POST OPERATIVE PROGNOSTIC PREDICTION OF ESOPHAGEAL CANCER CASES USING BAYESIAN NETWORKS AND SUPPORT VECTOR MACHINES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

NEGIN BAGHERZADI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE
IN
MEDICAL INFORMATICS

MAY 2014

POST OPERATIVE PROGNOSTIC PREDICTION OF ESOPHAGEAL CANCER CASES USING BAYESIAN NETWORKS AND SUPPORT VECTOR MACHINES

Submitted by **Negin Bagherzadi** in partial fulfillment of the requirements for the degree of **Master of Science in Medical Informatics Program, Middle East Technical University** by,

Prof. Dr. Nazife Baykal _____
Director, **Informatics Institute**

Asist.Prof. Dr. Yeşim Aydın Son _____
Head of Department, **Health Informatics, METU**

Asist.Prof. Dr Aybar Can Acar _____
Supervisor, **Health Informatics, METU**

Examining Committee Members:

Prof. Dr. Serdar Han _____
Medicine Department, Ufuk University

Asist.Prof. Dr Aybar Can Acar _____
Supervisor, Health Informatics, METU

Assoc. Prof. Dr. Tolga Can _____
Computer Engineering, METU

Assoc.Prof. Dr. Banu Günel _____
Information Systems, METU

Asist.Prof. Dr. Yeşim Aydın Son _____
Health Informatics, METU

Date: 06.05.2014

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: **Negin Bagherzadi**

Signature : _____

ABSTRACT

POST OPERATIVE PROGNOSTIC PREDICTION OF ESOPHAGEAL CANCER CASES USING BAYESIAN NETWORKS AND SUPPORT VECTOR MACHINES

Negin Bagherzadi
M.Sc, Medical Informatics Program
Supervisor: Assist. Prof. Dr. Aybar Can Acar

May 2014, 79 pages

The objective of this thesis is to develop and analyze the performances of a number of classifiers in prognosis classification based on a medical history data set. Generally, data mining uses algorithms originating in different disciplines such as artificial intelligence, statistics, optimization, database theory etc., to clarify available data. In this study, Support Vector Machine and Bayesian Network methods have been used. The data analyzed are clinical pathology records of patients with esophageal cancer who received an esophagectomy operation between 2003 and 2011. A large number of prognostic factors have been considered to classify the patients on prognosis. These factors found to be predictive were age, sex, dysphagia, odynophagia, lost weight, vomit, nausea, pathological N, pathological T, FEV1, tumor grade and tumor length. Classification trials for Support Vector Machine have shown %72.38 training accuracy with a generalization accuracy of %70.58, which was established by cross-validation. Support Vector Machine was used as the first method for data and SVM was found to achieve good accuracy on a data set of 119 patients when used in conjunction with PCA; SVM can be helpful for black-box analysis when data are irregularly distributed and produce accurate classifiers. Bayesian Network is the second method that is used in this study to solve missing data problem in the previous method. Bayesian Network can classify several different types of variables simultaneously. The quality of the results of the network that has been created for this study depends on the quality of the model. Performance of the model with Bayesian Network that is used in this study is 73.10% for 119 patients.

Keywords: Support Vector Machine, Bayesian Networks, Classification, Medical Data Analysis, Esophageal Cancer.

ÖZ

ESOFAGUS KANSER VAKALARININ BAYES AĞLARI VE DESTEK VEKTÖR MAKİNALARI KULLANILARAK AMELIYAT SONRASI PROGNOSTİK TAHMİNİ

NEGIN BAGHERZADI
Yüksek Lisans, Tıp Bilişimi
Tez Danışmanı: Assist. Prof. Dr. Aybar Can Acar

