

DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

**DATA MINING BASED DECISION SUPPORT
SYSTEM USING DATA WAREHOUSE**

by
Ezgi DEMİR

July, 2017
İZMİR

DATA MINING BASED DECISION SUPPORT SYSTEM USING DATA WAREHOUSE

**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

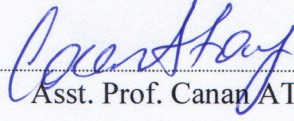
Ezgi DEMİR

July, 2017

İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled “**DATA MINING BASED DECISION SUPPORT SYSTEM USING DATA WAREHOUSE**” completed by **EZGİ DEMİR** under supervision of **ASST. PROF. CANAN ATAY** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.



Asst. Prof. Canan ATAY

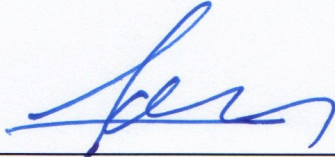
Supervisor

Yard. Doç. Dr. Semih VTKU


(Jury Member)

Doç. Dr. Mehmet ÜNLÜTÜRK


(Jury Member)


Prof. Dr. Emine İlknur CÖCEN

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere appreciation to my advisor Asst. Prof. Canan ATAY for her continuous support, guidance, patience and immense knowledge.

I thank to the American National Cancer Institute (NCI) for sharing this lung cancer dataset with us.

I thank to Prof. Dr. Vildan MEVSİM, from Family Medicine Department of Dokuz Eylul University Hospital, for her tolerance, understanding and contribution in interpreting this dataset.

I would like to express my profound gratitude to my parents and my brother for their endless support and patience.

I also thank to my friend, Gamze ÖZÇELİK who shared her knowledge and motivated me throughout the study.

I would like to offer my special thanks to Ulaş Devrim KARAMAN for his continuous support and encouragement.

Ezgi DEMİR

DATA MINING BASED DECISION SUPPORT SYSTEM USING DATA WAREHOUSE

ABSTRACT

Data warehousing and data mining have begun to be used in every aspect of life because of the increasing amount of data. Especially the health sector has a lot of data that needs to be converted into knowledge and to be evaluated. With the integration of data warehouse and data mining techniques, it is possible to expose information that supports the decision-making process of doctors and administrators.

In this study, we have created a data warehouse for lung cancer data that received from American National Cancer Institute (NCI). The lung cancer data have brought into certain formats, cleaned from errors and repetitions. The star schema model was used; hence the data warehouse is designed to respond to even the most complex queries. Then Naive Bayes, K-Nearest Neighbors, C4.5 classification algorithms were applied for the data in the data warehouse. By comparing the results of the applied algorithms, it is seen that the C4.5 algorithm is more successful than other algorithms. The classification process is based upon the survival of the lung cancer patient. The decision tree produced by C4.5 has provided various rules regarding the likelihood of survival of lung cancer patients. At the next stage to the study, a web-based prototype system was developed to provide support to physicians in the decision-making process using the decision tree model obtained.

When considering the size and complexity of cancer data, a decision-support system built with both data warehouse and data mining can shed light on the treatment process for doctors.

Keywords: Data warehouse, data mining, decision support system, lung cancer data

VERİ AMBARI KULLANARAK VERİ MADENCİLİĞİ TABANLI KARAR DESTEK SİSTEMİ

ÖZ

Günden güne artan veri miktarı nedeniyle veri ambarı ve veri madenciliği, yaşamın her alanında kullanılmaya başlanmıştır. Özellikle sağlık sektörü bilgiye dönüştürülmeyi ve değerlendirilmeyi bekleyen çok fazla miktarda veri barındırmaktadır. Veri ambarı ve veri madenciliği tekniklerinin entegrasyonu ile bu verilerden doktorların ve yöneticilerin karar verme sürecini destekleyen bilgilerin açığa çıkarılması mümkün olabilmektedir.

Bu çalışmada Amerikan Ulusal Kanser Enstitüsü'nden (NCI) alınan akciğer kanseri verileri hatalardan ve tekrarlardan arındırılarak veri ambarı modelinde tutulmuştur. Veri ambarı tasarlanırken çok boyutlu veri modelleme yaklaşımlarından yıldız şema modeli kullanılmıştır. Oluşturulan modelin yazılan karmaşık sorgulara yanıt verebildiği görülmüş ve sonuçları sunulmuştur. Veri ambarındaki veriler üzerinde Naive Bayes, K-Nearest Neighbors, C4.5 sınıflandırma algoritmaları uygulanmıştır. Uygulanan algoritmaların sonuçları karşılaştırılarak C4.5 algoritmasının daha başarılı sonuç verdiği görülmüştür. Sınıflandırma işlemi akciğer kanseri hastasının hayatta kalma durumuna göre yapılmıştır. C4.5 ile üretilen karar ağacından akciğer kanseri hastalarının hayatta kalma olasılığı ile ilgili çeşitli kurallar elde edilmiştir. Çalışmanın sonraki aşamasında, elde edilen karar ağacı modeli kullanılarak doktorlara karar verme aşamalarında destek sağlayacak web tabanlı örnek bir sistem geliştirilmiştir.

Kanser verilerinin büyüklüğü ve karmaşıklığı düşünüldüğünde veri ambarı ve veri madenciliğinin birlikte kullanılmasıyla kurulan bu karar destek sistemi, doktorların tedavi süreçlerini şekillendirmesinde onlara ışık tutabilecektir.

Anahtar kelimeler: Veri ambarı, veri madenciliği, karar destek sistemi, akciğer kanseri verileri

CONTENTS

	Page
M.Sc. THESIS EXAMINATION RESULT FORM.....	ii
ACKNOWLEDGEMENTS.....	iii
ABSTRACT.....	iv
ÖZ.....	v
LIST OF FIGURES.....	ix
LIST OF TABLES.....	xi
CHAPTER ONE - INTRODUCTION.....	1
1.1 General	1
1.2 Purpose	2
1.3 Organization of the Thesis	3
CHAPTER TWO - LITERATURE REVIEW	4
2.1 Data Mining Studies in the Field of Cancer	4
2.2 Data Warehouse Studies in the Field of Cancer	5
2.3 Decision Support System Studies in the Field of Cancer	6
CHAPTER THREE - DATA WAREHOUSING, DATA MINING AND DECISION SUPPORT SYSTEMS.....	8
3.1 Data Warehousing	8
3.1.1 Brief Characterization of Data Warehousing	8
3.1.2 Data Warehouse Architecture	9
3.1.3 Data Warehouse Modeling	11
3.2 Data Mining	13
3.2.1 Data Mining Process	14
3.2.2 Classification Algorithms	15

3.2.2.1 Naive Bayes.....	15
3.2.2.2 K-Nearest Neighbors	16
3.2.2.3 C4.5 Decision Tree Algorithm	17
3.3 Challenges of Data Warehousing and Data Mining in the Healthcare	19
3.4 Decision Support Systems	20
CHAPTER FOUR - USED TECHNOLOGIES.....	22
4.1 Visual Studio 2012.....	22
4.2 ASP.NET MVC 4	23
4.3 MS Sql Server Management Studio 2014.....	24
4.4 SAS University Edition	24
4.5 Weka	25
4.6 Used Programming Languages	25
CHAPTER FIVE - PROPOSED PROJECT.....	26
5.1 Dataset Description	27
5.2 Data Warehousing Stage	29
5.2.1 Data Processing	30
5.2.2 Data Storage..	32
5.2.3 Example of Querying in Data Warehouse.....	33
5.3 Data Mining Stage	37
5.3.1 Preparation of Arff File	37
5.3.2 Implementantion of Classification Algorithms..	39
5.3.3 Cross Validation.....	41
5.3.4 Implementantion of Naive Bayes Algorithm	41
5.3.5 Implementantion of KNN Algorithm.....	42
5.3.6 Implementantion of C4.5 (J48) Algorithm.....	43
5.3.7 Implementantion of C4.5 (J48) Algorithm.....	48
5.4 Development of Prototype Decision Support System	48

CHAPTER SIX - CONCLUSION AND FUTURE WORK.....	57
6.1 Conclusion	57
6.2 Future Work	58
REFERENCES.....	59
APPENDICES.....	65
APPENDIX 1 Dataset Attribute Explanations	65



LIST OF FIGURES

	Page
Figure 3.1 Data warehousing layers	10
Figure 3.2 Star schema about medical records	12
Figure 3.3 Snowflake schema about medical records	13
Figure 3.4 Overview of the steps constituting the data mining process	14
Figure 3.5 KNN algorithm	15
Figure 3.6 Bayes theorem	16
Figure 3.7 Learning phase of the decision tree algorithm	17
Figure 3.8 Classification phase of the decision tree algorithm	18
Figure 3.9 Basic algorithm for decision tree induction	18
Figure 5.1 Proposed project architecture	26
Figure 5.2 Data transfer from SAS data table to Excel spreadsheets	30
Figure 5.3 Temporary lung cancer database	31
Figure 5.4 Lung cancer data warehouse	31
Figure 5.5 Example for conversion operations	32
Figure 5.6 Lung cancer data warehouse star schema model	33
Figure 5.7 Arff file definition part	38
Figure 5.8 Arff file data part	39
Figure 5.9 Imported dataset	39
Figure 5.10 Distribution of class according to attributes	40
Figure 5.11 10-fold cross validation	41
Figure 5.12 Naive Bayes classification result	42
Figure 5.13 KNN classification result	43
Figure 5.14 C4.5 classification result	44
Figure 5.15 Decision tree created by the C4.5 algorithm	46
Figure 5.16 Sample section from decision tree-1.....	47
Figure 5.17 Sample section from decision tree-2.....	47
Figure 5.18 Comparison of classifiers	48
Figure 5.19 MVC architecture	49
Figure 5.20 Lung cancer decision support system MVC architecture	49
Figure 5.21 Lung cancer decision support system use case diagram	50

Figure 5.22 Lung cancer decision support system home page	50
Figure 5.23 Lung cancer decision support system our service page	51
Figure 5.24 Lung cancer decision support system make contact page	51
Figure 5.25 Patient information page	52
Figure 5.26 Medical history page	52
Figure 5.27 Cigarette history page	53
Figure 5.28 Cancer characteristics page	53
Figure 5.29 Treatment page	54
Figure 5.30 Risk of death is low	55
Figure 5.31 Risk of death is high	55
Figure 5.32 Information guide page	56



LIST OF TABLES

	Page
Table 3.1 Characteristics of conventional vs clinical data warehouses	19
Table 5.1 Some attributes examples from the dataset	29
Table 5.2 Result of Query 1	34
Table 5.3 Result of Query 2	35
Table 5.4 Result of Query 3	36



CHAPTER ONE

INTRODUCTION

1.1 General

Healthcare organizations need to use their data more consciously in the decision-making process, reduce costs by increasing productivity, and better manage resources. The health sector has a lot of data that needs to be turned into knowledge to accomplish all these goals. By taking advantage of information technology, it is necessary to diagnose illnesses in a short time, to manage existing data more easily and to select and apply the treatment method quickly.

Since healthcare data are extremely complex, heterogeneous, and separate from each other, it is difficult to integrate reliably. It is very time consuming and laborious to get consistent results because unintegrated data prevents doctors from being aware of each other's decisions. For the solution of this problem, it is necessary to keep the data in a certain format, which does not contain any mistakes and uncertainties, under a single roof called the data warehouse. The data warehouse provides a powerful solution for data integration, allowing for various analyzes and queries on the data.

Retrieving information from data warehouses through traditional query methods does not allow the hidden and important rules on the information to be revealed. Therefore, it is inevitable to apply data mining techniques that are used in this area for knowledge discovery from data warehouses. This discovery of knowledge through data mining helps healthcare organizations make strategic decisions.

When doctors' decision-making process is examined, past knowledge and experiences seem to be influential. Therefore, decisions may not be made where necessary due to inexperience, humanitarian situations and similar instant or permanent problems. As a result, undesirable results can be encountered in the

medical field where the error tolerance is very low (Kallmeyer & Venkat, 2002). The use of data warehousing and data mining techniques for decision support has emerged as a new direction in the healthcare field. Decision support systems built using data warehousing and data mining techniques are information system applications that help make the best decisions about patients by providing the most up-to-date information for doctors.

According to Globocan 2012 (Globocan, 2012) data, a total of 14.1 million new cases of cancer have occurred in the world during 2012 and there are 8.2 million cancer-related deaths. The most diagnosed cancers were lung 13%, breast 11.9% and colon 9.7% while cancer deaths were mostly lung 19.4%, liver 9.1%, stomach 8.8%. If the rate of cancer growth continues in this way, there will be a total of 19.3 million new cases of cancer in 2025 due to the increase in the world population and the aging of the population (Kanser, n.d.).

Cancer is a complicated disease which has many different types of variables and fluctuates rapidly depending on time, and there are many factors that should not be overlooked during diagnosis, follow-up and treatment. Through the use of the decision support system established with data warehouse and data mining in cancer disease, it will be possible to shorten the time that is valuable for disease processes, to provide better quality services and to obtain positive results.

1.2 Purpose

This thesis focuses on lung cancer, one of the most serious diseases of our age. This cancer type was chosen in this study since lung cancer is the first in terms of incidence and mortality rate in the world.

The data required in this study were obtained from the American National Cancer Institute (NCI, 2017). These data included demographic information of the patient, past medical history, general health status, smoking history, treatment modalities,

and cancer characteristics. Lung cancer data was kept into the data warehouse model by bringing into certain formats and purifying from mistakes and repetitions. The star schema modeling approach was used; hence the data warehouse is designed to support the decision-making process by meeting even the most complex queries.

The data stored in the data warehouse was converted to arff format and made ready for Weka. Naive Bayes, K-nearest neighbors and C4.5 data mining algorithms have been applied for classification using Weka program. As the performance measure criteria, the classification accuracy has been used to evaluate the classifier algorithms and then to select the best method. 10-fold cross validation was used in the test and the most accurate results were obtained from the C4.5 decision tree algorithm by 78.84%. The classification operation is based on whether the patient's cause of death is lung cancer. The algorithm builds a decision tree from the data of patients who died or survived due to lung cancer.

Using this decision tree, a prototype decision support system was developed to provide physicians an idea of the survival probabilities of future lung cancer patients. The aim of this study is to show that it may be possible to provide a better decision support environment for cancer disease by using the benefits of data warehousing and data mining methods.

1.3 Organization of the Thesis

This thesis includes six chapters and the remaining of this thesis is organized as follows. Chapter 2 summarizes the related literature and previous studies. Chapter 3 provides information on data warehousing, data mining and decision support systems. Chapter 4 explains used technologies during the study. Chapter 5 describes the data set used, the stages of data warehouse preparation, the implementation of data mining and the development of a prototype decision support system. Chapter 6 contains some concluding remarks and future studies.

CHAPTER TWO

LITERATURE REVIEW

There are many successful data mining, data warehousing and decision support system work examples in the field of cancer worldwide. Below, some application summaries of related works are given.

2.1 Data Mining Studies in the Field of Cancer

Danacı, Çelik & Akkaya (2010) give brief information about the most common breast cancer among women. Then, with the help of the Xcyt pattern recognition program, general data about the tissue were obtained and the breast cancer cells were estimated and diagnosed using the Weka program.

Bellaachia & Güven (2006) present an analysis of the prediction of survivability rate of breast cancer patients using data mining techniques. They investigated three data mining techniques: the Naive Bayes, the back-propagated neural network, and the C4.5 decision tree algorithms and they found out that C4.5 algorithm has a much better performance than the other two techniques.

Krishnaiah, Narsimha & Chandra (2013) developed a prototype lung cancer disease prediction system using data mining classification techniques. The most effective model to predict patients with Lung cancer disease appears to be Naive Bayes followed by if-then rule, decision trees and neural network. For diagnosis of lung cancer disease Naive Bayes gave better results than decision trees.

Zubi & Saad (2013) used some data mining techniques such as neural networks for detection and classification of lung cancers in X-ray chest films to classify problems aiming at identifying the characteristics that indicate the group to which each case belongs.

2.2 Data Warehousing Studies in the Field of Cancer

With their work, researchers presented the evaluation of the architecture of healthcare data warehouse specific to cancer diseases. They build the cancer data warehouse to integrate between the operational database and medical files and therefore the analysis on data makes easy by using OLAP cubes and viewing multilevel of details from the data (Sheta & Eldeen, 2013).

Pedersen & Jensen (1998) identified that clinical data warehouse needs to support for complex-data modeling features, advanced temporal support, advanced classification structures, continuously valued data, dimensionally reduced data and the integration of very complex data. Hence, clinical data warehouse requires advanced data modeling than conventional multidimensional data warehousing approaches.

Arous et al. (2013) have integrated two different operational health systems that include pancreatic cancer data. They mentioned the challenges they faced as the data types in the databases differed.

In their study, Hu et al. (2004) took a data warehouse approach because databases designed for online transaction processing are not amenable for use in scanning large volumes of data to find answers to complex queries. In developing comprehensive breast cancer data warehouse, they developed a hybrid system (integrated and federated) in order to optimize the advantages and minimize the disadvantages.

Wah & Sim (2009) reviews the development and use of a clinical data warehouse specific to the Lymphoma or Lymph Node cancer, which could be used by doctors, physicians and other health professionals. They proposed a 5-stage sequential methodology for the clinical data warehouse development.

Choi et al. (2015) developed prostate cancer research database system which incorporates information about a prostate cancer research including demographics, medical history, operation information, laboratory, and quality of life surveys. Their system includes three different ways of clinical data collection to produce a comprehensive data base; direct data extraction from electronic medical record (EMR) system, manual data entry after linking EMR documents like magnetic resonance imaging findings and paper-based data collection for survey from patients.

2.3 Decision Support System Studies in the Field of Cancer

The vast majority of decision support systems for cancer are still at the research level and only a few are used in clinical practice. PAPNET is a decision support system clinically used in cervical cancer. PAPNET uses artificial neural networks to identify abnormal cell appearances on vaginal smear slides (Boon & Kok, 2001).

Forgionne, Gangopadhyay & Adya (2000) discussed how data warehousing, data mining, and decision support systems can reduce the national cancer burden or the oral complications of cancer therapies, especially as related to oral and pharyngeal cancers. An information system is presented. The system organized relevant claims data, detected cancer patterns in general and special populations, formulated models that explain the patterns and evaluated the efficacy of specified treatments and interventions with the formulations.

Ramachandran, Girija & Bhuvaneshwari (2014) developed clinical data warehouse which integrates data by automatically performing the ETL procedure i.e. extracting the data from different sources, transforming and gets itself loaded to supports the data mining system which could be used by doctors and medical analysts as a decision support system, to predict cancer in its earlier stages and provide the needed treatment.

Goletsis et al. (2011) performed data integration between medicine and molecular biology, by developing a framework where, clinical and genomic features are appropriately combined in order to handle cancer diseases. Through this integration, real time conclusions can be drawn for early diagnosis, staging and more effective cancer treatment. The integrated data (clinical and genomic) are analysed in order to discover valuable knowledge that can be used for decision support purposes.

Fernandes et al. (2010) proposed a breast cancer decision support system which comprises three different prognostic modelling methodologies: the clinically widely used Nottingham prognostic index, the Cox regression modelling and a partial logistic artificial neural network with automatic relevance determination.

