (2001). Computer-based patientrecord systems. Shortliffe**.**, Edward, H., Cimino, James, J., (Ed). *Medical Informatics: Computer Applications in Health Care and Biomedicine.* (1th ed.). London: Springer.

Tomar, D., & Agarwal, S. (2013). A survey on data mining approaches for healthcare, *International Journal of Bio-Science and Bio-Technology*, *5*(5), 241-266.

Weka 3: Machine learning software in java (n.d.). Retrieved April 10, 2019, from http://www.cs.waikato.ac.nz/ml/weka

WHO Report on the Global Tobacco Epidemic (n.d.). Retrived 2009, from https://www.who.int/tobacco/mpower/2009/GTCR_2009-web.pdf

Yalçın, Ö. (2013*). Veri madenciliği yöntemleri*. Çölkesen, R. (Ed.). (2th ed.) İstanbul: Papatya Yayınevi

Yıldırım, Ö. (2000*). Kalp hastalıklarının teşhisinde kullanılan bir uzman sistem uygulaması.* MSc Thesis, Ege University, İzmir.

Yıldırım, P., Uludağ, M., Görür, A. (2008). *Hastane bilgi sistemlerinde veri madenciliği*. Akademik Bilişim, Çanakkale Onsekiz Mart University, Çanakkale.