Mayıs 2014, 79 pages

Bu tezin amacı, tıbbi bir veri setini sınıflandırmak için bir dizi etkili sınıflandırma modelini geliştirmek ve performanslarını analiz etmektir. Genellikle, veri madenciliğinde, mevcut verileri açıklamak için, yapay zeka, istatistik, optimizasyon, veri tabanı teorisi gibi çeşitli alanlardan gelen algoritmaları kullanılır. Bu çalışmada, Destek Vektör Makinesi ve Bayes Ağ yöntemleri kullanılmaktadır. Analiz verileri, 2003 ve 2011 yılları arasında bir özefajektomi operasyonu geçirmiş özofagus kanserli hastaların klinik patoloji kayıtlarıdır. Hastaların prognozlarına göre sınıflandırılabilmeleri için prognostik olarak etkili olabilecek büyük grup faktör analiz edilmiştir. Bu faktörlerden tahminde etkili bulunanlar yaş, cinsiyet, disfaji, odinofaji, kilo kaybı, kusma, bulantı, patolojik N, patolojik T, FEV1, tümör derecesi, tümör boyu olmuştur. Yapılan sınıflandırma deneylerinde, Destek Vektör Makinesi %72.38 öğrenme doğruluğu ve çapraz-doğrulama ile %70.58 genelleme başarısı görülmüştür. Destek Vektör Makinesi ilk yöntem olarak kullanıldı ve PCA ile desteklendiğinde en iyi doğruluk oranı elde edildi; SVM düzensiz dağılan ve tam modellenemeyen verilerin analizi için doğru sınıflandırıcı üretebilir. Bayes Ağı önceki yöntemde eksik veri sorununu çözmek için, bu çalışmada kullanılan ikinci yöntemdir. Bayes Ağları aynı anda birkaç farklı değişken türde sınıflandırabilir. Bu çalışma için oluşturulmuş ağ sonuçlarının kalitesi modelin kalitesine bağlıdır. Bu çalışmada kullanılan Bayes ağı modelinin doğruluğu 119 hasta için %73.10 dir.

Keywords: Destek Vektör Makinesi, Bayesian Ağlari, sınıflandırma, Tıbbi Veri Analizi, Özofagus Kanseriz.

All my love to my family,

For finding me the light,

However they were far away.

ACKNOWLEDGEMENTS

I am sincerely thankful to my supervisor, Assist.Prof. Dr Aybar Can Acar his guidance, motivation, expertise, understanding, and patience, added considerably to my graduate experience. I appreciate his vast knowledge and skill in many areas

I would like to thank to Prof. Dr. Serdar Han for answering all my never ending questions and thanks to Assist.Prof. Dr. Yeşim Aydın Son, Assoc. Prof. Dr. Tolga Can, and Asist.Prof. Dr. Banu Günel for reviewing my work.

I would like to thank to Assist.Prof. Dr. Reza Hassanpour and Dr. Roya Choupani, for their kindness and support they have shown during the past two years it has taken me to finalize this thesis.

Special thanks go to my colleagues and my lovely friends for their support during my study.

I would like to especially thank to Iman Ashtiani Abdi, Mona Zolfaghari Borra, Yeghaneh Baghi, Parisa Sharif, Yousef RezaieTabar and Sara Razzaghi for being always there when I need courage, help and for their endless support, encouragement, and suggestions since the beginning of this research.

I would like to thank to Informatics Institute members Hakan Güler and Yaşar Sayın for their support.

I also would like to thank Gökçe Oğuz. This thesis would not have been possible unless her supports. I will never have words to thank her enough.

Special thanks to my family for their endless support, love, trust, patience and encouragement during this study. They devoted their life on me

Specially thanks to my twin sister Haleh and Laleh Bagherzadi for their endless affection. They never left me alone, even for a second.

TABLE OF CONTENTS

LIST OF FIGURES:

LIST OF TABLES:

CHAPTER 1

1. INTRODUCTION

1.1. Motivation

Data mining and Machine Learning Techniques

Nowadays, computer systems and programs are one of the developed fields in artificial intelligent that designed to help to experts where the human knowledge are invalid and it is not enough. There are some trends in medical decision system that shows the need to obtain formal and logic reasoning, for this purpose using intelligent data analysis systems are useful for extraction of statistical knowledge from patient's data that they are sorted in medical records [1].

Data mining is a process to extract the meaningless data from huge data sets. In last year's researchers and doctors try to solve many medical analyzing problems by data mining approaches, for this purpose they have to understand the structure of the main and real data and statistical view point of model.