Chiang, Shieh, Hsu & Wong (2005) built decision support system by using fuzzy classification trees which integrate decision tree techniques and fuzzy classifications, provide the efficient way to classify the data in order to generate the model for polyp screening.

CHAPTER THREE

DATA WAREHOUSING, DATA MINING, DECISION SUPPORT SYSTEMS

This section briefly describes the data warehouse, data mining, decision support systems and their concepts.

3.1 Data Warehousing

Data warehouse usage in health care services has slowed down due to the lack of understanding of the advantages of technology. Clinical data warehousing will become very important in the near future; doctors and health enterprises need to gain more information their clinical and administrative data, in order to improve quality of treatments and reduce costs. Healthcare organisations practicing evidence-based medicine strive to unite their data assets in order to achieve a wider knowledge base for more sophisticated research as well as to provide a matured decision support service for the care givers. The central point of such an integrated system is a data warehouse, to which all participants have access (Stolba, Banek & Tjoa, 2006).

3.1.1 Brief Characterization of Data Warehousing

The term “Data Warehouse” was first used by Devlin & Murphy (1988), but Inmon (1996) has won the most acclaim for introducing the concept, defined as follows. “A Data Warehouse is a *subject oriented, integrated, non-volatile and time-variant* collection of data in support of *management’s decisions*.” Let us have a closer look at these interesting properties.

- **Subject-oriented** means that all relevant data about a subject is gathered and stored as a single set in a useful format. Information is presented according to specific subjects or areas of interest.

- **Integrated** refers to data being stored in a globally acceptable fashion with consistent naming conventions, measurements, encoding structures, and physical attributes, even when the underlying operational systems store the data differently.
- **Non-volatile** means stable information that does not change each time an operational process is executed.
- **Time-variant** means that the data warehouse contains a history of the subject, as well as current information. Data warehouse data represents long-term data from five to ten years in contrast to the 30 to 60 day time period of operational data.

Data warehousing is a process requiring a set of hardware and software components that can be used to analyse the massive amounts of data that organisations, companies and research disciplines are accumulating to make better operational and/or strategic decisions. The data warehousing process does not consist of just adding data to the data warehouse, but also requires the architecture and tools to collect, query, analyse and present information. "Data warehousing is a process, not a product, for assembling and managing data from various sources for the purpose of gaining a single, detailed view of part or all of a business" (Stephen,1998).

3.1.2 Data Warehouse Architecture

According to Kimball & Casertam (2004) "a data warehouse is a system that extracts, cleans, conforms, and delivers source data into a dimensional data store and then supports and implements querying and analysis for the purpose of decision making". In general, there are three basic layers in the data warehouse architecture (Figure 3.1).

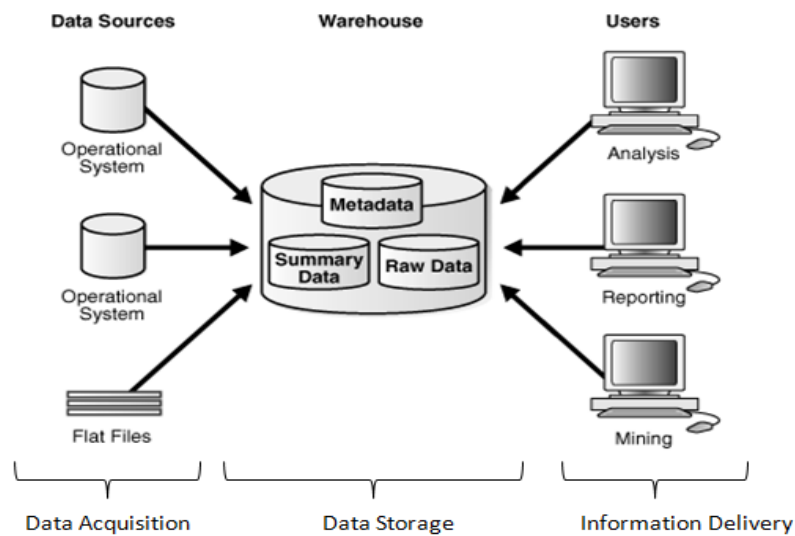


Figure 3.1 Data warehouse layers

- **Data Acquisition Layer:** In this layer, medical records, blood tests, urinalysis results, x-ray results of patients collected from different sources, have been corrected and brought to certain formats. At this stage, the data are made ready for loading into the data warehouse by filtering out noise, inconsistencies and irregularities. This is the longest and time-consuming phase of design.
- **Data Storage Layer:** In this layer, the data from a lower layer is made meaningful for the analysis studies and uploaded to the data warehouse. The data are held in the data warehouse according to multidimensional modeling approaches (star, snowflake, galaxy schema model, etc.). In this way, it is possible to approach the data from the different perspective.
- **Information Delivery:** In this layer, the stored data is used for purposes such as analysis, inquiry, reporting, data mining, decision support. The data in the data warehouse can be divided into separate lower storages according to the purposes to be used. Multidimensional analysis of data, OLAP operations, reports such as tables and graphs, information discovery from hidden patterns in the data is performed at this stage.

3.1.3 Data Warehouse Modeling

The data warehouse schema is composed as a set of levels organized into dimensions, as well as a set of fact relationships. A fact table is connected to three or more dimension tables. The term “fact” represents a business measure. A typical fact table has two types of columns, two or more foreign keys that connect primary keys and numeric attributes of the dimension tables (known as measurements). The measurement attributes contain numerical data that are analyzed using the different perspectives represented by the dimensions. These can be classified as additive, semi-additive, or non-additive (Malinowski & Zimanyi, 2007). Whereas additive facts can be aggregated through simple arithmetical addition, semi-additive facts can be aggregated along some of the dimensions but not along others. Non-additive facts, such as ratios, cannot be added.

Two of the most popular and well-known schemas for implementing the multidimensional model are the star schema and the snowflake schema. A star schema adopts the relational model for the representation of multidimensional data consisting of a fact table and a number of denormalized dimension tables. Figure 3.2 illustrates an example of a star schema about medical records. The granularity of the measured values in the “Admission” fact table rows in Figure 3.2 is patient, diagnosis, therapy, and date. A single row then records all patients with a single diagnosis from a single therapy type that was applied on a single day. Other possible granularities for the time dimension include weeks, months, and years. The choice of granularity affects the level of detail for the recorded information and the size of the database.

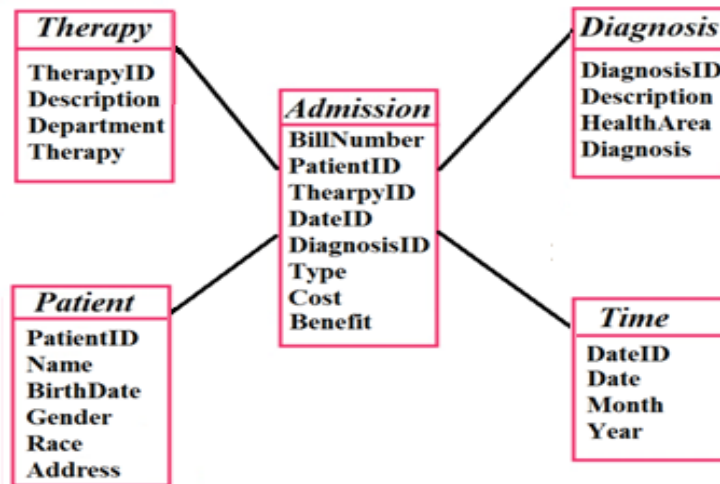


Figure 3.2 Star schema about medical records

The dimension is a structure that shares a common semantic within the domain being modeled, and is composed of one or more hierarchies. Dimensional attributes are descriptive, textual values for describing the dimensional value. Dimension tables often have fewer rows than fact tables, but are wide with many text columns. Each dimension is defined by a single primary key that serves as the basis for the referential integrity with any given fact table to which it is joined (Kimball & Ross, 1996).

A snowflake schema is obtained from a star schema, and is made up of a fact table and many dimensions related to that fact. In this type of schema, dimension data are grouped into multiple (totally or partially) normalized tables instead of one large table. An example of a patient admission snowflake schema is shown in Figure 3.3. While this saves space, it increases the number of dimension tables and requires other foreign key joins.

A hierarchy refers to a dimension and consists of many-to-one associations between the numbers of attributes within these dimensions (Garani & Helmer, 2012). A dimension can be composed of more than one hierarchy, which use ordered levels to organize the data. A level represents a position in a hierarchy. Dimension hierarchies group the levels from general to granular.

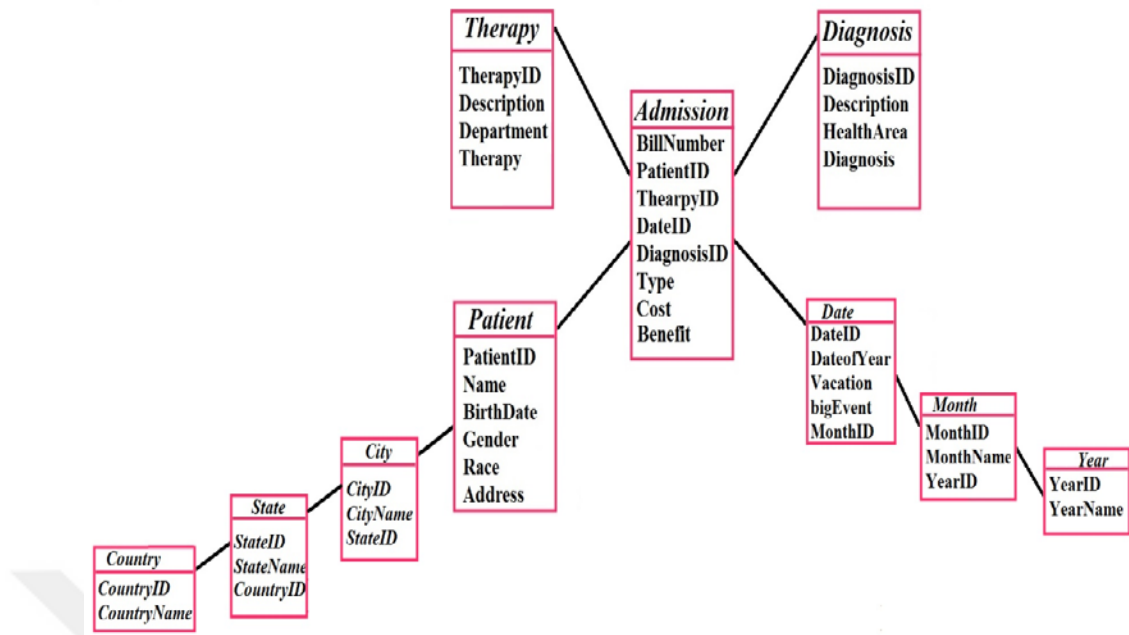


Figure 3.3 Snowflake schema about medical records

3.2 Data Mining

In recent years, the use of data mining has increased in the healthcare field as in every sector. Data mining tries to reveal hidden patterns in data that are difficult to perceive with traditional statistical methods. The health industry has a lot of data that holds complex information about patients and their medical condition. Since data in the health field is huge, multidimensional and scattered, it is difficult to understand the full potential of data without data mining.

The usage of data mining in healthcare can be divided into four general categories.

- **Treatment effectiveness:** Data mining applications can develop to evaluate the effectiveness of medical treatments. Data mining can deliver an analysis of which course of action proves effective by comparing and contrasting causes, symptoms, and courses of treatments.

- **Healthcare management:** Data mining applications can be developed to identify, track chronic disease states for the high-risk patients, design appropriate interventions, and reduce the number of hospital admissions and claim to aid healthcare management. Data mining might also be to analyze massive volumes of data and statistics to search for patterns that might indicate an attack by bio-terrorists.
- **Customer relationship management:** Customer relationship management is a core approach for managing interactions between commercial organizations; typically banks and retailers-and their customers. Customer interactions may occur through call centers, hospitals, billing departments, inpatient settings, and ambulatory care settings.
- **Fraud and abuse:** Detect fraud and abuses establish norms and then identify unusual or abnormal patterns of claims by physicians, clinics, or others attempt in data mining applications. Data mining applications fraud and abuse applications can highlight inappropriate prescriptions or referrals and fraudulent insurance and medical claims. (Subbarao , Khan & Kumar, 2016)

3.2.1 Data Mining Process

According to Fayyad, Shapiro & Smyth (1996), the data mining process involves numerous steps. Namely; learning the application domain, creating a target data set, data cleaning and preprocessing, data reduction and projection, choosing the function of data mining, choosing the data mining algorithm, data mining, interpretation and using discovered knowledge, as shown in Figure 3.4.

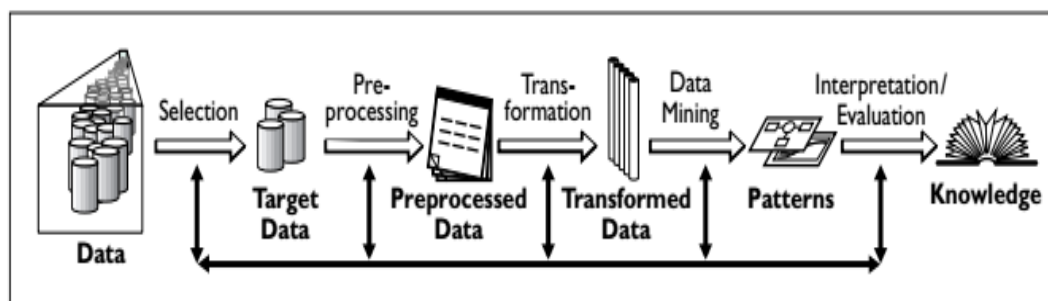


Figure 3.4 Overview of the steps constituting the data mining process

3.2.2 Classification Algorithms

There are special data mining algorithms for different purposes. The main objectives are classifying the new data, clustering similar cases, finding outliers and discovering interesting patterns specific to the data. In this work, the data mining's classification task was used. Classification algorithms have been used to evaluate lung cancer patients according to their chances of survival. The algorithms applied to the lung cancer data set are Naive Bayes, K-Nearest Neighbors and C4.5 Decision Tree.

3.2.2.1 Naive Bayes

Naive Bayes classifier is a probabilistic algorithm based on the Bayes theorem. Rather than predictions, the Naive Bayes algorithm produces probability estimates. For each class value they estimate the probability that a given instance belongs to that class. Requiring a small amount of training data to estimate the parameters necessary for classification is the advantage of the Naive Bayes classifier. It assumes that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence (Han & Kamber, 2000).

Given a hypothesis h and data D which concerned with the hypothesis:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

The diagram illustrates the components of Bayes' theorem. The equation $P(h|D) = \frac{P(D|h)P(h)}{P(D)}$ is shown. Arrows point from labels to the corresponding parts of the equation: 'posterior probability' points to $P(h|D)$, 'likelihood' points to $P(D|h)$, 'prior probability' points to $P(h)$, and 'data evidence' points to $P(D)$.

Figure 3.5 Bayes theorem

The purpose of Bayes Theorem (Figure 3.5) is to determine the most probable hypothesis from the given data D.

Prior probability of h, P(h): is the probability of being h is a correct hypothesis.

Prior probability of D, P(D): is the probability of training data D will be observed.

Conditional Probability of observation D, P(D|h): is the probability of observing data.

D given some world in which hypothesis h holds. (Barber, 2010)

3.2.2.2 K-Nearest Neighbors (KNN)

This algorithm looks at the nearest k of previously categorized points to classify a new point. When calculating the distance, the Euclidean distance formula is usually used. Euclidean distance calculated as the square root of the sum of the squares of the differences between points.

$$\text{Euclidean Distance} = \sqrt{\sum_{i=1}^k (x_i - y_i)^2} \quad (3.1)$$

If the k nearest neighbor is in which class, the new point is assigned to that class. For example; if k is taken as 8 in Figure 3.6, there are four patterns in category 1, two patterns in category 2, two patterns in category 3. Pluralities of patterns are in category 1 so decide x is in category 1.

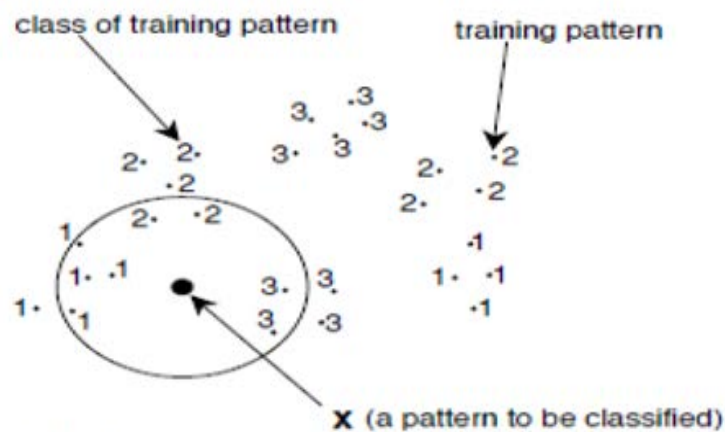


Figure 3.6 KNN Algorithm

3.2.2.3 C4.5 Decision Tree Algorithm

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller subsets while at the same time an associated decision tree is incrementally developed. The result is a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data. (Decision Tree, n.d.)

The data classification process is performed in two steps, training and classification. In the learning phase, the data is trained by the classification algorithm and some rules are derived in Figure 3.7.

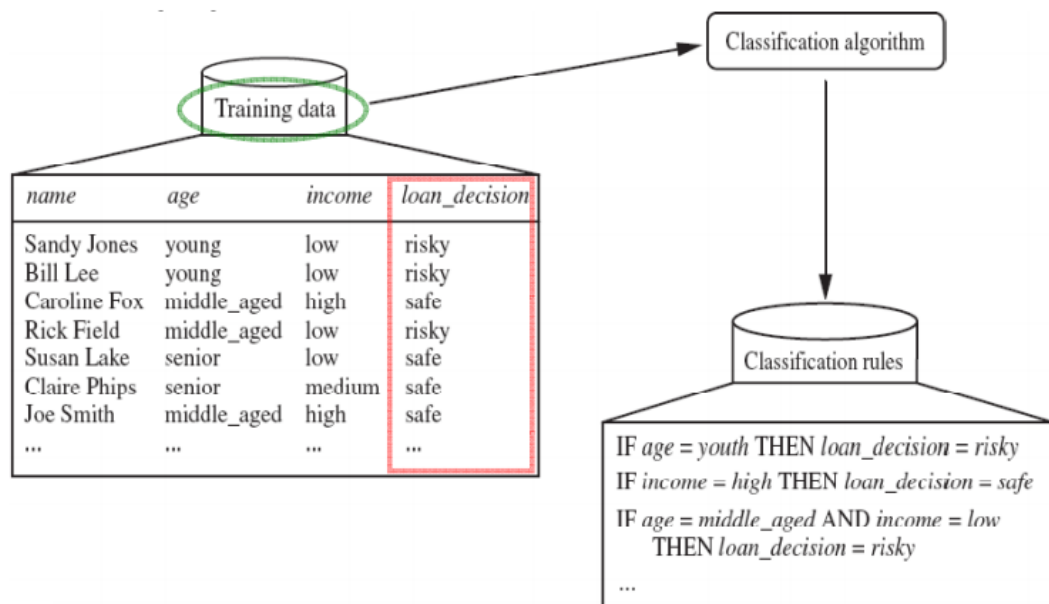


Figure 3.7 Learning phase of the decision tree algorithm

In the classification phase, the rules obtained in the previous step are tested. If the accuracy is acceptable, the model can be used to classify the new data in Figure 3.8.