There are techniques to create expert systems like decision theory [2] symbolic reasoning technology [3] and probabilistic belief networks [4] and machine learning. Clustering is another technique for intelligent data analysis.

We can classify machine learning methods into three branches such as:

Pattern Recognition methods have many important role in machine learning such as (K-nearest neighbors and Bayesian classifiers [5] and [6]) and Neural Networks [7] and decision trees [8].

We can use machine learning methods to improve medical decisions because these methods can be applied to most medical domains such as diagnostic and prognostic cases in oncology [9] esophagus cancer [10] and many other diseases. By learning the past experiences we can improve diagnosis and prognosis problems.

Prognostic models are important tools in medicine. Given a set of patient specific parameters, they predict the future incidence of a medical event or outcome for specific diseases (e.g. expected survival for cardiovascular diseases and cancer). The models are used for prediction purposes. Predictions help doctors and patients to make treatment choices. Prognostic models describe the relationship between predictor and outcome variables. The standard methodology to achieve an objective explanation of this relationship to build predictive models from a set

of observed patient data and outcomes. All patient data that are available are then taken into account for model development. Subsequently, variables that are found to have predictive value for the outcome are selected for inclusion to the model (feature selection).

Medical specialist need to be assisted in their decisions and there is a fact that the cost of medical care is increasing and medical errors are increasing too, so using computer assistance to support medical decision making will lead doctors to make fewer mistakes.

In this study the main question is from patient to doctors. They focus on asking questions such as “How long do you predict my life span?” or “What is my life expectancy after surgery?”

1.2. Goal

In order to optimize the doctor’s answers for these kinds of questions we use two methods “Support vector machine” and “Bayesian Networks” to help to doctors for their decisions and statistical analyzing.

Medical applications of Support Vector Machine and Bayesian Networks

Support Vector Machine is one of the important methods for Classification. SVM has successful applications in many complex problems in data mining. Support Vector Machine is a useful method in solving medical problems.

Bayesian networks are used in many medical problems as a classification. One reason is the popularity of Bayesian networks in medicine because the structure of networks is composed of directed acyclic graphs (DAGs) and suggest causal relationships among the input variables.

Bayesian networks have advantages such as analyzing the network is a potential benefit for the medical community and conditional probability distribution is another advantage of Bayesian Networks.

Bayesian network can works in various medical applications and allows for the users to evaluate with other data sets with existing hypotheses.

1.3. Outline of Thesis

This thesis consists of 5 chapters as follows:

In chapter 1, we briefly give a review of data mining and machine learning, the aim of this thesis study and the outline of the thesis.

In chapter 2, we describe the background of disease and statistical distribution of this disease around the world and data reduction is described for decreasing feature numbers that it used for our first method, our methods “Support Vector Machine” and “Bayesian Networks” is introduced that we applied for classification on esophagus cancer database. Another topic that we mentioned in this chapter is about software that assists us to classify data and analyze the results as a statistical view point.

In chapter 3, we present the dataset which is given from Ankara Numune Teaching and Research Hospital between 2003 and 2011 and the use of PCA and SVM and Bayesian Networks at this study described in detail in this chapter.

In chapter 4, the results are provided and discussed. We used Confusion Matrix and Receiver Operating Characteristic (ROC) for analyzing our first method performance to obtain best model of SVM for classification and for second method we used a lot of networks to find best relation between features.

In chapter 5, the results and discussions of this thesis work are summarized and we present the main conclusions.

CHAPTER 2

2. BACKGROUND AND RELATED WORKS

2.1. BACKGROUND OF DISEASE

In early twentieth century, scientists have been interested in esophageal carcinoma. Epidemiologic studies have shown the "esophageal cancer belt" which has generated a drastic concern in etiology and other risk factors [11]. By combining different therapy modalities and introducing new agents similar to "hyperthermia" and "biologic response modifier", approaches to treatment have taken long steps. Contrary to these efforts, the overall prospects of survival in esophageal carcinoma remain grim [12] and [13].

International variations in the occurrence of squamous cell carcinoma of the esophagus are perhaps greater than any other tumour. It does not spread uniformly across the world. As Figure 2.1 represents, in the United States and Britain 0.02% of people are suffering from the carcinoma of the Esophagus, this number increases to 0.16% in parts of South Africa and the Honan Province of China and reaches to 0.54% in the Guriev district of Kazakhstan. [14, 15]

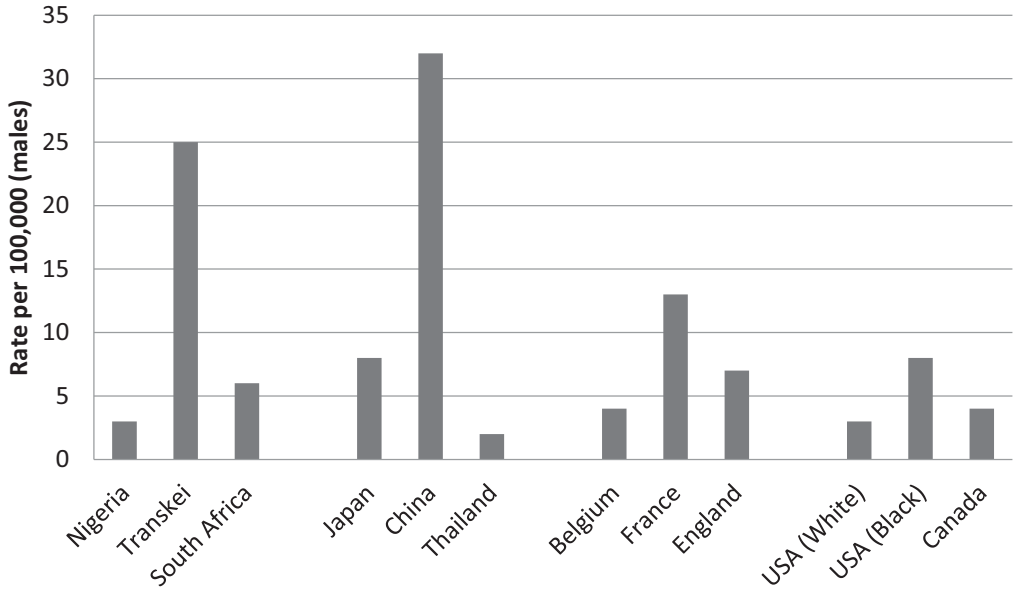

Figure 2.1: Geographic variation in distribution of esophageal cancer

The geographical variation in esophageal carcinoma is probably related to variation in life style. In western countries, two main accepted factors, which are linked with carcinoma of the Esophagus, are smoking and alcohol consumption. The risk for carcinoma in people who quit smoking is reduced. The amount smoked also seemed to correlate with risk. Regular drinkers, defined as those who drink beer, wine or sprits at least once a week all showed a significant dose-reponse relationship [16].

Other factors suspected as being important in esophageal carcinoma are diet and nutrition. Studies show existence of nitroso- compounds in local foodstuffs, zinc and molybdenum deficiency could increase the risk of esophageal carcinoma. Selenium concentration and high body-iron stores have also been reported as a risk factor in carcinoma. A recent study reported that unsaturated fatty acids influenced the development of chemically induced tumors in the gastrointestinal tract of experimental animal models [17].

Occupational factors can be related to esophageal carcinoma. Studies shows that long-term occupational (at least 19 years) exposure to metal dust (nickel, beryllium and chromium) has a serious dependency to carcinoma of esophagus. The influence of the genetic factors on esophageal carcinoma has been rarely reported. However, the risk of upper gastrointestinal carcinoma is increased in patients with the familial polyposis syndrome [18].

Regarding the age and sex as influential factors, none of reported cases seems to be age or sex linked. In Britain, the male—female ratio is 17:1, whereas in the high incidence areas of Iran the ratio is reversed! [19]

2.2. DIMENTIONALITY REDUCTION (PCA):

Principal Component Analysis was first introduced by Karl Pearson (1901). He found lines and planes which best fit a set of points in p-dimensional space that is the way to optimize geometric topics. Harold Hotelling (1933) reinvented it in different point view, for this purpose he published a paper on *Journal of Educational Psychology* that it focus on algebraic optimization. Hotelling choose components to maximize their successive contribution to the total variance [20].