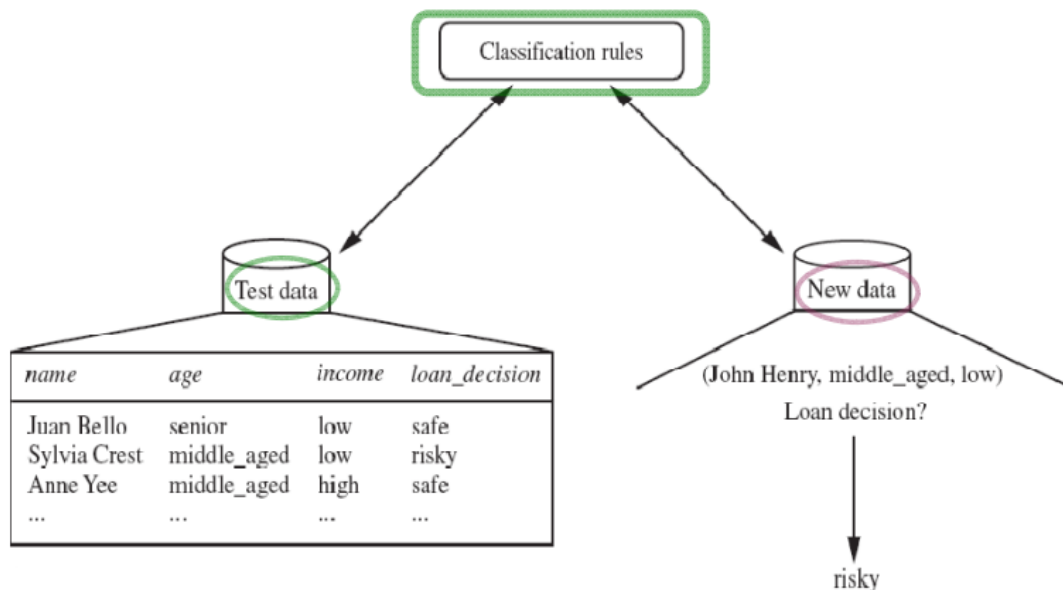


Figure 3.8 Classification phase of the decision tree algorithm

Decision tree J48 implements Quinlan's C4.5 algorithm for generating a pruned or unpruned C4.5 tree. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by J48 can be used for classification. J48 builds decision trees from a set of labeled training data using the concept of information entropy. It uses the fact that each attribute of the data can be used to make a decision by splitting the data into smaller subsets (Quinlan, 1993)

Input: The training samples, samples, represented by discrete-valued attributes; the set of candidate attributes, attribute-list.
Output: A decision tree.
Method:

```

create a node N;
if samples are all of the same class, C then
    return N as a leafnode labeled with the class C;
if attribute-list is empty then
    return N as a leafnode labeled with the most common class in samples; // majority voting
select test-attribute, the attribute among attribute-list with the highest information gain;
label node N with test-attribute;
for each known value ai of test-attribute // partition the samples
    grow a branch from node N for the condition test-attribute=ai;
    let si be the set of samples in samples for which test-attribute=ai; // a partition
    if si is empty then
        attach a leaf labeled with the most common class in samples;
    else attach the node returned by Generate decision tree(si, attribute-list - test-attribute);

```

Figure 3.9 Basic algorithm for decision tree induction (Han & Kamber,2000).

This algorithm uses the “Entropy”, a measure of the disorder of the data. The Entropy is calculated by this formula.

$$\text{Entropy}(S) = \sum_{i=1}^n - P(C_i) \times \log_2 P(C_i) \quad (3.2)$$

where S is the dataset, n is the class number in S , C_i is the i^{th} class in S , $P(C_i)$ is the frequency of C_i in S .

Gain is defined by

$$\text{Gain}(S,A) = \text{Entropy}(S) - \sum_{i=1}^m P(A_i) \text{Entropy}(S_{A_i}) \quad (3.3)$$

where $\text{Gain}(S,A)$ is the gain of S after a split on attribute A , m is the number of values of attribute A in S , $P(A_i)$ is the frequency of cases that have A_i value in S , $\text{Entropy}(S_{A_i})$ is the subset of S with items that have A_i value.

3.3 Challenges of Data Warehousing and Data Mining in the Healthcare

Developing a data warehouse in the healthcare field is a study that requires much more effort than other sectors. Clinical data warehouse requires advanced temporal support, classification structures, dimensional reduction of data, overcoming continuously changed and complex data. Table 3.1 compares the characteristics of conventional and clinical data warehousing.

Table 3.1 Characteristics of conventional vs clinical data warehouses (Pedersen & Jensen, 1998)

Characteristics	Conventional	Clinical
Data Model	Simple	Complex
Temporal	Medium	Advanced
Classifications	Simple	Advanced
Continuously Valued Data	No	Yes
Dimensionally Reduced	No	Yes

Table 3.1 Characteristics of conventional vs clinical data warehouses (Pedersen & Jensen, 1998) (Cont.)

Data		
Very Complex Data	No	Yes
Advanced Business Rules	Maybe	Yes (Protocols)
Data Mining	Maybe	Yes (Medical Research)

The biggest challenge in implementing data mining in the healthcare sector is the large amount of multidimensional and heterogeneous medical data. These data are interviews with patients, laboratory results, doctor's opinions collected from various sources; therefore the whole process is very time-consuming and troublesome to integrate them reliably. Inconsistent data coming from various data sources in different formats constitutes a big challenge for successful data mining. Doctors' clinical interpretations are not structured, which causes non-standard data additionally. Patients' not willing to share health data for study and missing entries is obstacle for data mining. Furthermore, health professionals and IT specialist may need to work together when domain knowledge is required for a successful data mining in the healthcare field.

3.4 Decision Support Systems

According to a study conducted by the Institute of Medicine in 2000, 98.000 patients die each year due to medical errors in the United States. The aim of the Institute is to minimize this number with the clinical decision support systems developed (Koç, Şengül & Özkaya, 2012).

A clinical decision support system (CDSS) is any computer designed to help health professionals make clinical decisions. In a sense, any computer system that deals with clinical data or medical knowledge is intended to provide decision support. It is accordingly useful to consider three types of decision support functions; tools for information management, tools for focusing attention, tools for patient-specific consultation (Shortliffe, 1987).

According to Perreault & Metzger, (1999) the four basic functions of the decision support system are summarized:

- Administrative: Supporting clinical coding and documentation, authorization of procedures, and referrals.
- Managing clinical complexity and details: Keeping patients on research and chemotherapy protocols; tracking orders, referrals follow-up and preventive care.
- Cost control: Monitoring medication orders; avoiding duplicate or unnecessary tests.
- Decision support: Supporting clinical diagnosis and treatment plan processes; and promoting the use of best practices, condition-specific guidelines, and population-based management."

CDSS provide an increase in the number of alternatives that are tested with proficient features such as early warning, rapid response, instant analysis, cost reduction, correct decision, effective team work, time savings and good use of data resources in unexpected situations in the clinical departments (Özata & Aslan, 2004).

Many potential benefits from CDSS have been widely reported throughout the literature (Johnson & Feldman, 1995). The claims made fall into three broad categories (Sintchenko et al., 2002).

- Improved patient safety, e.g. through reduced medication errors and adverse events and improved medication and test ordering.
- Improved quality of care, e.g. by increasing clinicians' available time for direct patient care, increased application of clinical pathways and guidelines.
- Improved efficiency in health care delivery e.g. by reducing costs through faster order processing, reductions in test duplication, decreased adverse events, and changed patterns of drug prescribing favouring cheaper but equally effective generic brands.

CHAPTER FOUR

USED TECHNOLOGIES

This chapter explains used technologies during the development process throughout the project.

4.1 Visual Studio 2012

Microsoft Visual Studio is an integrated development environment (IDE) from Microsoft. It is used to develop computer programs for Microsoft Windows superfamily of operating systems, as well as web sites, web applications and web services. Visual Studio uses Microsoft's software development platforms such as Windows API, Windows Forms, Windows Presentation Foundation, Windows Store and Microsoft Silverlight. It can produce both native code and managed code. Visual Studio includes a code editor supporting IntelliSense as well as code refactoring. The integrated debugger works both as a source-level debugger and a machine-level debugger. Other built-in tools include a forms designer for building GUI applications, web designer, class designer, and database schema designer. It accepts plug-ins that enhance the functionality at almost every level—including adding support for source-control systems (like Subversion) and adding new toolsets like editors and visual designers for domain-specific languages or toolsets for other aspects of the software development lifecycle (like the Team Foundation Server client: Team Explorer).


Visual Studio supports different programming languages and allows the code editor and debugger to support (to varying degrees) nearly any programming language, provided a language-specific service exists. Built-in languages include C, C++ and C++/CLI (via Visual C++), VB.NET (via Visual Basic .NET), C# (via Visual C#), and F# (as of Visual Studio 2010). Support for other languages such as M, Python, and Ruby among others is available via language services installed

separately. It also supports XML/XSLT, HTML/XHTML, JavaScript and CSS (Visual Studio 2010, n.d.).

4.2 ASP.NET MVC 4

The ASP.NET MVC is a web application framework developed by Microsoft, which implements the model–view–controller (MVC) pattern. It is open-source software, apart from the ASP.NET Web Forms component which is proprietary.

ASP.NET MVC allows software developers to build a web application as a composition of three roles: Model, View and Controller. The MVC model defines web applications with three logic layers:

- 
- Model (business layer)
 - View (display layer)
 - Controller (input control)

A model represents the state to a particular aspect of the application. A controller handles interactions and updates the model to reflect a change in state of the application, and then passes information to the view. A view accepts necessary information from the controller and renders a user interface to display that information.

The view engine used in this project is Razor. Razor is an ASP.NET programming syntax used to create dynamic web pages with the C# programming language. The Razor syntax is template markup syntax, based on the C# programming language, that enables the programmer to use an HTML construction workflow. Instead of using the ASP.NET Web Forms (.aspx) markup syntax with `<%= %>` symbols to indicate code blocks, Razor syntax starts code blocks with a `@` character and does not require explicit closing of the code-block.

The idea behind Razor is to provide an optimized syntax for HTML generation using a code-focused templating approach, with minimal transition between HTML and code. The design reduces the number of characters and keystrokes, and enables a more fluid coding workflow by not requiring explicitly denoted server blocks within the HTML code (MVC, n.d.).

4.3 MS Sql Server Management Studio 2014

Microsoft SQL Server is a relational database management system developed by Microsoft. As a database server, it is a software product whose primary function is to store and retrieve data as requested by other software applications, be it those on the same computer or those running on another computer across a network (including the Internet). There are at least a dozen different editions of Microsoft SQL Server aimed at different audiences and for workloads ranging from small single-machine applications to large Internet-facing applications with many concurrent users. Its primary query languages are T-SQL and ANSI SQL (Sql Server, n.d.).

4.4 SAS University Edition

SAS University Edition includes the SAS products Base SAS, SAS/STAT, SAS/IML, SAS/ACCESS Interface to PC Files, and SAS Studio. Teachers, students, adult learners, and academic researchers can access the SAS University Edition software for noncommercial learning purposes. You can download SAS University Edition for free, directly from SAS. After you download it to your PC, Mac, or Linux workstation, SAS works locally on your machine by using virtualization software and your browser, so no Internet access is required (SAS University Edition n.d.).

SAS Studio is a developmental web application for SAS that you access through your web browser. With SAS Studio, you can access your data files, libraries, and existing programs, and you can write new programs. You can also use the predefined tasks in SAS Studio to generate SAS code for you. When you run a program or task,

SAS Studio processes the SAS code on an SAS server. The SAS server can be a server in a cloud environment, a server in your local environment, or SAS installed on your local machine. After the code is processed, the results are returned to SAS Studio in your browser (SAS Studio n.d.).

4.5 Weka

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes (Weka n.d.).

4.6 Used Programming Languages

Decision support system was developed in C# programming language. C# is a simple, modern, general-purpose, object-oriented programming language. HTML is the standard markup language used to create web pages. CSS is a style sheet language used for describing the look and formatting of a document written in a markup language while most often used to style web pages and interfaces written in HTML. JavaScript is commonly used as part of web browsers, whose implementations allow client-side scripts to interact with the user, control the browser, communicate asynchronously, and alter the document content that is displayed. Lastly, the SAS programming language was also used.

CHAPTER FIVE

PROPOSED PROJECT

In this study, it is aimed to develop a data mining based prototype decision support system using a data warehouse. The data required for the study were obtained from the American National Cancer Institute's Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial. Within the scope of this thesis, study we only worked with lung cancer data. The study consists of three phases as in Figure 5.1.

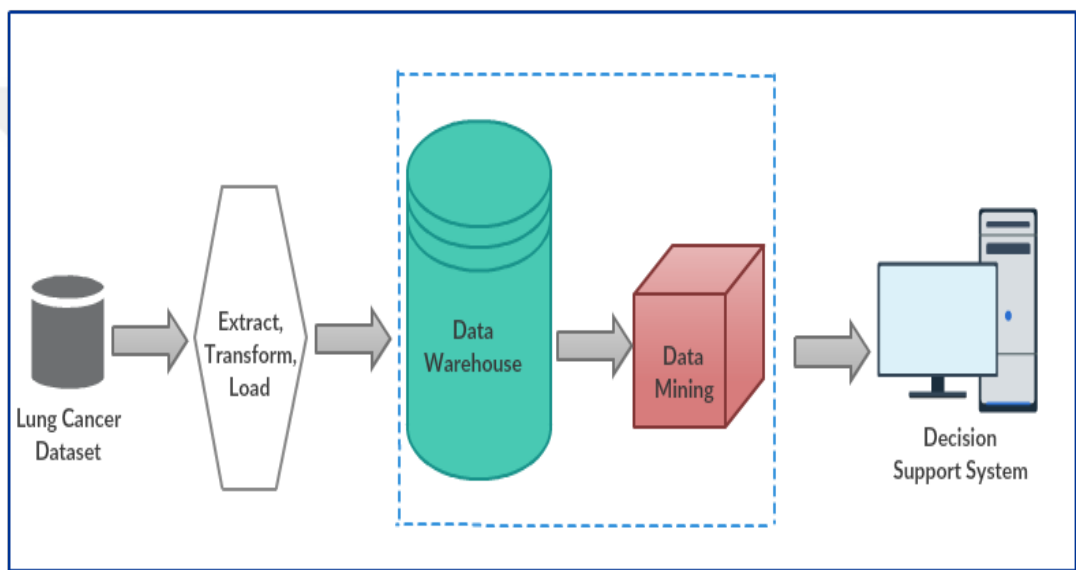


Figure 5.1 Proposed project architecture

At the first stage, lung cancer data were cleaned from mistakes, repetitions and was held in the data warehouse. It is ensured that the data warehouse conforms to the subject-oriented, integrated, time-variant and non-volatile conditions of the data warehouse approach presented by Inmon (1996). When designing the data warehouse, star schema model, which is a multidimensional data modeling approach, was used. The model is capable of answering complex queries and its results presented.

At the second stage, the data held in the data warehouse was converted to arff format for use in the data mining. Naive Bayes, KNN and C4.5 classification

algorithms were applied on the data in the arff file using the Weka program. Classification is based on whether patients die from lung cancer. According to the accuracy percentage, the best result was obtained by the C4.5 algorithm.

At the third stage, prototype decision support system was developed by using the rules of a decision tree obtained by the C4.5 algorithm. System can provide insight to the doctors so will shape the process of the treatment.

In this section, used data set is defined, data warehouse, data mining and decision support system processes in the project are explained.

5.1 Dataset Description

The data needed during the study were taken from the NCI (National Cancer Institute), the leading institution for cancer research in the United States. The NCI coordinates the National Cancer Program, which conducts and supports research, training, health information dissemination, and other programs with respect to the cause, diagnosis, prevention, and treatment of cancer, rehabilitation from cancer, and the continuing care of cancer patients and the families of cancer patients. The NCI provides research grants and cooperative agreements to coordinate and supports research projects conducted by universities, hospitals, research foundations, and businesses. The NCI supports education and training in fundamental sciences and clinical disciplines for participation in basic and clinical research programs and treatment programs relating to cancer. Because of the work of NCI scientists and cancer researchers throughout the United States and the rest of the world, real progress is being made against cancer. In the United States, the rate of new cancer cases overall has been declining since 1999, and the rate of cancer deaths overall has been decreasing for more than a decade. These trends reflect improvements in cancer treatment and advances in technology that have led to better tools for understanding, detecting, and diagnosing cancer. People with cancer are living longer and have a

better quality of life than ever before. In 2012, there were about 14 million cancer survivors in the United States (NCI, n.d.).

Firstly, the project idea was presented to NCI, which has large data sets of cancer, and approval was obtained. Then the dataset and data explanation guides were sent to us via the CDAS (Cancer Data Access System). CDAS is a website for requesting data recorded from some cancer studies. CDAS provides extensive public documentation for each study, including a trial summary, an overview of the data collected, and a searchable database of research projects and publications. If you are interested in obtaining study data, you may submit a project proposal. All projects are reviewed by NCI trial leadership. Upon approval, you will be granted access to the requested data for a limited period (CDAS, n.d.).

The PLCO (Prostate, Lung, Colorectal and Ovarian) data set was requested that is a large-scale, randomized study to determine whether certain screening tests will reduce the number of deaths from these cancers. PLCO is being conducted at ten sites, geographically and demographically disparate, around the U.S. There are approximately 155,000 male and female participants between the ages of 55 and 74 from 1993 to 2001. Data were collected on cancer diagnoses and deaths from all causes that occurred through December 31, 2009. Median follow-up time was 12.4 years.

Within the scope of this thesis, only the data on the patients with lung cancer were studied. This dataset contains information on 3594 lung cancer patients. These data include demographic information about the patient, past medical history, general health status, smoking history, treatment modalities, and cancer characteristics. In order to minimize the error margin and get clearer results, operations such as filtering, conversion, integration, reductions are applied to the data. As a result, we have worked on 69 attributes, one of which is a class. Some attribute examples from the data set are given in Table 5.1.

Table 5.1 Some attribute examples from the dataset

Attribute	Description	Text Format
agelevel	Patient Age Level	0="≤59", 1="60-64", 2="65-69", 3="≥70"
sqx_fh_lung	Family History of Lung Cancer	0="No", 1="Yes"
lung_stage_m	M Stage Component (Distant Metastases)	1="MX", 2="M0", 3="M1", 99="Not Available"
curative_radl	Had Radiation Treatment for Lung Cancer	0="No", 1="Yes"
bronchit_f	Did the participant ever have chronic bronchitis?	0="No", 1="Yes"
sqx_smk30days	Smoke in the Last 30 Days	1="Every Day", 2="Some Days", 3="Not at All"

5.2 Data Warehousing Stage

In recent years, the importance of data warehouses in medicine has become increasingly understood. Data warehouse is a repository that contains a large amount of clinical and administrative data used by physicians and healthcare institutions to increase the quality of care and at the same time reduce costs. Healthcare institutions are trying to bring their own data as integrated as possible in order to obtain a large data pool for research and to provide decision support services. With the data warehouse approach, data from different operational systems are collected in a common repository and made available for operations such as querying and analysis.