PCA is a powerful tool for analyzing data. It compresses the data by reducing its dimensions and remove some unnecessary information. The goal of PCA is to find more important data from data set. It describe in the better way by creating new data set by calculating eigenvectors and eigenvalues from original matrix.

First step is to calculate Covariance matrix, since eigenvectors and eigenvalues extracted from square symmetric matrix. Covariance matrix is m×m symmetric matrix, the diagonal of this matrix is the Variance of vectors and other numbers are covariance. This matrix is used to measure two random variables from dataset that how much the dimensions is vary from mean. Covariance extracts some features that their correlation is near together and instead of them it creates new feature based on following formula:

$$\text{Cov}(x, y) = \sum_{i=1}^{N} \frac{(x_i - x)(y_i - y)}{N} \qquad (2.2.1)$$

In this equation x is mean of the x and y is mean of the y and N is dimension of the dataset. Cov(x, y) matrix subtracts the mean of each sample so by this operation data centered on the mean. After this operation eigenvectors and eigenvalues are ready for analyzing. First eigenvectors should order by eigenvalues from highest to lowest. The eigenvector with the highest eigenvalue is the first principle component of the data set (PC1) and second principal component is the eigenvector with second highest eigenvalue (PC2). The next step is deleting the less important components to have fewer dimensions than original dataset. Therefore, new dataset are calculated by multiplying each row of the eigenvectors with the sorted eigenvalues. After that principal components multiplying with transpose of original data and new data is create.

The best way to test minimal error between the original dataset and the PCA result is to calculate following formula:

$$\frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{N} \lambda_i} > \theta \text{ , } (\theta \text{ is } 0.95) \qquad (2.2.2)$$

Here, K is new dataset's dimension and N is original dataset's dimension and θ is a threshold, λ is an eigenvectors of the covariance matrix.

2.3. SUPPORT VECTOR MACHINES

History of SVM

In 1936, R. A. Fisher suggested the first algorithm for pattern recognition [21], In 1957 Frank Rosenblatt invented a linear classifier called the perceptron the simplest kind of feed forward neural network. After six years Aizerman, Braverman and Rozonoer introduced the geometrical interpretation of the kernels as inner products in a feature space. In 1964, by the passage of time the necessity of extistance the better algorithm lead to development of the Generalized Portrait algorithm "Support Vector Machines"which was introduced by Vladimir Vapnik and Chervonenkis [22].
Support vector machines (SVM) are important in machine learning and pattern recognition methods for classification. Classification is one of the most important tasks of data mining in our days. Every day the data numbers are increasing therefore, the need of use of classification is highly important. With large number of data set the first step is to create training set with features, the goal is to classify unseen data with a model that is reliable. SVM was applied in many different areas such as data mining.

Though SVM theory is accepted, some unpredictable results were shown during data classification by this algorithm. In other word, in some applications, data become increasingly large and SVM suffers from large, noisy and unbalanced data. In order to solve such problems, SVM was expanded to robust SVM that contains non-linearly separation and multiclass classification.

2.3.1. Two Class Linearly Separable

Support vector learning machine means to determine functions that can be used to classify data points. Data points are viewed as an n-dimensional vectors and our question is to know how we can separate points with n-1 dimensional hyper planes. For this reason this type of SVM called a linear classifier.

There are many hyper planes; however, the best one is the one that represent the largest margin between the two classes.

The case we are looking for is the hyper plane with maximum distance between it and nearest data point on each side.

In formula 2.3.1.1, $\{X_1, \dots, X_n\}$ is our data set or training set and $y_i \rightarrow \{+1, -1\}$ is the class label of X_i.

$$x^i, y_i \in IR^n \times +1, -1 , \quad i = 1, \dots, l \tag{2.3.1.1}$$

Many decision boundaries can separate these two classes. In formula 2.3.4, f_l is linear decision boundary function where w is decision hyper plane normal vector. Y is class of our dataset (+1, -1) and X is dataset:

$$f_l(x) : w^T x_i + b = 0 \tag{2.3.1.2}$$

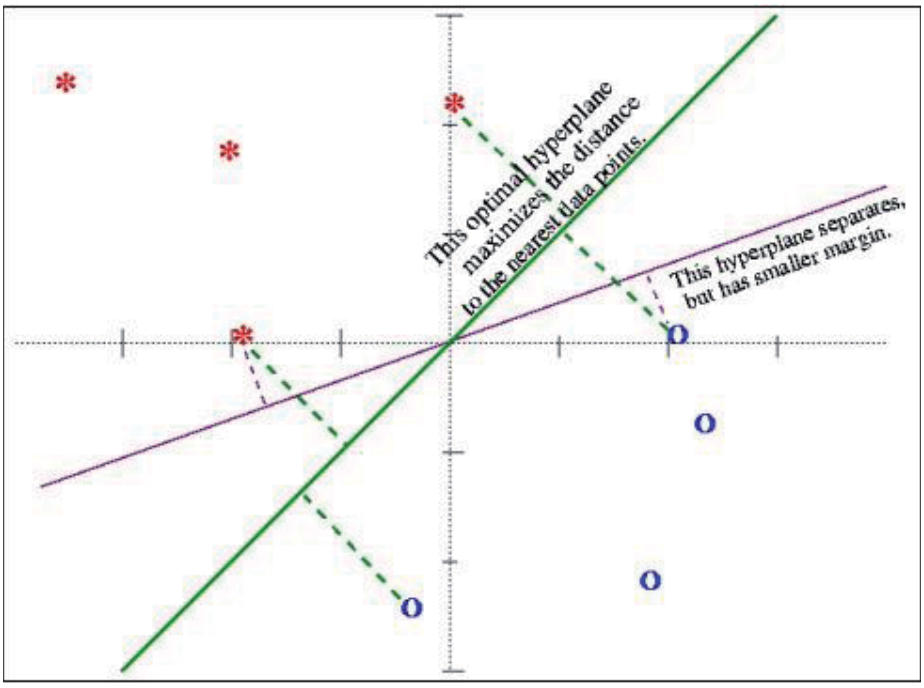

Figure 2.2: Maximum-margin hyper plane and margins [23]

The optimal decision boundary should be as far away from the data of both classes as possible. The decision boundary should classify all points correctly (Figure 2.2). Using assumptions of statistical learning theory the desired classifier have to define as below:

$$w.\ x_i\text{-}b \geq +1 \qquad (2.3.1.3)$$

And

$$w.\ x_i\text{-}b \geq -1$$

By using geometry, the distance between two hyper planes is $\frac{2}{||w||}$ and the purpose is to minimize ||w||. Some data point maybe falling into the margin so we rewrite above functions like as following:

For each *i*:

$w.\ x_i\text{-}b \geq +1$ for x_i of the first class (2.3.1.4)

$w.\ x_i\text{-}b \geq -1$ for x_i of the second class

2.3.2. Two Class Non-Linearly Separable

In some cases data points spread non-linearly, it doesn't mean that there isn't any way to classify the data. By using polynomial curves and circles it might be easier to classify data. However it is difficult in some cases to find optimal curves or circles. In 1992 Bernhard E. Boser and Valdimir N. Vapnik et al., find a way to classify non-linear dataset by using kernel trick.

The general idea is the original feature space can always transfer to upper dimensional feature space. Figure 2.3 shows spreading the data in a feature space [23].