Cancer is one of the leading causes of death today. Most of these deaths are lung cancer. Therefore, using a data warehouse structure for serious disease, such as lung cancer, will allow doctors to look for the disease from a wider perspective. Time is very important element in the treatment process of this disease; the sooner the process starts, the better the results will be. Data warehouse will prevent loss of time in patient's history files. In this way, the doctor will be able to analyze the patient's

past data as he wishes, and more efficient treatment methods will be implemented since the disease can be monitored more easily.

5.2.1 Data Processing

Data processing is the most important and time-consuming part of the design of the data warehouse, which includes processes such as filtering, cleaning, transforming to ensure better quality and accurate results. In order to create a data warehouse model in the medical field, it is necessary to define the data very well. For this reason, the opinions of specialists on lung cancer have been taken primarily, and the data have become more understandable and easily interpretable by us.

In order to be able to examine the data better, it was transferred from the SAS data tables to Microsoft Excel spreadsheets using SAS University Edition. (Figure 5.2)

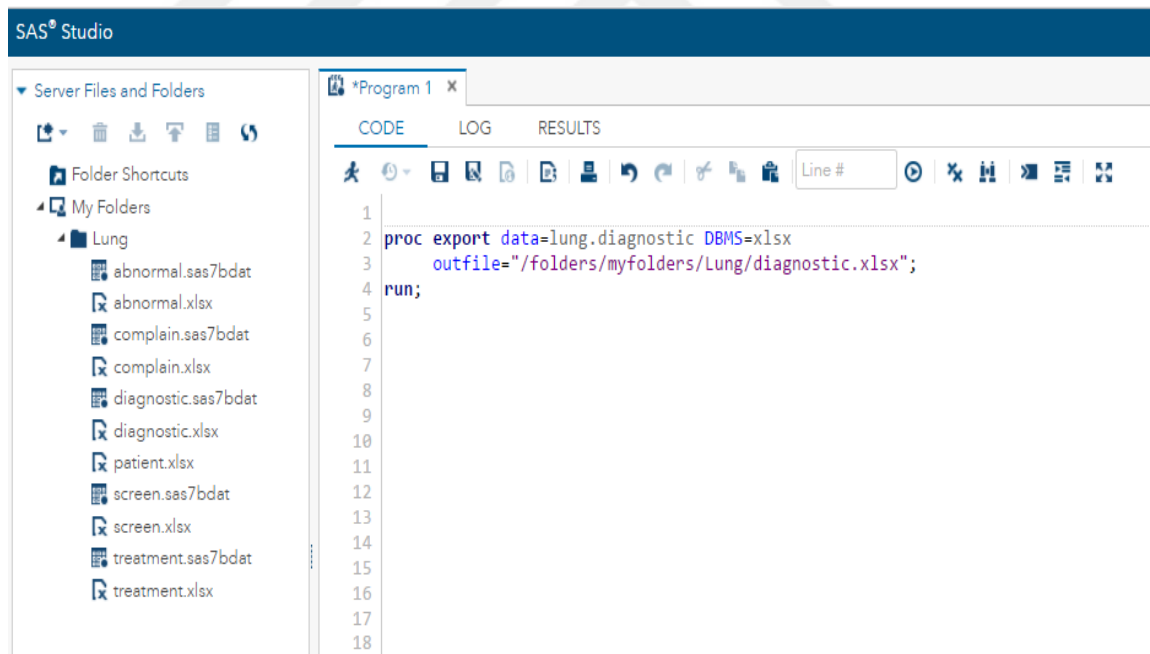


Figure 5.2 Data transfer from SAS data table to Excel spreadsheets

A temporary database has been created using Microsoft SQL Server because it will be easier to extract data with SQL queries. Using the import and export wizard, the data in the Excel spreadsheets is transferred to this database (Figure 5.3).

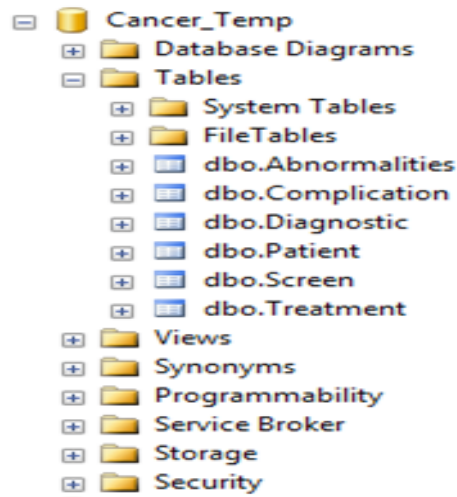


Figure 5.3 Temporary lung cancer database

After this phase, the data became better able to be examined. In this database, the data was cleared from the missing and incorrect ones and some transformation processes were applied to make it ready to upload to the data warehouse (Figure 5.4).

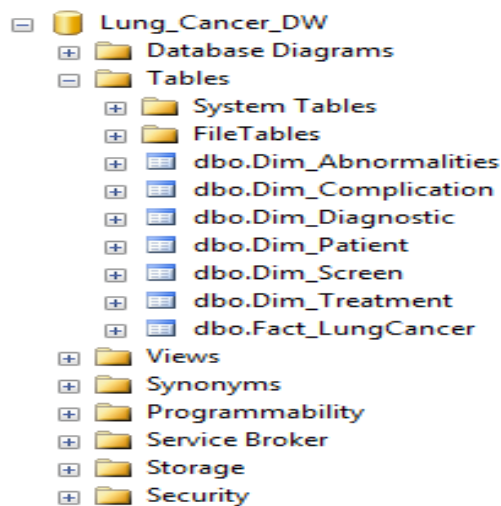


Figure 5.4 Lung cancer data warehouse

Example for conversion operations; in the Treatment table in the database, the treatment value is coded as 1 = chemotherapy, 2 = laser therapy, while in the Dim_Treatment dimension table of the data warehouse, this value is coded by converting it to 1 if the patient has received treatment and other treatment types are 0. In the same way, when the data is transferred from the database to the data warehouse, similar transformations have been made (Figure 5.5).

```

SELECT 'Trt_Num1_'+ cast(trt_num1 as varchar) as Trt_Num1, treatment_id as TreatmentID
INTO #temp FROM Cancer_Temp..Treatment
WHERE Trt_Num1 != '' or Trt_Num1 !=0

DECLARE @Trt_Num1 VARCHAR(50)
DECLARE @TreatmentID VARCHAR(50)
DECLARE db_cursor CURSOR FOR SELECT * FROM #temp

OPEN db_cursor
FETCH NEXT FROM db_cursor INTO @Trt_Num1, @TreatmentID

WHILE @@FETCH_STATUS = 0
BEGIN
    DECLARE @QRY1 NVARCHAR(MAX)
    SET @QRY1 = 'UPDATE Lung_Cancer_DW..Dim_Treatment set '+@Trt_Num1+'=1 where treatment_id='+@TreatmentID
    EXEC SP_EXECUTESQL @QRY1
    FETCH NEXT FROM db_cursor INTO @Trt_Num1, @TreatmentID
END

CLOSE db_cursor
DEALLOCATE db_cursor

```

Figure 5.5 Example for conversion operations

5.2.2 Data Storage

This phase includes the process of loading and storing the filtered data into the data warehouse. The data is stored in a star schema model, a multidimensional modeling approach, for performing OLAP operations. The multidimensional data model is based on concepts such as cube, dimension and hierarchy. This model allows users to visualize and sort out the results of the lung cancer data warehouse in a variety of forms. The star schema model used is easy to understand and suitable for query performance. This model was created by connecting small sub-tables (dimension table) , one for each dimension, to a large central table (fact table). The

lung cancer data warehouse model created using Microsoft SQL Server has six dimension tables (patient information, complications, diagnoses, treatments, x-ray results, and x-ray anomalies) and one lung cancer fact table. As shown in Figure 5.6, the fact table contains keys representing each of the dimension tables.

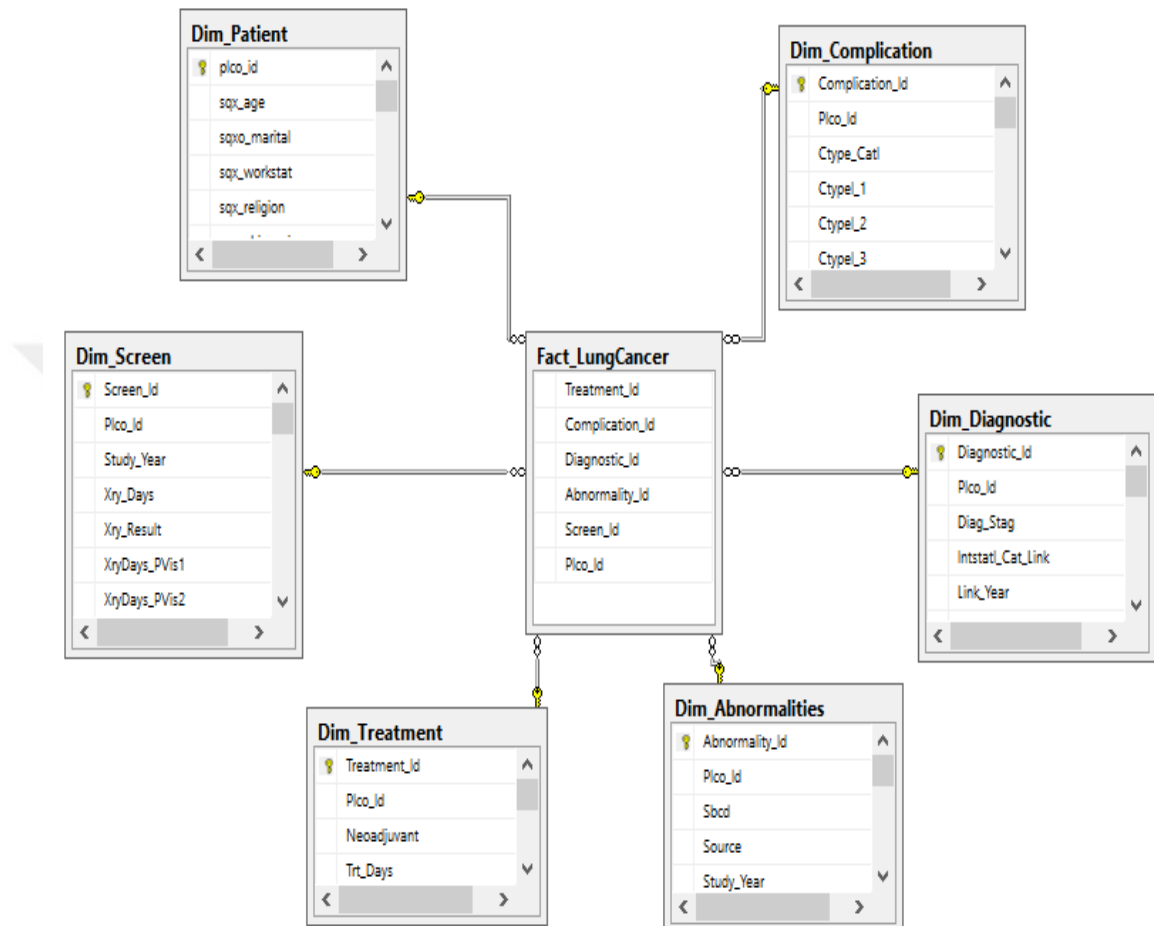


Figure 5.6 Lung cancer data warehouse star schema model

5.2.3 Examples of Querying in Data Warehouse

The established data warehouse is designed to respond for large and complex queries. It is intended to help medical analysts and doctors to make decisions through a diverse perspective. Below, the results for the sample queries written using the established data warehouse are shown in Table 5.2-5.4 respectively.

Query 1: In which patients who smoked more than three cigarettes a day and whose family has lung cancer history, was lung cancer detected?

```
select * from Dim_Patient
where sqx_amt_smk>3 and
sqx_fh_lung=1 and plco_id in
(select plco_id
from Dim_Diagnostic
where Proc_Res_3=1)
order by sqx_age
```

Table 5.2 Result of Query 1

	plco_id	sqx_age	sqx_amt_smk	sqx_height	sqx_fh_lung	sqxo_marital	sqx_workstat	sqx_religion	sqxbq_race_nh6	sqxo_race7	sqx_bmi_curr	sqx_wt_curr
1	R-000494-5	60	4	64	1	1	2	1	5	1	18,364501953125	107
2	R-064016-1	61	4	79	1	4	7	1	5	1	34,5811568658869	307
3	Q-141127-2	61	5	57	1	3	7	1	5	1	21,6374269005848	100
4	J-069569-5	61	6	72	1	1	7	1	5	1	22,3755787037037	165
5	A-121976-7	62	5	71	1	1	2	1	5	1	32,0749851219996	230
6	B-081981-5	62	4	62	1	1	2	1	5	1	25,9693028095734	142
7	E-094665-0	62	5	66	1	1	2	6	5	1	26,1446280991736	162
8	S-050814-5	62	5	67	1	1	2	4	5	1	25,83983069726	165
9	Q-037402-5	62	4	70	1	1	7	1	5	1	26,5418367346939	185
10	O-016601-8	62	5	70	1	1	7	1	5	1	33,5718367346939	234

Query 2: How is the distribution of patients with pneumothorax complication and chemotherapy treatment according to age?

```
select sqx_age as AGE,  
       count(plco_id) as COUNT from Dim_Patient P  
where plco_id in  
       (select T.Plco_Id from Dim_Treatment T  
        inner join Dim_Complication C on  
        T.Plco_Id= C.Plco_Id and  
        T.Trt_Familyl_4=1 and  
        C.Ctypel_3=1)  
group by sqx_age
```

Table 5.3 Result of Query 2

	AGE	COUNT
2	64	1
3	65	2
4	66	2
5	67	7
6	68	5
7	69	7
8	70	1
9	71	4
10	72	2
11	73	2
12	74	6
13	75	3
14	76	2
15	78	3
16	79	2
17	80	1

Query 3: How is the distribution of patients with nodules and biopsies according to the number of daily cigarette?

```
select sqx_amt_smk as CIGARETTE,  
count(plco_id) as COUNT  
from Dim_Patient P where plco_id in  
(select A.Plco_Id  
from Dim_Abnormalities A  
inner join Dim_Diagnostic D  
on A.Plco_Id= D.Plco_Id  
and A.Desc_1=1  
and D.Biop=1)  
group by sqx_amt_smk  
order by sqx_amt_smk asc
```

Table 5.4 Result of Query 3

	CIGARETTE	COUNT
2	1	5
3	2	7
4	3	11
5	4	1
6	5	5

5.3 Data Mining Stage

The established data warehouse, which can store large amounts of multidimensional data, facilitates the decision-making process by responding to complex queries and makes data available for the data mining techniques. Data mining has been defined as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. (Frawley, Shapiro, Christopher & Matheus, 1992)

Data mining techniques can be deployed for discovering the patterns of clinical pathways. Based on the patient record data, administrative data, clinical log data and evidence based rules, mining process is applied. With their usage, we can detect the structure of clinical paths and the sequence among activities, which human beings could hardly find. The development of clinical pathways is knowledge intensive, and it requires the cooperation among knowledge workers, clinicians, nurses and clinical management. (Stolba & Tjoa, 2007)

Throughout data mining, data in the data warehouse were converted into arff format and different classification algorithms were applied in Weka program. Various rules have been obtained from the survival probability of lung cancer patients by comparing the results of the applied algorithms.

5.3.1 Preparation of Arff File

An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for use of the Weka machine learning software (Weka n.d.).

ARFF files have two distinct sections. The first section is the Header information, which is followed the Data information. The Header of the ARFF file contains the

name of the relation, a list of the attributes (the columns on the data), and their types. The relation name is defined as the first line in the ARFF file. The format is: @relation <relation-name>, where <relation-name> is a string.

Attribute declarations take the form of an ordered sequence of @attribute statements. Each attribute in the data set has its own @attribute statement which uniquely defines the name of that attribute and its data type. The order the attributes are declared indicates the column position in the data section of the file. The format for the @attribute statement is: @attribute <attribute-name> <datatype> where the <attribute-name> must start with an alphabetic character. The definition part example of the arff file used in the project is as in Figure 5.7.

```

@relation LungCancer

@attribute num_cancel {1,2}
@attribute fin_smcl {0,1,2,9,10}
@attribute lung_stage {1,2,3,4,5,6,7,8,9,93,99}
@attribute lung_stage_t {1,4,5,6,7,99}
@attribute lung_stage_n {1,2,3,4,5,99}
@attribute lung_stage_m {1,2,3,99}
@attribute lung_clinstage {2,3,5,6,7,8,9,11,99}
@attribute lung_clinstage_t {1,4,5,6,7,99}
@attribute lung_clinstage_n {1,2,3,4,5,99}
@attribute lung_clinstage_m {1,2,3,99}
@attribute lung_pathstage {2,3,5,6,7,8,9,11,99}
@attribute lung_pathstage_t {1,4,5,6,7,99}
@attribute lung_pathstage_n {1,2,3,4,5,99}
@attribute lung_pathstage_m {1,2,3,99}
@attribute lung_grade {1,2,3,4,5,99}
@attribute lung_histtype {2,3,4,5,7,8,9,10,11,12,13,14,15,18,30,31,32,33,99}
@attribute lung_cancer_type {1,2}
@attribute lung_is_carcinoid {0,1}
@attribute curative_pneuml {0,1}
@attribute curative_well {0,1}
@attribute curative_chemol {0,1}
@attribute curative_radl {0,1}
@attribute neoadjuvantl {0,1}
@attribute primary_trtl_NSC {0,1,2,3,4,5,6,10,12}
@attribute primary_trtl_small {0,1,2,3,4,5,10,12}
.
.
.
@attribute sqx_smk_exp_child {1,2,3}
@attribute sqx_smk_exp_adult {1,2,3}
@attribute sqx_smk_exp_work {1,2,3}
@attribute class {0,1}

```

Figure 5.7 Arff file definition part

The ARFF Data section of the file contains the data declaration line and the actual instance lines. The @data declaration is a single line denoting the start of the data segment in the file. The format is: @data.

Each instance is represented on a single line, with carriage returns denoting the end of the instance. Attribute values for each instance are delimited by commas. They must appear on the order that they were declared in the header section (i.e. the data corresponding to the n^{th} @attribute declaration was always the n^{th} field of the attribute). Missing values are represented by a single question mark (Weka n.d.). The data part example of the arff file used in the project is as in Figure 5.8.

```
@data
1,10,2,4,2,2,2,4,2,2,2,4,2,2,2,4,2,1,0,0,1,0,0,0,1,12,0,0,0,0,0,0,0,0,0,1,1,3,1,1,3,1,1,3,?,?,3,?,?,21,?,?,2,0,1,1,1,2,2,1,4,1,2,1,0,1,?,1,1,2,0
1,10,6,5,3,2,6,5,3,2,6,5,3,2,99,7,1,0,0,1,0,0,0,1,12,0,0,0,0,0,0,0,1,1,0,2,1,2,3,1,1,9,?,?,3,?,?,21,?,?,1,2,1,1,1,3,2,3,?,?,1,0,0,0,1,3,2,0
1,10,8,6,5,2,7,6,4,2,8,1,5,2,3,7,1,0,0,0,1,1,0,5,12,0,0,0,0,0,0,0,0,1,1,2,1,1,2,0,2,2,?,3,3,?,16,21,?,2,2,0,1,1,3,2,1,3,2,2,0,0,0,?,3,1,2,1
1,10,6,6,2,2,2,4,2,2,6,6,2,2,4,7,1,0,0,1,1,0,0,3,12,2,2,1,1,0,0,0,0,1,0,5,3,2,3,1,1,9,?,?,3,?,?,21,?,?,1,1,?,1,1,3,2,3,?,?,?,1,0,0,0,3,3,1,0
1,10,8,6,5,2,7,6,4,2,8,1,5,2,3,7,1,0,0,0,1,1,0,5,12,0,0,0,0,0,0,0,0,1,1,2,1,2,0,2,2,?,3,3,?,16,21,?,2,2,0,1,1,3,2,1,3,2,2,0,0,0,?,3,1,2,1
1,10,9,5,3,3,9,5,3,3,99,99,99,99,99,7,1,0,0,0,1,1,0,5,12,0,0,0,0,0,0,0,0,1,1,2,2,1,3,1,1,9,?,?,3,?,?,21,?,?,1,0,0,?,1,3,2,1,2,2,2,0,0,0,?,1,3,3,1
```

Figure 5.8 Arff file data part

Prepared arff file loaded to the Weka program, as shown in Figure 5.9.

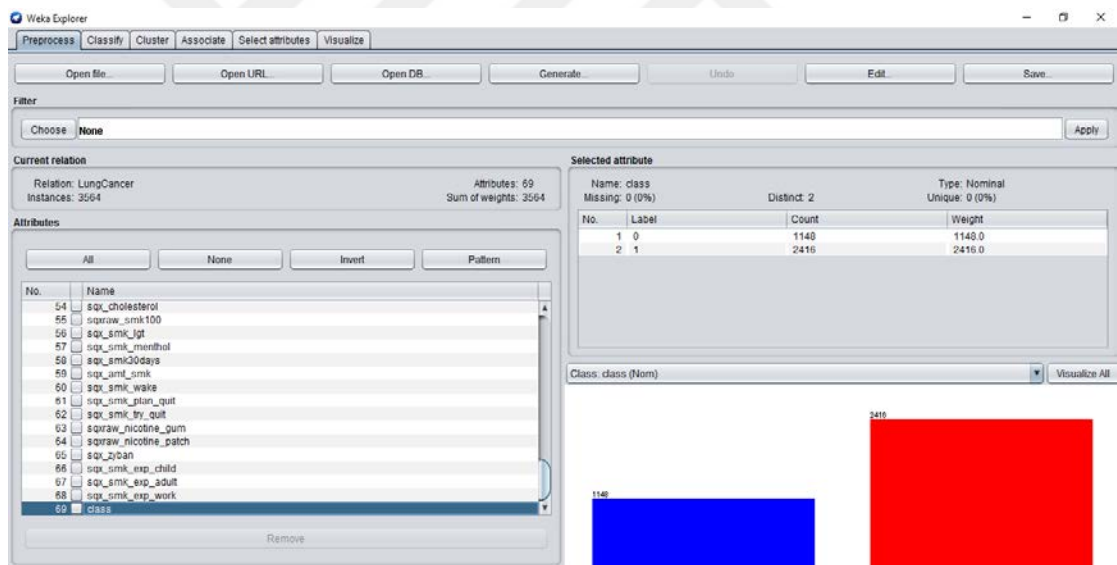


Figure 5.9 Imported dataset

5.3.2 Implementation of Classification Algorithms

After we have created the Arff file, we can view the distribution to the class according to attributes by the weka program. The red ones are those patients who died of lung cancer, and the blue ones are living patients Figure 5.10.

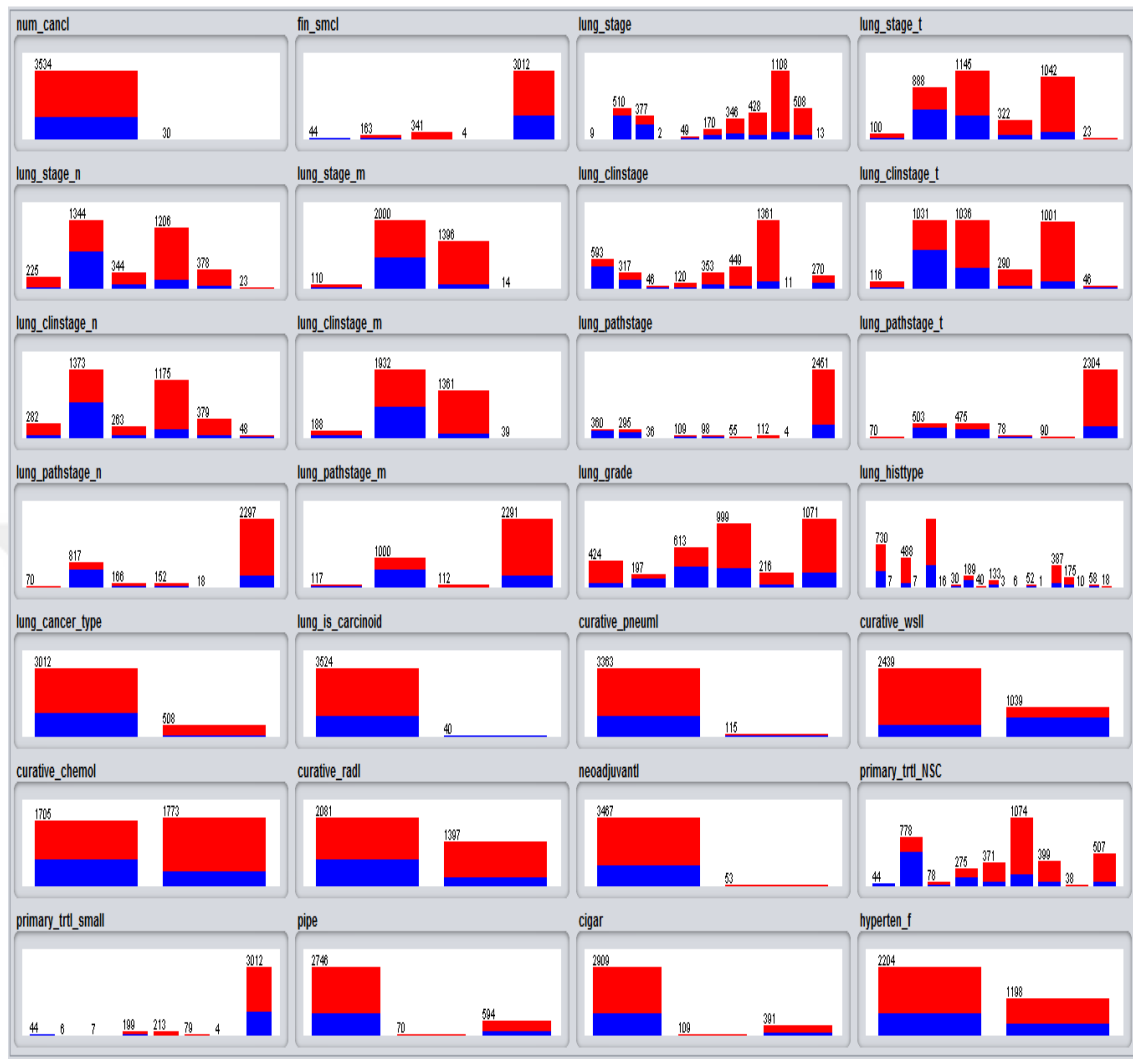


Figure 5.10 Distribution of class according to attributes

Classification is an important data mining technique with a wide range of applications for the classification of various data used in almost every area of our lives. Three algorithms have been applied to classify lung cancer patients according to their chances of survival. These are Naive Bayes, KNN and J48 (C4.5) algorithms. Each algorithm was evaluated by 10-fold cross validation method, and their output was presented.

5.3.3 Cross Validation

Cross validation allows us to use the entire data set for computation. The data set is divided into k subset and each time one of the k subclasses is used as a test set, the other k-1 subclasses are combined to form a training set. Then the average error in all k experiments is calculated. How the data is divided less important in this method, which is an advantage. Each data point enters exactly once into the test set, and training set k-1 times. 10-fold cross validation was used in this study.



Figure 5.11 10-fold cross validation

5.3.4 Implementation of Naive Bayes Algorithm

The Naive Bayes algorithm is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combinations of values in a given data set. The algorithm uses Bayes theorem and assumes all attributes to be independent given the value of the class variable. This conditional independence assumption rarely holds true in real world applications, hence the characterization as Naive yet the algorithm tends to perform well and learn rapidly in various supervised classification problems (Dimitoglou, Adams & Jim, 2012).

```

=== Summary ===

Correctly Classified Instances      2781           78.0303 %
Kappa statistic                    0.5081
Mean absolute error                 0.2198
Root mean squared error             0.457
Relative absolute error             50.3312 %
Root relative squared error         97.8085 %
Total Number of Instances          3564

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.702   0.183   0.646     0.702   0.673     0.509   0.827    0.718    0
                0.817   0.298   0.852     0.817   0.835     0.509   0.827    0.890    1
Weighted Avg.   0.780   0.261   0.786     0.780   0.783     0.509   0.827    0.835

=== Confusion Matrix ===

  a  b  <-- classified as
806 342 |  a = 0
441 1975 |  b = 1

```

Figure 5.12 Naive Bayes classification result

Naive Bayes classifier applied to the lung cancer dataset estimated the survival of lung cancer with 78.03% accuracy rate.

5.3.5 Implementation of K-Nearest Neighbors (KNN) Algorithm

The k-nearest neighbor is a classification (or regression) algorithm that combines the classifications of k nearest points to determine the class of a point.

```

=== Summary ===

Correctly Classified Instances      2603           73.0359 %
Kappa statistic                    0.4262
Mean absolute error                0.2651
Root mean squared error            0.4826
Relative absolute error            60.7049 %
Root relative squared error        103.2853 %
Total Number of Instances          3564

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
          0.726   0.268   0.563     0.726   0.634     0.435   0.769    0.587    0
          0.732   0.274   0.849     0.732   0.786     0.435   0.768    0.845    1
Weighted Avg.  0.730   0.272   0.757     0.730   0.737     0.435   0.769    0.762

=== Confusion Matrix ===

  a  b  <-- classified as
834 314 |  a = 0
647 1769 |  b = 1

```

Figure 5.13 KNN classification result

KNN classifier applied to the lung cancer data set estimated the survival of lung cancer with 73.03% accuracy rate. K value was taken as 1.

5.3.6 Implementation Of C4.5 (J48) Algorithm

The algorithm first identifies the dominant attribute of the training set and places it as the root of the tree. Second, it creates a leaf for each of the possible values the root can take. Then, for each of the leaves it repeats the process using the training set data classified by this leaf. The core function of the algorithm is determining the most appropriate attribute to best partition the data into various classes (Peddabachigari, Abraham, Grosan & Thomas, 2007).

```

=== Summary ===

Correctly Classified Instances      2810           78.844 %
Kappa statistic                    0.4974
Mean absolute error                 0.3018
Root mean squared error            0.4014
Relative absolute error            69.1066 %
Root relative squared error        85.9022 %
Total Number of Instances         3564

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC    ROC Area  PRC Area  Class
          0.603   0.123   0.699     0.603   0.647     0.500   0.780    0.611    0
          0.877   0.397   0.823     0.877   0.849     0.500   0.780    0.834    1
Weighted Avg.  0.788   0.309   0.783     0.788   0.784     0.500   0.780    0.762

=== Confusion Matrix ===

  a  b  <-- classified as
692 456 |  a = 0
298 2118 |  b = 1

```

Figure 5.14 C4.5 classification result

C4.5 classifier applied to the lung cancer data set estimated the survival of lung cancer with 78.84% accuracy rate.

J48 pruned tree

```

curative_wsl1 = 0
|   lung_is_carcinoid = 0
|   |   lung_clinstage = 2
|   |   |   sex = 1
|   |   |   |   lung_grade = 1: 0 (8.01/2.0)
|   |   |   |   lung_grade = 2: 0 (5.0)
|   |   |   |   lung_grade = 3: 0 (13.01/6.0)
|   |   |   |   lung_grade = 4: 0 (12.71/2.7)
|   |   |   |   lung_grade = 5: 1 (4.0/1.0)
|   |   |   |   lung_grade = 99: 1 (22.02/4.02)
|   |   |   sex = 2: 0 (43.17/14.7)
|   |   lung_clinstage = 3
|   |   |   neoadjuvant1 = 0
|   |   |   |   primary_trtl_NSC = 0: 0 (0.12)
|   |   |   |   primary_trtl_NSC = 1: 1 (0.0)
|   |   |   |   primary_trtl_NSC = 2: 1 (26.0/9.0)
|   |   |   |   primary_trtl_NSC = 3: 0 (13.0/3.0)
|   |   |   |   primary_trtl_NSC = 4: 0 (23.0/8.0)
|   |   |   |   primary_trtl_NSC = 5: 1 (20.0/3.0)
|   |   |   |   primary_trtl_NSC = 6: 1 (11.0/4.0)
|   |   |   |   primary_trtl_NSC = 10: 1 (0.0)
|   |   |   |   primary_trtl_NSC = 12: 1 (9.0/2.0)
|   |   |   neoadjuvant1 = 1: 1 (2.0/0.0)
|   |   lung_clinstage = 5: 0 (14.02/5.0)
|   |   lung_clinstage = 6
|   |   |   -
|   |   |   num_confirmed = 1: 1 (67.75/13.04)
|   |   |   num_confirmed = 2: 0 (8.04/3.0)
|   |   |   num_confirmed = 3: 0 (2.0/1.0)
|   |   |   num_confirmed = 4: 1 (0.0)
|   |   |   lung_clinstage = 7: 1 (290.43/78.03)
|   |   lung_clinstage = 8
|   |   |   confirmed_icdbeh1 = 1: 0 (1.0)
|   |   |   confirmed_icdbeh1 = 2: 0 (5.0/1.0)
|   |   |   confirmed_icdbeh1 = 3: 1 (417.88/55.18)
|   |   lung_clinstage = 9: 1 (1318.13/133.9)
|   |   lung_clinstage = 11: 0 (4.0/1.0)
|   |   lung_clinstage = 99: 1 (130.95/32.85)
|   lung_is_carcinoid = 1: 0 (28.05)
curative_wsl1 = 1
|   lung_stage_n = 1: 1 (9.51/3.42)
|   lung_stage_n = 2
|   |   curative_pneuml = 0
|   |   |   neoadjuvant1 = 0
|   |   |   |   lung_stage_m = 1: 0 (21.56/5.0)
|   |   |   |   lung_stage_m = 2: 0 (746.4/185.59)
|   |   |   |   lung_stage_m = 3: 1 (23.89/6.29)
|   |   |   |   lung_stage_m = 99: 0 (0.0)
|   |   |   neoadjuvant1 = 1: 1 (17.21/6.21)
|   |   curative_pneuml = 1: 1 (4.06/0.05)
|   lung_stage_n = 3
|   |   confirmed_can_type1 = 3: 0 (1.0)
|   |   confirmed_can_type1 = 6: 0 (2.03/1.0)
|   |   confirmed_can_type1 = 7: 1 (0.0)

```

```

| | confirmed_can_type1 = 7: 1 (0.0)
| | confirmed_can_type1 = 10: 1 (0.0)
| | confirmed_can_type1 = 11: 1 (0.0)
| | confirmed_can_type1 = 13: 1 (0.0)
| | confirmed_can_type1 = 14: 1 (0.0)
| | confirmed_can_type1 = 15: 1 (0.0)
| | confirmed_can_type1 = 16: 1 (0.0)
| | confirmed_can_type1 = 17: 0 (1.0)
| | confirmed_can_type1 = 18: 0 (1.0)
| | confirmed_can_type1 = 19: 1 (0.0)
| | confirmed_can_type1 = 20: 1 (0.0)
| | confirmed_can_type1 = 21
| | |   cigar = 0
| | |   |   emphys_f = 0
| | |   |   |   sex = 1: 1 (44.21/13.13)
| | |   |   |   sex = 2
| | |   |   |   |   agelevel = 0
| | |   |   |   |   |   curative_rad1 = 0: 0 (4.92)
| | |   |   |   |   |   curative_rad1 = 1: 1 (5.14/1.14)
| | |   |   |   |   |   agelevel = 1: 0 (12.63/4.3)
| | |   |   |   |   |   agelevel = 2: 1 (12.5/1.2)
| | |   |   |   |   |   agelevel = 3: 1 (1.1/0.1)
| | |   |   |   |   |   emphys_f = 1: 0 (7.87/1.84)
| | |   |   |   |   |   cigar = 1: 0 (2.14/0.04)
| | |   |   |   |   |   cigar = 2: 1 (11.64/1.41)
| | |   |   |   |   |   confirmed_can_type1 = 24: 0 (2.0)
| | |   |   |   |   |   confirmed_can_type1 = 25: 0 (1.0)
| | |   |   |   |   |   confirmed_can_type1 = 26: 0 (2.0/1.0)
| | |   |   |   |   |   confirmed_can_type1 = 29: 1 (0.0)
| | |   |   |   |   |   confirmed_can_type1 = 31: 0 (7.1/2.0)
| | |   |   |   |   |   confirmed_can_type1 = 33: 1 (0.0)
| | |   |   |   |   |   confirmed_can_type1 = 34: 1 (0.0)
| | |   |   |   |   |   confirmed_can_type1 = 37: 1 (0.0)
| | |   |   |   |   |   confirmed_can_type1 = 38: 1 (0.0)
| | |   |   |   |   |   confirmed_can_type1 = 39: 1 (0.0)
| | |   |   |   |   |   confirmed_can_type1 = 43: 1 (0.0)
| | |   |   |   |   |   confirmed_can_type1 = 44: 1 (0.3)
| | |   |   |   |   |   confirmed_can_type1 = 47: 1 (0.0)
| | |   |   |   |   |   confirmed_can_type1 = 50: 1 (0.0)
| | |   |   |   |   |   confirmed_can_type1 = 51: 1 (0.0)
| | |   |   |   |   |   confirmed_can_type1 = 53: 0 (1.0)
| | |   |   |   |   |   confirmed_can_type1 = 60: 1 (0.0)
| | |   |   |   |   |   confirmed_can_type1 = 80: 1 (1.0)
| | |   |   |   |   |   confirmed_can_type1 = 999: 1 (0.0)
| | |   |   |   |   |   lung_stage_n = 4: 1 (112.57/32.99)
| | |   |   |   |   |   lung_stage_n = 5: 1 (6.88/0.38)
| | |   |   |   |   |   lung_stage_n = 99: 1 (1.01/0.01)