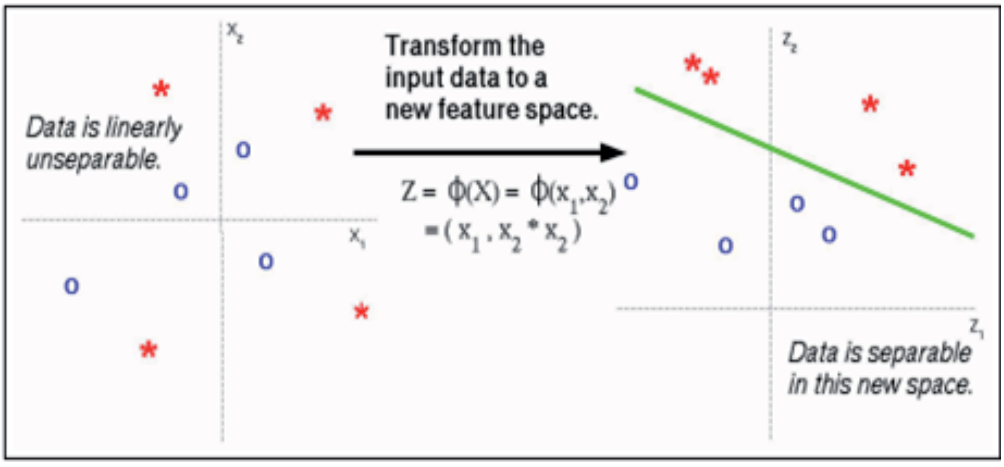

Figure 2.3: Separating the Data in a Feature Space [23]

To do this, the linear classifier relies on inner product between vectors K ($\mathbf{x}_i$, $\mathbf{x}_j$) = $\mathbf{x}_i^T\mathbf{x}_j$

After mapping all data points by Φ: $\mathbf{x} \rightarrow \varphi(\mathbf{x})$ transformation into high dimensional space, the inner products becomes:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\varphi}(\mathbf{x}_i)^T\boldsymbol{\varphi}(\mathbf{x}_j) \qquad (2.3.2.1)$$

Kernel functions:

Linear Kernel: K ($\mathbf{x}_i$, $\mathbf{x}_j$) = $\mathbf{x}_i^T\mathbf{x}_j$

Mapping Φ: $\mathbf{x} \rightarrow \varphi(\mathbf{x})$, where $\boldsymbol{\varphi}(\mathbf{x})$ is $\mathbf{x}$ itself

Polynomial of power p: $K(\mathbf{x}_i, \mathbf{xj}) = (1+ \mathbf{x}_i^T\mathbf{x}_j)^p$

Mapping Φ: $\mathbf{x} \rightarrow \varphi(\mathbf{x})$, where $\boldsymbol{\varphi}(\mathbf{x})$ has $\binom{d+p}{p}$ dimensions

Gaussian Kernel: $K(\mathbf{x}_i,\mathbf{x}_j) = e^{-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{2\sigma^2}}$

Mapping Φ: $\mathbf{x} \rightarrow \boldsymbol{\varphi}(\mathbf{x})$, where $\boldsymbol{\varphi}(\mathbf{x})$ is infinite-dimensional: every point is mapped to *a* function (a Gaussian); combination of functions for support vectors is the separator.

2.3.3. Multi Class SVM

SVM framework works for binary (+1,-1 or 0, 1) and multi class problems. The main idea for multiclass classification problem is to decompose into multiple binary classification problems. There are simple but effective approaches:

One-against-all (OvA)

The first solution for N>2 classes, N SVM classifiers are constructed. The i^{th} class as a train set labeled the samples as positive examples and all the rest as test set labeled as negative examples.

The disadvantage of this process is the complexity of training set, the number of samples in training set is large and each of the N classifiers is trained using all available samples [24].

One-against-One (OvO)

This strategy also known as "pairwise coupling", this algorithm works $\frac{N(N-1)}{2}$ two-class classifiers. Each classifier trained first class as positive examples and the second class as negative examples. The advantage of this method is the lower number of instances that causes lower nonlinearity, resulting in shorter training time. The disadvantage of this

method is in the test samples because every test samples has to present to the large number of classifier $\frac{N(N-1)}{2}$. This results in slower testing, especially when the number of the classes in the problem is big [24].

2.4. BAYESIAN NETWORKS

Bayesian Networks (BN) were first introduced in the 1980s as a probabilistic model of the problem [25]. Pearl et al. in 1985 indicated that they can sketch a graph that the nodes could indicate random variables. These nodes are drawn like circles that are labeled by their names and links or edges which represent direct dependence among the variables. Moreover, it was suggested that the network could be improved by given weights to links that hold out type and power of dependencies between connections [25]. These weights are to be used as conditional probabilities later.