```

Figure 5.15 Decision tree created by the C4.5 algorithm

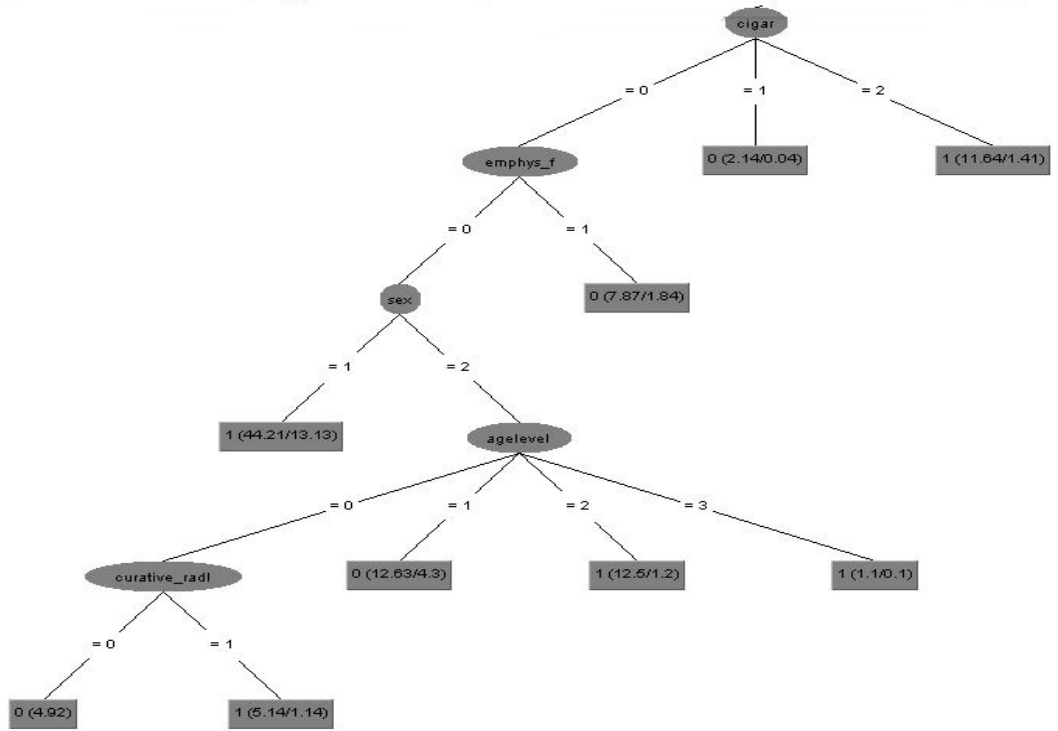


Figure 5.16 Sample section from decision tree-1

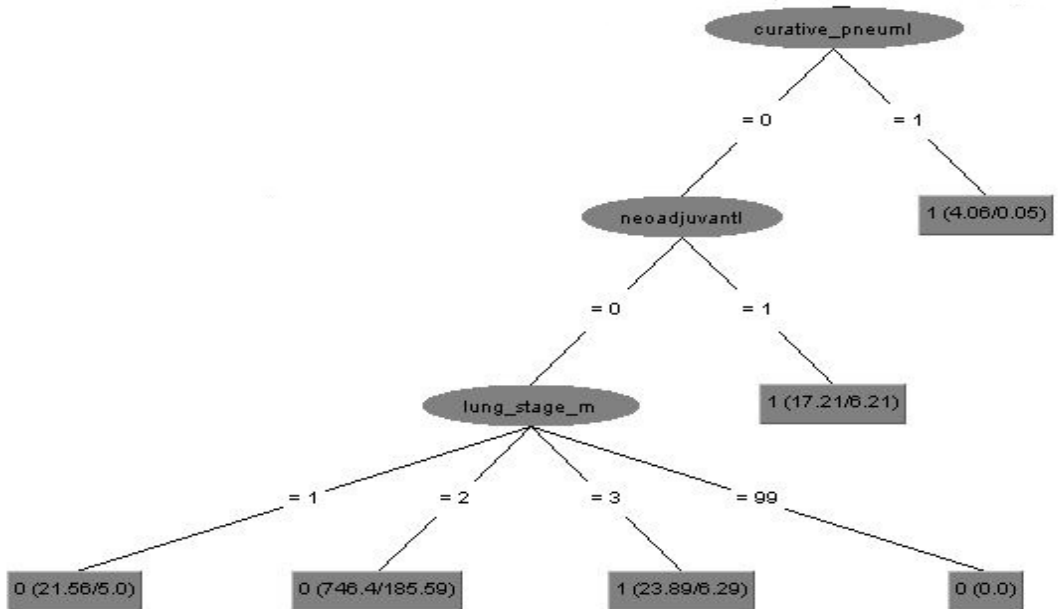


Figure 5.17 Sample section from decision tree-2

5.3.7 Evaluation of Classification Algorithms

Using 10-fold cross validation, the performance of the classification algorithms was evaluated. Figure 5.18 depicts that the C4.5 has higher classification accuracy than Naive Bayes and KNN.

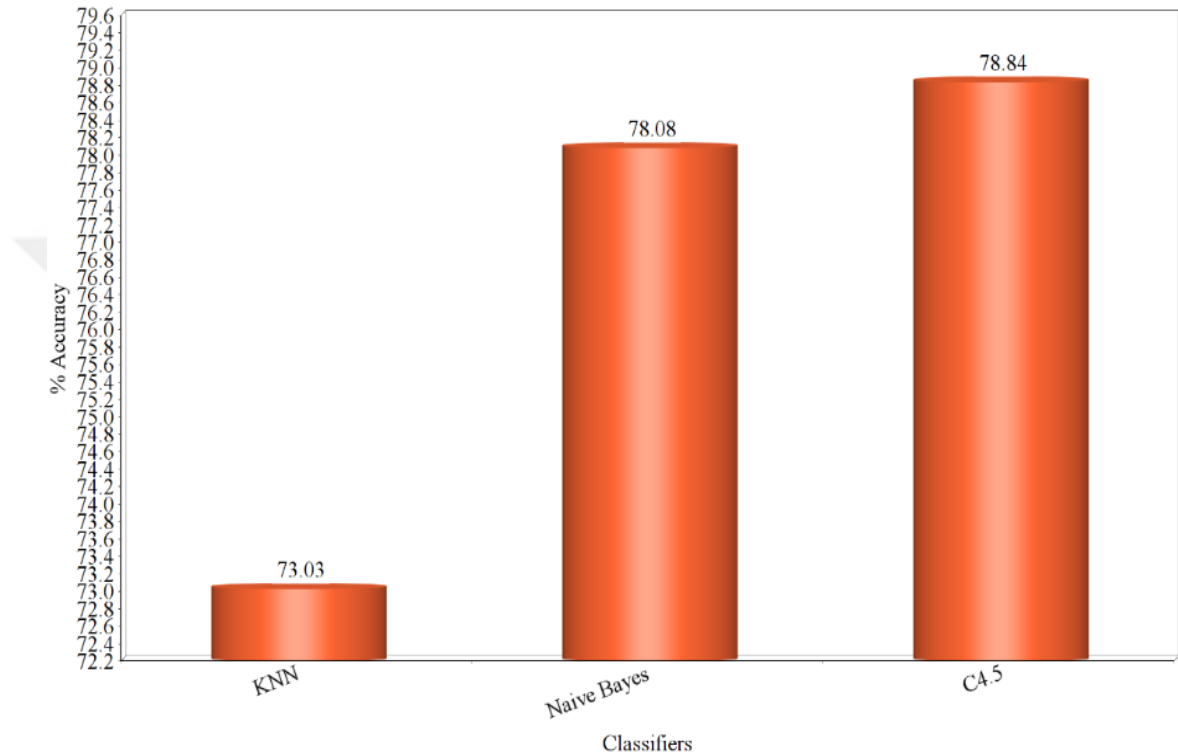


Figure 5.18 Comparison of classifiers

Since the accuracy rate is obtained from the C4.5 algorithm at most, it is decided to use this algorithm in the decision support system.

5.4 Development of Prototype Decision Support System

A decision support system has been developed to help doctors to make a prediction about future lung cancer patients' chances of survival using data from past lung cancer patients. The decision support system was developed in Visual Studio environment using C# programming language and MVC software architectural pattern.

MVC is a software architecture that separates the business logic from the application user interface and prevents the application's parts serving different purposes from interfering with each other. The MVC consists of three parts: model, view, controller as Figure 5.19 Model represents the data used in the application and is the part where the data is processed. View is simply the part of your application that your users see. Controller provides communication between model and view.

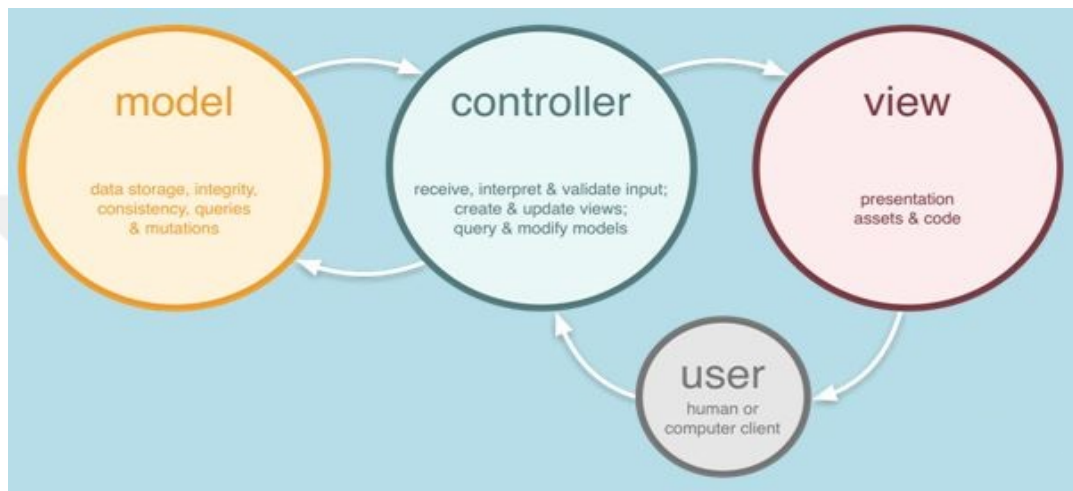


Figure 5.19 MVC architecture

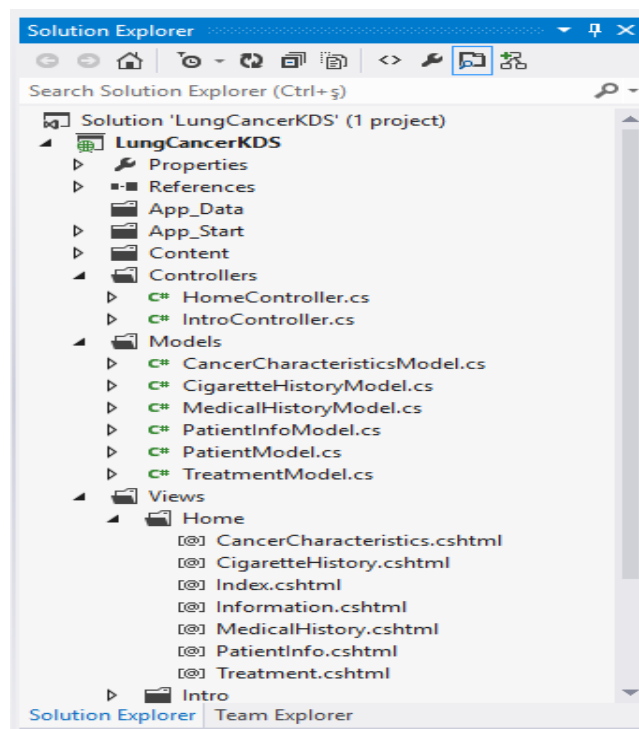


Figure 5.20 Lung cancer decision support system MVC architecture

The use case diagram of the Lung Cancer Decision Support System is shown in the Figure 5.21.

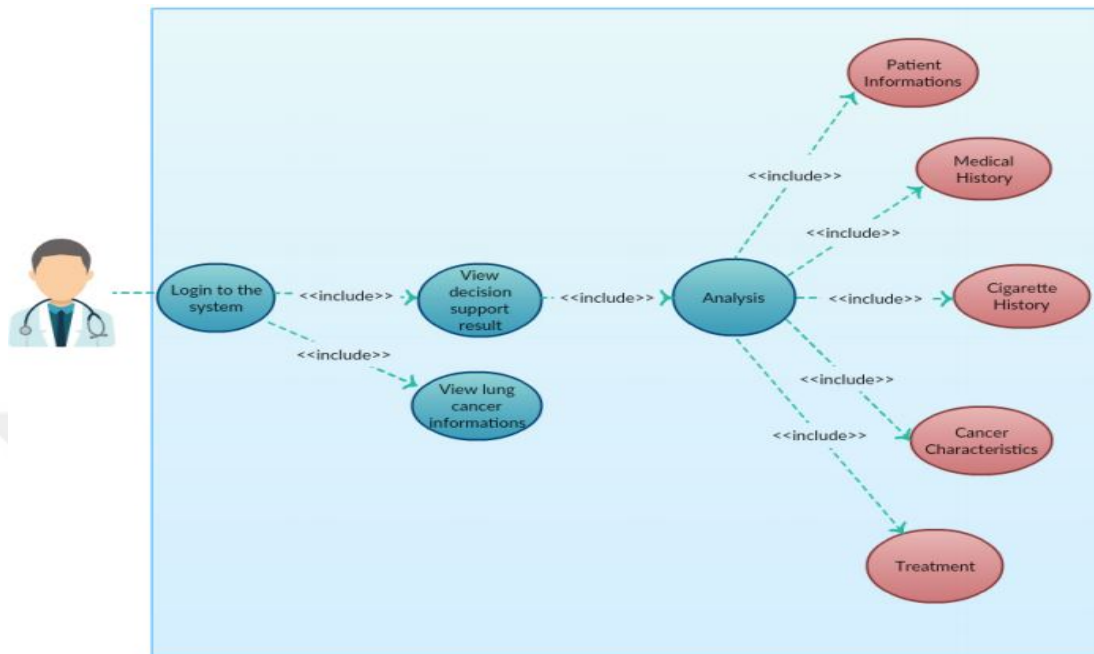


Figure 5.21 Lung cancer decision support system use case diagram

When the doctor opens the program, an entry screen is displayed. At the top of this page, there are five buttons that point to different parts Figure 5.22.

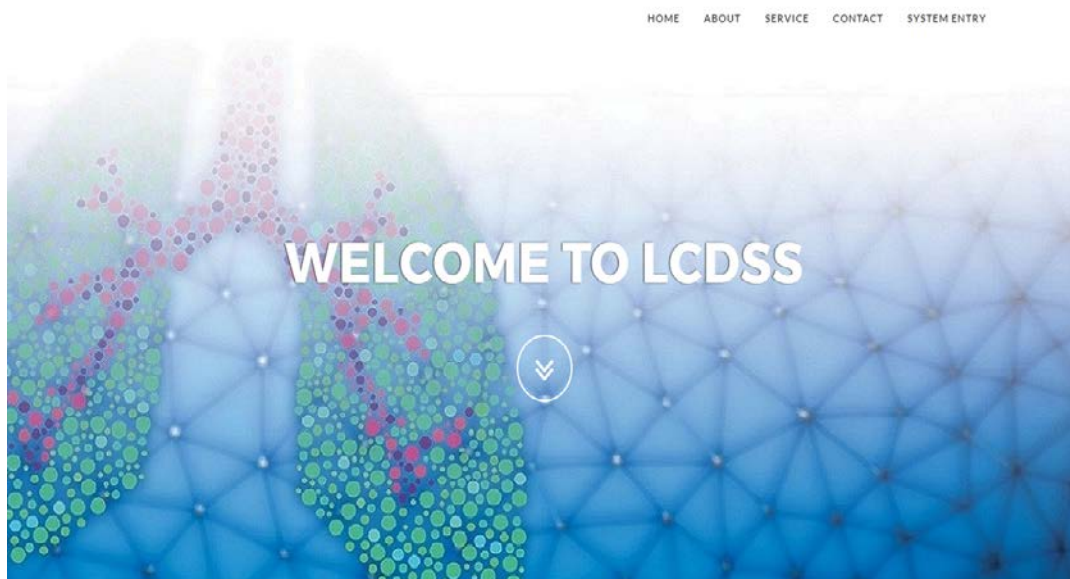


Figure 5.22 Lung cancer decision support system home page

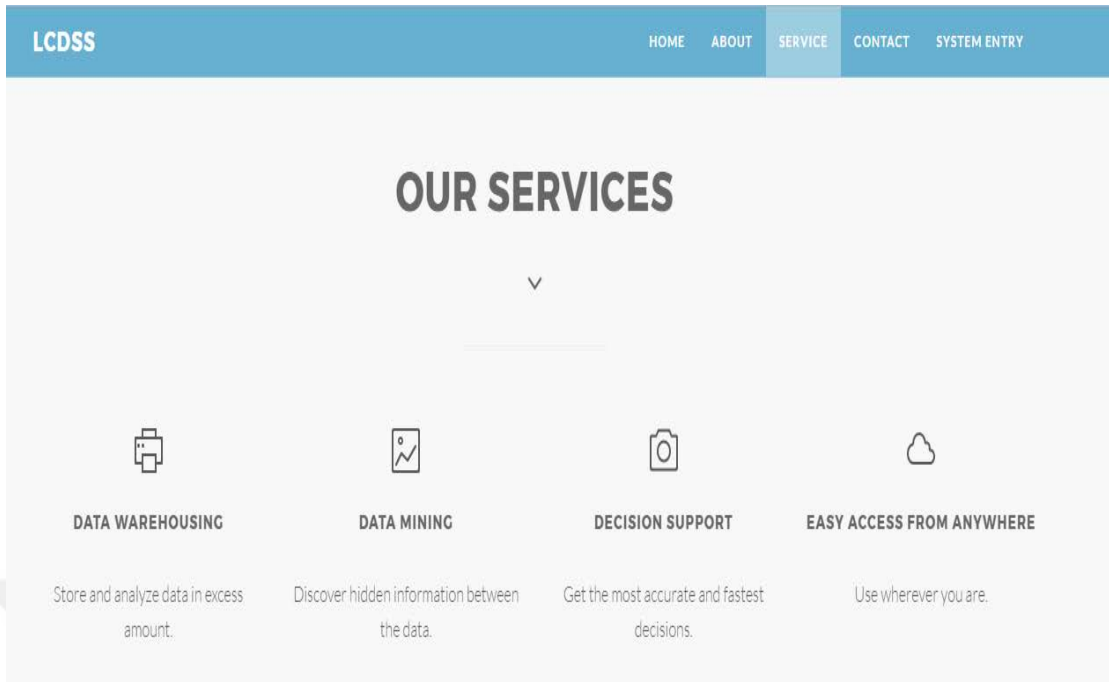


Figure 5.23 Lung cancer decision support system our service page

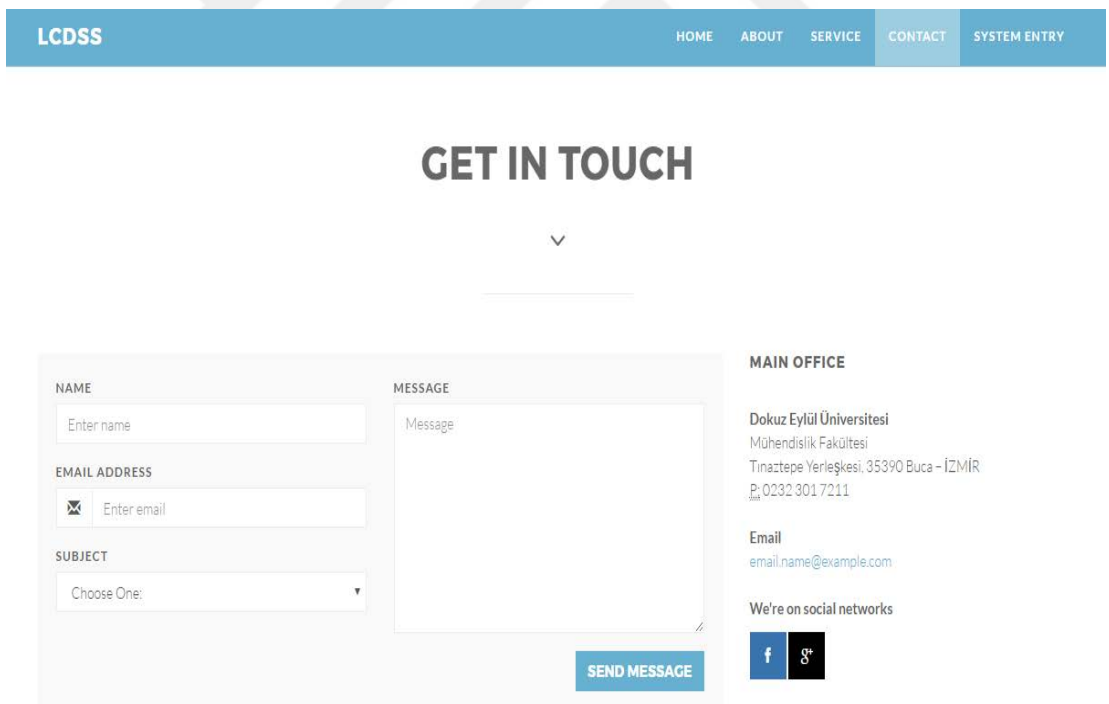


Figure 5.24 Lung cancer decision support system make contact page

The doctor will login to the system when he presses the "System Entry" button. Here a screen displays where the doctor will enter the patient's information as depicted in Figure 5.25.

Figure 5.25 Patient information page

After filling out the information on this page, when you click on the "Next" button, the patient's medical history screen is displayed in Figure 5.26.

Figure 5.26 Medical history page

After clicking "Next" button Cigarette History Page is displayed as in Figure 5.27.

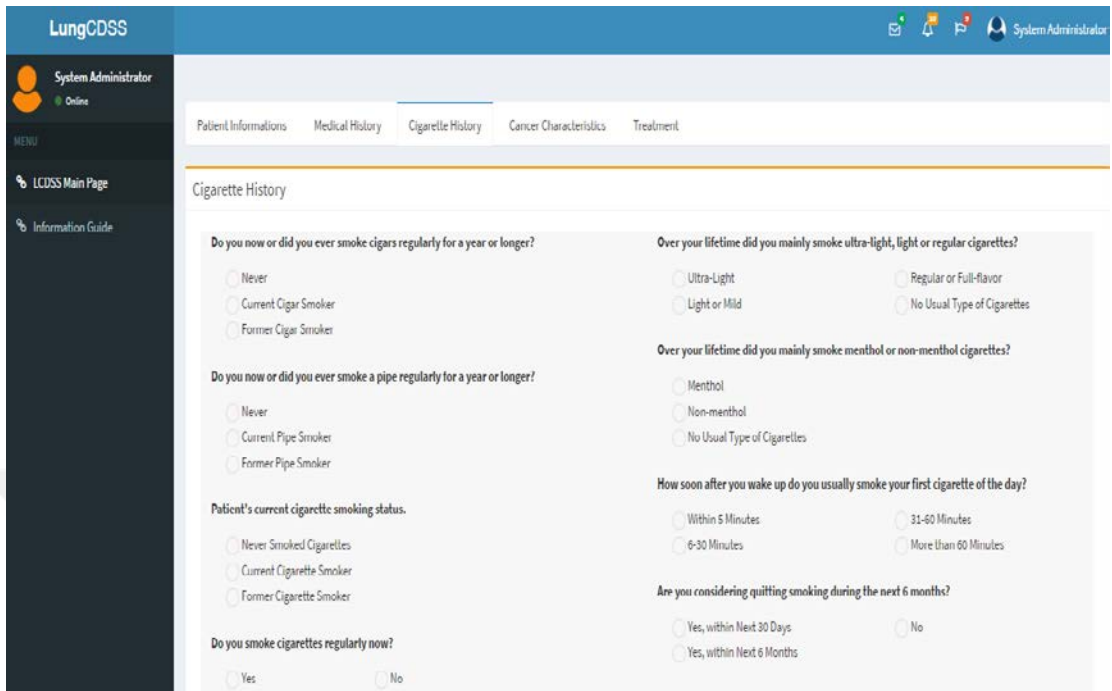


Figure 5.27 Cigarette history page

After clicking "Next" button Cancer Characteristics Page is displayed as Figure 5.28.

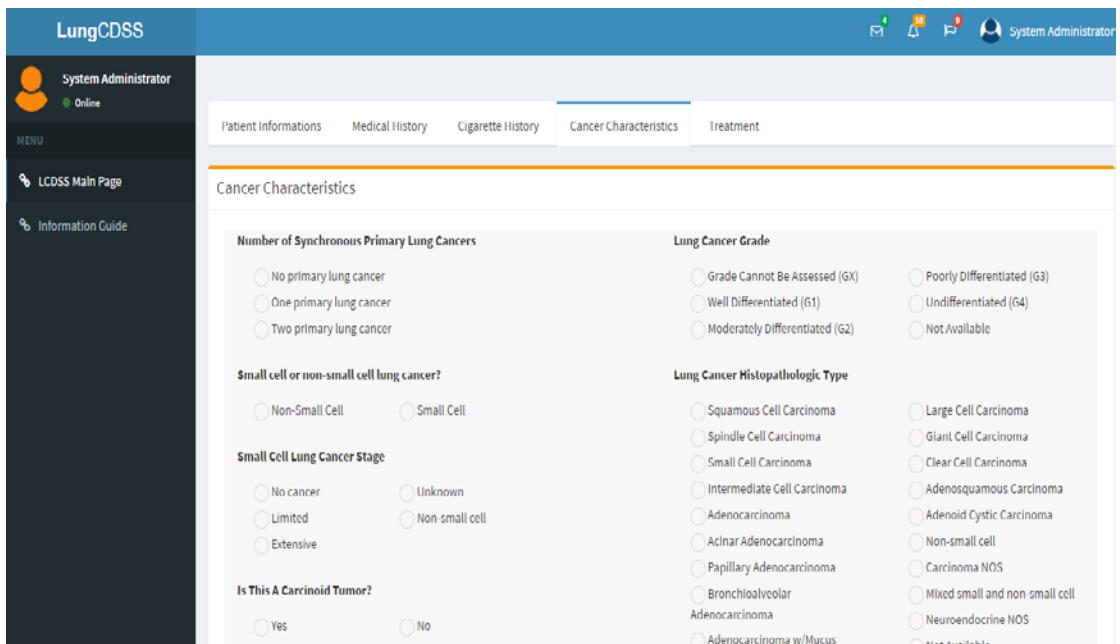


Figure 5.28 Cancer characteristics page

The doctor clicks on the "Analysis" button after entering the information of the treatments applied to the patient on the Treatment Page. When this button is clicked, the system converts the data entered by the doctor into the arff format. Then, according to the decision tree that we have created by the C4.5 algorithm in the data mining phase, the system runs this arff file. Weka Libraries were used to do this in Visual Studio.

Treatment

Had Pneumonectomy or Bilobectomy for Lung Cancer
 No Yes

Had Radiation Treatment for Lung Cancer
 No Yes

Had Chemotherapy for Lung Cancer
 No Yes

Had Wedge Resection, Segmental Resection, or Lobectomy for Lung Cancer
 No Yes

Had Neoadjuvant Treatment for Lung Cancer
 No Yes

Known Primary Treatment for Non-Small Cell Lung Cancer
 No Yes

Known Primary Treatment for Small Cell Lung Cancer
 No Yes

Analysis

Figure 5.29 Treatment page

After this step, the decision tree returns information about the patient's chances of survival according to entered data by doctor. When the "Show Result" button is clicked, the results are reflected on the screen as shown in the Figure 5.30 and Figure 5.31.

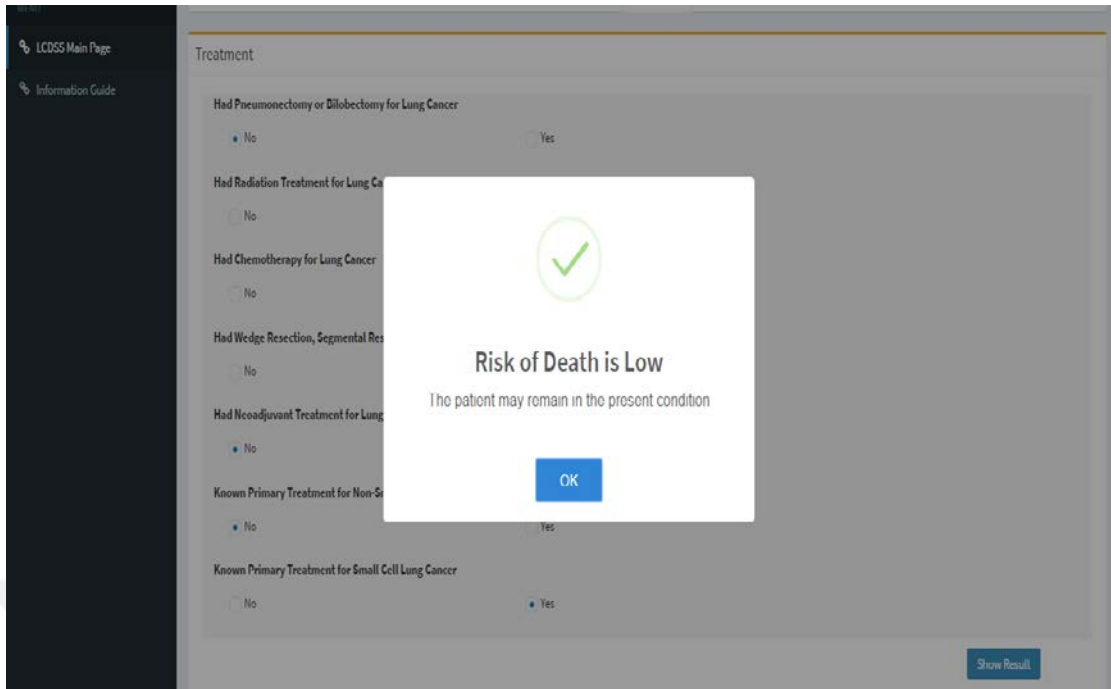


Figure 5.30 Risk of death is low

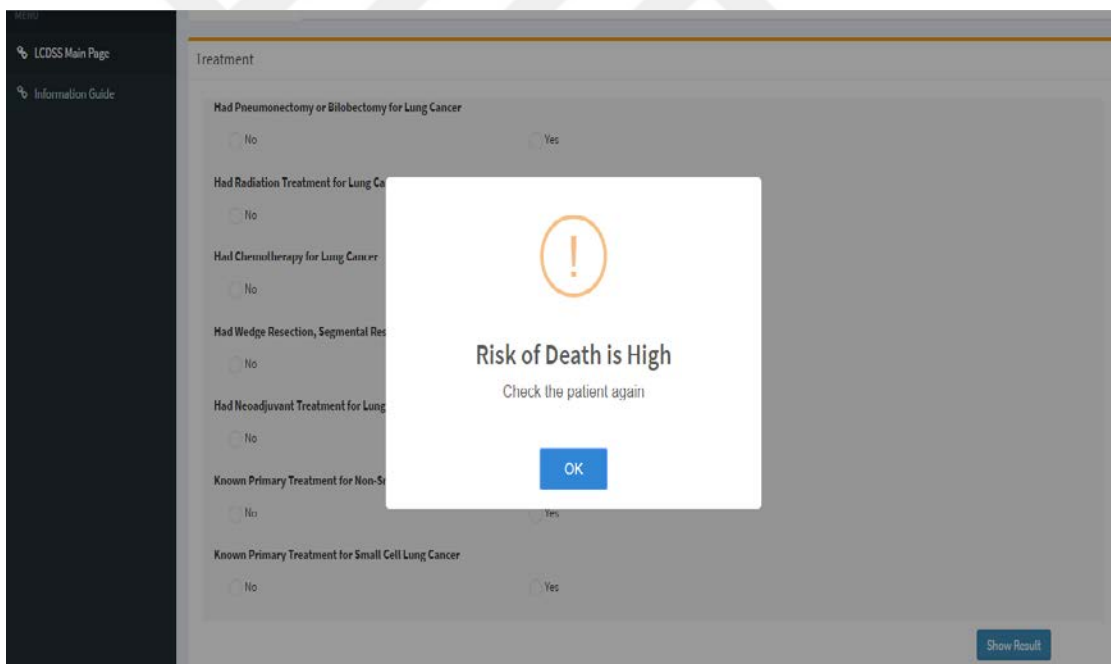



Figure 5.31 Risk of death is high

After this step, the doctor may shape the patient's treatment according to the results. On the left side, when you enter the Information Guide, you will see a section with information that the doctor can use. This section contains information on lung cancer, as well as conferences, publications and researches.


Let's beat cancer sooner

Find out everything you need about lung cancer.



Lung cancer is the second most common cancer in both men and women (not counting skin cancer), and is by far the leading cause of cancer death among both men and women. Each year, more people die of lung cancer than of colon, breast, and prostate cancers combined.


Most lung cancers could be prevented, because they are related to smoking (or secondhand smoke), or less often to exposure to radon or other environmental factors. But some lung cancers occur in people without any known risk factors for the disease. It is not yet clear if these cancers can be prevented.



Who is at Risk?

Some risk factors increase the risk of developing lung cancer. These are smoking tobacco, exposure to radon gas, chemicals and workplace risks, air pollution, previous lung disease, family history of lung cancer, previous radiotherapy treatment. But having any of these risk factors doesn't mean that you will definitely develop cancer.

[Find out more](#)



Signs of Lung Cancer

A cough that won't go away, a change in a cough you have had for a long time, being short of breath, coughing up blood, an ache or pain in the chest or shoulder, loss of appetite, losing weight, feeling very tired, persistent chest infections.

[Find out more](#)

Conferences

- 18th World Conference On Lung Cancer (Yokohama, Japan)
- ELCC 2017 European Lung Cancer Conference (Geneva, Switzerland)
- The Lung Cancer Centre of Excellence Conference 2017 (Manchester, UK)

Important Links

- <https://www.cancer.gov>
- <https://www.cdc.gov>
- <https://www.cancer.org>
- <http://kanser.gov.tr>
- <https://www.cancerresearchuk.org>

Researches

- Combining Targeted Therapies for Newly Diagnosed Metastatic Lung Cancer
- More Immunotherapy Options Approved for Lung Cancer
- Crizotinib Approval Expanded for Advanced Lung Cancer
- Lung Cancer Precision Medicine Trials: Adapting to Progress
- Using Gene Expression to Diagnose Lung Cancer More Accurately

Figure 5.32 Information guide page

CHAPTER SIX

CONCLUSION AND FUTURE WORK

6.1 Conclusion

There is a large amount of data awaiting evaluation in the health field. Healthcare institutions need to be able to use these data more consciously during decision-making, to develop more effective treatment methods, and to reduce costs by ensuring better management of resources.

The first step in achieving these goals is to keep data, coming from different systems and purifying from repetition, errors and uncertainties through the selection, preprocessing, conversion processes, in the data warehouse. The integrated data warehouse is used for analysis, querying and reporting. However, data warehousing alone can not reveal confidential information from the data.

Data mining algorithms to be applied to this data will be able to extract hidden, valuable, usable informations. These informations are used for the purposes such as choosing appropriate treatment method, determining drug interactions, classifying patient data according to certain factors, predicting risk factors related to diseases.

The use of data warehousing and data mining techniques for decision support is a new direction in the healthcare field. The developed decision support systems enable users to get the data quickly and easily, help them to make decisions on time and improve productivity and quality of the decisions.

In this study, a data warehouse for lung cancer data from the American National Cancer Institute (NCI) was established. It is ensured that the data warehouse conforms to the subject-oriented, integrated, time-variant and non-volatile conditions of the data warehouse approach presented by Inmon (1996). Data were integrated through preprocessing, transformation and selection operations. When designing the

data warehouse, the star schema model, multidimensional data modeling approach, was used. This model succeeded in responding to complex queries and the results were presented.

Naive Bayes, K-Nearest Neighbors, the C4.5 classification algorithms were then applied to the data in the data warehouse. By comparing the results of the algorithms applied, it is seen that the C4.5 algorithm is more successful than the others by 78.84%. The classification process is based upon the survival of the lung cancer patient. The decision tree produced by C4.5 has provided various rules regarding the likelihood of survival of lung cancer patients.

At the next stage of the study using a decision tree model obtained, web based prototype system was developed to provide support to physicians in decision making stages. This study showed that it could be possible to provide a better decision support environment for cancer disease by taking advantage of data warehouse and data mining methods.

6.2 Future Work

We intend to improve the decision support system developed by using data warehouse and data mining by adding the data from other cancer types in the future. We will use star schema with dimensional model for each type of cancer and combine all cancer types into a galaxy schema; therefore many fact table and some common dimensional tables will be shared. More various algorithms can be applied to the data mining phase. Cancer is a complex disease that changes rapidly with time, and many factors should not be ignored during diagnosis, follow-up and treatment. Such a system would provide great convenience to doctors who are dealing with cancer disease.

REFERENCES

- Arous E., McDade T., Smith J., Ng S., Sullivan M., & Zottola R. (2014). Electronic medical record: research tool for pancreatic cancer. *Journal Of Surgical Research*, 187, 466-470.
- Barber, D. (2010). *Bayesian reasoning and machine learning*, Retrieved April 10, 2017, from <http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/090310.pdf>.
- Bellaachia A., & Guven E. (2010). Predicting breast cancer survivability using data mining techniques. *2nd International Conference on Software Technology and Engineering*, V2-227-V2-231.
- Boon, M.E., & Kok, L.P. (2001). Using artificial neural networks to screen cervical smears: How new technology enhances health care. *Clinical Applications Of Artificial Neural Networks*, 81–89.
- CDAS (n.d.). Retrieved June 10, 2017, <https://biometry.nci.nih.gov/cdas/>
- Chiang I., Shieh M., Hsu J., & Wong J. (2005). Building a medical decision support system for colon polyp screening by using fuzzy classification trees. *Applied Intelligence*, 22(1), 61-75.
- Choi I., Park S., Park B., Chung B., Kim C., & Lee H. (2013). Development of prostate cancer research database with the clinical data warehouse technology for direct linkage with electronic medical record system. *Prostate International*, 1(2), 59-64.
- Danacı M., Çelik M., & Akkaya E. (2010). Veri madenciliği yöntemleri kullanılarak meme kanseri hücrelerinin tahmin ve teşhisi. *Akıllı Sistemlerde Yenilikler ve Uygulamaları Sempozyumu*.

Decision Tree, (n.d.). Retrieved May 21, 2017, from http://www.saedsayad.com/decision_tree.htm.

Devlin P.B.B.A., & Murphy P. T. (1998). An architecture for a business and information system. *IBM Systems Journal*, 27(1), 60-80.

Dimitoglou G., Adams J.A., & Jim C.M. (2012). Comparison of the C4.5 and a Naive Bayes classifier for the prediction of lung cancer survivability, Retrieved May 3, 2017, from <https://arxiv.org/ftp/arxiv/papers/1206/1206.1121.pdf>.

Fayyad U., Shapiro G., & Smyth P. (1996). The knowledge discovery and data mining process of extracting useful knowledge form volumes of data. *Communications of the ACM*, 39(11), 27-34.

Fernandes A., Alves P., Jarman I., Etchells T., Fonseca J., & Lisboa P. (2010). A clinical decision support system for breast cancer patients. *IFIP Advances in Information and Communication Technology*, 314, 122-129.

Frawley J.W., Shapiro G., Christopher J., & Matheus C. (1992). Knowledge discovery in databases: An overview. *AI Magazine*, 13(3).

Forgionne G., Gangopadhyay A., & Adya M. (2000). A decision technology system to advance the diagnosis and treatment of breast cancer. *Managing Healthcare Information Systems With Web-Enabled Technologies*, 141-150.

Garani G., & Helmer S. (2012) Integrating star and snowflake schemas in data warehouses. *International Journal of Data Warehousing and Mining (IJDWM)*, 8(4), 22-40

Globocan (n.d.). Retrieved May 10, 2017, from <http://globocan.iarc.fr/>

- Goletsis Y., Exarchos T., Giannakeas N., Tsipouras M., & Fotiadis D. (2011). Retrieved May 5, 2017, from https://www.researchgate.net/publication/228360777_Integration_of_clinical_and_genomic_data_for_decision_support_in_cancer.
- Han J., & Kamber M. (2000). Retrieved April 1, 2017, from http://ccs1.hnue.edu.vn/hungtd/DM2012/DataMining_BOOK.pdf.
- Hu H., Brzeski H., Hutchins J., Ramaraj M., Qu L., & Xiong R. (2004). Biomedical informatics: development of a comprehensive data warehouse for clinical and genomic breast cancer research. *Pharmacogenomics*, 5(7), 933–94.
- Inmon W.H. (1996). *Building the Data Warehouse* (2nd ed.). U.S.: Wiley Computer Publishing
- Johnson K.B., & Feldman M.J. (1995). Decision-support systems. *Archives of Pediatrics & Adolescent Medicine*, 149(12), 1371-1380.
- Kallmeyer V., & Venkat K. (2002). Beyond e-health: health and information technology converge, *Siliconindia*, 6(4), 42.
- Kanser (n.d.). Retrieved April 10, 2017, from <http://kanser.gov.tr/>
- Kimball R. & Ross M. (1996). *The data warehouse toolkit: The definitive guide to dimensional modeling* (2nd ed.). U.S.: Wiley Computer Publishing.
- Kimball R., & Casertam J. (2004). *The data warehouse etl toolkit: Practical techniques for extracting, cleaning, conforming, and delivering data* (2nd ed.). U.S.: Wiley Computer Publishing.

- Krishnaiah V., Narsimha G., & Chandra S. (2013). Diagnosis of lung cancer prediction system using data mining classification techniques. *International Journal of Computer Science and Information Technologies*, 4(1) , 39 - 45.
- Koç E., Şengül A.Y., & Özkaya U.A. (2012). Klinik karar destek sistemlerinin sağlık hizmetleri verimliliğine etkileri. *6.Sağlık ve Hastane İdaresi Kongresi*.
- Malinowski E., & Zimányi E. (2008). *Advanced data warehouse design: From conventional to spatial and temporal applications* (1st ed.). Springer.
- MVC (n.d.). Retrieved June 10, 2017, from https://en.wikipedia.org/wiki/ASP.NET_MVC.
- NCI (n.d.), Retrieved January 10, 2017 from <https://www.cancer.gov/>
- Özata M., & Aslan Ş. (2014). Klinik karar destek sistemleri ve örnek uygulamalar. *Afyon Kocatepe Üniversitesi Tıp Dergisi*, 5(1), 11-17.
- Peddabachigari S., Abraham A., Grosan G., & Thomas, J. (2007). Modeling intrusion detection system using hybrid intelligent systems. *Journal Of Network And Computer Applications*, 30(1), 114-132.
- Perreault L., & Metzger J.A. (1993). Pragmatic framework for understanding clinical decision support. *Journal of Healthcare Information Management*, 13(2), 5-21.
- Pedersen T.B., & Jensen C.S. (1998). Research issues in clinical data warehousing. *Tenth International Conference on Scientific and Statistical Database Management*, 43-52.
- Quinlan R. (1993). *C4.5: Programs for machine learning* (1st ed.). CA: Morgan Kaufmann Publishers.

Ramachandran P., Girija N., & Bhuvanewari T. (2014). Early detection and prevention of cancer using data mining techniques. *International Journal of Computer Applications*, 97(1), 48-53.

SAS *Studio* (n.d.). Retrieved June 10, 2017, from <http://support.sas.com/software/products/sas-studio>.

SAS *University Edition* (n.d.). Retrieved June 10, 2017, from <http://support.sas.com/software/products/university-edition>.

Sheta O., & Eldeen A. (2013). Evaluating a healthcare data warehouse for cancer diseases. *International Journal of Computer Science and Information Technology & Security*, 3(3), 237-241.

Shortliffe E.H. (1987). Computer programs to support clinical decision making. *Journal of the American Medical Association*, 258, 61-66.

Sintchenko V., & Garsden H. (2002). Retrieved May 2, 2017 from https://www.researchgate.net/publication/242184454_CLINICAL_DECISION_SUPPORT_NEW_APPROACHES_TO_USABILITY_STUDY.

Sql Server (n.d.). Retrieved June 10, 2017, from http://en.wikipedia.org/wiki/SQL_Server_Management_Studio.

Stephen R. (1998). Building the data warehouse. *Communications of the ACM*, 41(9), 52-60.

Stolba N., & Tjoa A.M. (2007). The relevance of data warehousing and data mining in the field of evidence-based medicine to support healthcare decision making. *World Academy Of Science, Engineering And Technology*, 1(11), 659-664.

- Stolba N., Banek M., & Tjoa A.M. (2006). The security issue of federated data warehouses in the area of evidence-based medicine. *First International Conference on Availability, Reliability and Security*, 329-339.
- Subbarao G., Sameeruddin K.S., & Kumar V. (2016). Case study on data mining application in health care monitoring systems. *International Journal of Engineering And Science*, 6(5), 79-82.
- Wah T., & Sim O. (2009). Development of a data warehouse for lymphoma cancer diagnosis and treatment decision support. *Transactions On Information Science And Applications*, 3(6), 530-543.
- Weka* (n.d.). Retrieved June 10, 2017, <http://www.cs.waikato.ac.nz/ml/weka>.
- Visual Studio 2010* (n.d.). Retrieved June 10, 2017, from http://en.wikipedia.org/wiki/Microsoft_Visual_Studio.
- Zubi Z., & Saad R. (2014). Improves treatment programs of lung cancer using data mining techniques. *Journal of Software Engineering and Applications*, 7, 69-77.

APPENDICES

APPENDIX 1: Dataset Attribute Explanations

Attribute	Description	Text Format
num_canc1	Number of Synchronous Primary Lung Cancers	0="No primary lung cancer", 1="One primary lung cancer", 2="Two primary lung cancers"
fin_smcl	Small Cell Lung Cancer Stage	0="Nocancer", 1="Limited", 2="Extensive", 9="Unknown", 10="Non-small cell"
lung_stage	Lung Cancer Stage	1="Stage I", 2="Stage IA", 3="Stage IB", 4="Stage II", 5="Stage IIA", 6="Stage IIB", 7="Stage IIIA", 8="Stage IIIB", 9="Stage IV", 93="Small Cell", 99="Not Available"
lung_stage_m	M Stage Component (Distant Metastases)	1="MX", 2="M0", 3="M1", 99="Not Available"
lung_stage_n	N Stage Component (Nodal Involvement)	1="NX", 2="N0", 3="N1", 4="N2", 5="N3", 99="Not Available"
lung_stage_t	T Stage Component (Primary Tumor)	1="TX", 4="T1", 5="T2", 6="T3", 7="T4", 99="Not Available"
lung_clinstage	Lung Stage Based On Clinical TNM	2="Stage IA", 3="Stage IB", 5="Stage IIA", 6="Stage IIB", 7="Stage IIIA", 8="Stage IIIB", 9="Stage IV", 11="Occult Carcinoma", 99="Not Available"

lung_clinstage_m	Clinical M (Distant Metastases)	1="MX", 2="M0",3="M1", 99="Not Available"
lung_clinstage_n	Clinical N (Nodal Involvement)	1="NX", 2="N0",3="N1",4="N2",5="N3", 99="Not Available"
lung_clinstage_t	Clinical T (Primary Tumor)	1="TX", 4="T1",5="T2",6="T3",7="T4", 99="Not Available"
lung_pathstage	Lung Stage Based On Pathologic TNM	2="Stage IA", 3="Stage IB", 5="Stage IIA", 6="Stage IIB", 7="Stage IIIA", 8="Stage IIIB", 9="Stage IV", 11="Occult Carcinoma", 99="Not Available"
lung_pathstage_m	Pathologic M (Distant Metastases)	1="MX", 2="M0", 3="M1", 99="Not Available"
lung_pathstage_n	Pathologic N (Nodal Involvement)	1="NX", 2="N0", 3="N1", 4="N2",5="N3", 99="Not Available"
lung_pathstage_t	Pathologic T (Primary Tumor)	1="TX", 4="T1", 5="T2", 6="T3", 7="T4", 99="Not Available"
lung_grade	Lung Cancer Grade	1="Grade Cannot Be Assessed (GX)",2="Well Differentiated (G1)", 3="Moderately Differentiated (G2)", 4="Poorly Differentiated (G3)", 5="Undifferentiated (G4)", 99="Not Available"
lung_histtype	Lung Cancer Histopathologic Type	2="Squamous Cell Carcinoma", 3="Spindle Cell Carcinoma", 4="Small Cell Carcinoma", 5="Intermediate Cell Carcinoma", 7="Adenocarcinoma", 8="Acinar Adenocarcinoma",

		9="Papillary Adenocarcinoma", 10="Bronchioalveolar Adenocarcinoma", 11="Adenocarcinoma w/Mucus Formation", 12="Large Cell Carcinoma", 13="Giant Cell Carcinoma", 14="Clear Cell Carcinoma", 15="Adenosquamous Carcinoma", 18="Adenoid Cystic Carcinoma", 30="Non-small cell (recoded)", 31="Carcinoma NOS (recoded)", 32="Mixed small and non-small cell (recoded)", 33="Neuroendocrine NOS (recoded)", 99="Not Available"
lung_cancer_type	Lung Cancer Type	1="Non-Small Cell Lung Cancer", 2="Small Cell Lung Cancer"
lung_is_carcinoid	Is This A Carcinoid Tumor?	0="No", 1="Yes"
curative_pneuml	Had Pneumonectomy or Bilobectomy for Lung Cancer	0="No", 1="Yes"
curative_wsll	Had Wedge Resection, Segmental Resection, or Lobectomy for Lung Cancer	0="No", 1="Yes"
curative_chemol	Had Chemotherapy for Lung Cancer	0="No", 1="Yes"
curative_radl	Had Radiation Treatment for Lung Cancer	0="No", 1="Yes"

neoadjuvantl	Had Neoadjuvant Treatment for Lung Cancer	0="No", 1="Yes"
primary_trtl_NSC	Known Primary Treatment for Non-Small Cell Lung Cancer	1="Resection of one lobe or less without chemotherapy", 2="Resection of two lobes or more without chemotherapy", 3="Resection and chemotherapy", 4="Radiation only", 5="Chemotherapy without resection", 6="Has treatment form, no known treatment with curative intent", 10="MDF for the treatment form", 12="Small cell lung cancer"
primary_trtl_small	Known Primary Treatment for Small Cell Lung Cancer	1="Chemotherapy, resection, and radiation", 2="Chemotherapy and resection without radiation", 3="Chemotherapy and radiation without resection", 4="Chemotherapy only", 5="Has treatment form, no known treatment with curative intent", 10="MDF for the treatment form", 12="Non-small cell lung cancer"
pipe	Ever Smoked a Pipe?	0="Never", 1="Current Pipe Smoker", 2="Former Pipe Smoker"
cigar	Ever Smoked Cigars?	0="Never", 1="Current Cigar Smoker", 2="Former Cigar Smoker"
hyperten_f	Hypertension	0="No", 1="Yes"
hearta_f	Heart Attack	0="No", 1="Yes"
stroke_f	Stroke	0="No", 1="Yes"
emphys_f	Emphysema	0="No", 1="Yes"
bronchit_f	Bronchitis	0="No", 1="Yes"

diabetes_f	Diabetes	0="No", 1="Yes"
smoked_f	Ever Smoke Regularly >= 6 Months?	0="No", 1="Yes"
rsmoker_f	Smoke Regularly Now?	0="No", 1="Yes"
cigpd_f	# of Cigarettes Smoked Per Day	0="0", 1="1-10", 2="11-20", 3="21-30", 4="31-40" 5="41-60", 6="61-80", 7="81+"
filtered_f	Usually Filtered or Non-Filtered?	1="Filter", 2="Non-Filter", 3="About Equal"
cig_stat	Cigarette Smoking Status	0="Never Smoked Cigarettes", 1="Current Cigarette Smoker", 2="Former Cigarette Smoker"
bmi_curc	Current BMI	1="0-18.5", 2="18.5-25", 3="25-30", 4="30+"
fh_cancer	Has Family History of Any Cancer?	0="No", 1="Yes"
num_confirmed	Number of Confirmed Cancers	Numeric
confirmed_icdgrd(1-3)	[X]th Confirmed Cancer Grade	1="Well Differentiated; Grade I", 2="Moderately Differentiated; Grade II", 3="Poorly Differentiated; Grade III", 4="Undifferentiated; Grade IV", 5="T Cell; T Precursor", 6="B Cell; Pre B; B Precursor", 7="Null Cell; Non T, Non B", 9="Not Determined/Stated/or Applicable"
confirmed_icdbeh(1-3)	[X]th Confirmed Cancer Behavior	1="Uncertain, Borderline, or LMP", 2="In Situ", 3="Malignant, Primary Site", 9="Malignant, Uncertain Primary"

		or Metastatic"
confirmed_can_type(1-3)	[X]th Cancer Type	1="Abdomen", 2="Adrenal Glands", 3="Bladder", 4="Bone", 5="Brain", 6="Breast", 7="Cervix", 9="Diaphragm and Connective Tissue of Thorax", 10="Digestive System, Other and Unspecified", 11="Esophagus", 12="Falopian Tubes", 13="Female Genital, Other and Unspecified", 14="Hodgkins Disease", 15="Intestine", 16="Ill-Defined Sites", 17="Kidney and Renal Pelvis", 18="Larynx", 19="Leukemia", 20="Liver", 21="Lung", 22="Lymph Nodes", 23="Male Genital, Other and Unspecified", 24="Melanoma", 25="Lip, Oral Cavity, Pharynx", 26="Non-Hodgkin's Lymphoma", 28="Ovary", 29="Pancreas", 30="Peritoneum", 31="Prostate, 33="Skin", 34="Stomach", 35="Testis", 37="Thyroid", 38="Uterus", 39="Vagina", 40="Anus and Anal Canal", 41="Connective, Subcutaneous, and Other Soft Tissues and Peripheral Nervous System", 42="Endocrine Glands", 43="Endometrium", 44="Eye", 45="Gallbladder", 46="Heart, Mediastinum, and Pleura", 47="Hematopoietic and

		Reticulendothelial Systems", 48="Kaposi's Sarcoma", 49="Meninges", 50="Multiple Myeloma", 51="Nasopharynx, Nasal Cavity, Middle Ear, and Sinuses", 52="Pelvis", 53="Penis", 56="Spinal Cord and Cranial Nerves", 58="Thymus", 59="Trachea" 60="Ureter, Urinary Organs", 80="Colorectum",999="Not Ascertained"
sex	Sex	1="Male", 2="Female"
agelevel	Age Level	0="<= 59", 1="60-64", 2="65-69", 3=">= 70"
sqx_fh_lung	Family History of Lung Cancer	0="No", 1="Yes"
sqx_cholesterol	High Cholesterol	0="No", 1="Yes"
sqxraw_smk100	Smoked 100 Cigarettes	0="No", 1="Yes"
sqx_smk_lgt	Cigarette Type	1="Ultra-Light", 2="Light or Mild", 3="Regular or Full-flavor", 4="No Usual Type of Cigarettes"
sqx_smk_menth ol	Menthol or Non-Menthol	1="Menthol", 2="Non-menthol", 3="No Usual Type of Cigarettes"
sqx_amt_smk	Cigarettes Per Day	1="1-5 Each Day", 2="6 - Under 1 Pack Each Day", 3="About 1 Pack Each Day", 4="About 1 1/2 Packs Each Day", 5="About 2 Packs Each Day", 6="More than 2 Packs Each Day"
sqx_smk_wake	Wake Up Smoke	1="Within 5 Minutes", 2="6-30 Minutes", 3="31-60 Minutes", 4="More

		than 60 Minutes"
sqx_smk_plan_quit	Quit Smoking 6 Months (SQX)	1="Yes, within Next 30 Days", 2="Yes, within Next 6 Months", 3="No"
sqx_smk_try_quit	Serious Attempt to Quit (SQX)	0="No", 1="Yes"
sqxraw_nicotine_gum	Nicotine Gum (SQX)	0="No", 1="Yes"
sqxraw_nicotine_patch	Nicotine Patch (SQX)	0="No", 1="Yes"
sqx_zyban	Prescription Pills (SQX)	0="No", 1="Yes"
sqx_smk_exp_child	Lived with Smoker Prior to Age 18 (SQX)	1="Yes, During Most of your childhood", 2="Yes, During Some of your childhood", 3="No, Not at All"
sqx_smk_exp_adult	Lived with Smoker After Age 18 (SQX)	1="Yes, During Most of your adult life", 2="Yes, During Some of your adult life", 3="No, Not at All"
sqx_smk_exp_work	Worked with Smoker as Adult (SQX)	1="Yes, During Most of your work experience", 2="Yes, During Some of your work experience", 3="No, Not at All"
class	Is Lung Cancer The Underlying Cause Of Death?	0="No", 1="Yes"