In early 1990s, research has been focused on figuring out probabilities of the graphs in the medicine domain. From machine learning models, Bayesian Network can be considered as a beneficial model of this group. Through the usage of Bayesian Networks' tools, medical analysis can be easier due to the fact of the user friendly design of its tools. There are three steps for creating a meaningful structure. First step represent the reason under uncertainties in the medical problem; as a results, variables of network should partition into three types: query, evidence and intermediary variables. Query variables are used for raising questions about variable to compute its posterior border. An evidence variable is a variable that shows the assertion of that evidence. Finally, an intermediary variable is used for obtaining more detail about relationship between evidence and query variables so it is neither query nor evidence.

Following the first step, the second step is being used and graphs can be formed. The collective information about the graph is very useful since the edge of network can be determined and provides an effective reason about links. In this step, each variable (X) from the network can be analyzed.

Third step, calculate the dependencies of relations between features and variables that define as CPT (Conditional Probability Table). For a directed model of graph we must represented probabilities as a table if the variables are discrete. Every child node takes the probability that it is a combination of values of its parents. Considering all cases the value of label in all nodes are different such as (True or False), (0 or 1), (yes or no), or some nodes might have several stages, for example T_1,T_2,…, T_n, and so on.

Bayesian Network is a direct acyclic graphical (DAG) representation based on Bayes theory. All the links are directed without any cycle, which means it is not possible to traverse from the last node to the start point of the graph.

Figure 2.4, illustrates a directed Bayesian graph. This is a DAG graph because all the edges are directed and as mentioned, this is an acyclic graph (when you leave one node you cannot cycle back to the original vertex).

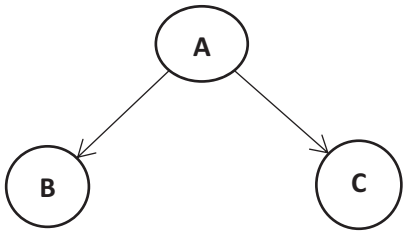

Figure 2.4: A Bayesian Network

In the above figure we can see edges between A, B and A, C so set of edges is E= {(A, B), (A, B)} [2]. In this example the joint distribution of variables is shown in the following formula:

$$P(A, B, C) = P(B|A).\ P(A)\ .\ P(C|A) \tag{2.4.1}$$

In total directed edges, if one edge from X vertex goes to another vertex Y, we called X is Y's parent. Each node has conditional distribution probability that shows numerical influences of other parent nodes to this node.

$$P(X_i|Parents(X_i)) \tag{2.4.2}$$

We assume that there are N an independent nodes, $X=X_1, \dots, X_n$, then according to the *chain rule* in probability can express any statistical distribution for Bayesian Network as:

$$f(x) = \prod_{i=1}^{n} f(x_i|X_1 = x_1, \dots, X_{i-1} = x_{i-1}) \tag{2.4.3}$$

Expansion of above formula is:

$$P(x_1, \dots, x_n) = P(x_n|x_{n-1}, \dots, x_1)P(x_{n-1}|x_{n-2}, \dots, x_1)\dots P(x_2|x_1)P(x_1) \tag{2.4.4}$$

By comparing formula (2.4.3) with (2.4.4) we can conclude that when Bayesian network is equal to statistical distribution while for each variable X_i in network we have:

$$P(X_i|X_{i-1}, \dots, X_1) = P(X_i|Parents(X_i)) \tag{2.4.5}$$

If we denote that the parents of each vertex x_i in the graph by π_i, $\pi_i \subseteq \{x_1, x_2, \dots, x_{i-1}\}$, we can rewrite (2.4.2) as [3]:

$$f(x) = \prod_{i=1}^{n} f(x_i|x_1, \dots, x_{i-1}) = \prod_{i=1}^{n} f(x_i|\pi_i) \tag{2.4.6}$$

By using this equation there is no need to define and calculate full multi-dimensional density functions because only calculated the conditional density function for each X_i is enough. It is so easier than to calculate the overall distribution. Local distribution function is determined and calculates the global distribution function too.

Different areas of use for a Bayesian Network in Medicine

Usage of Bayesian Network can be in several sectors including Medicine and Healthcare. The main fields of Bayesian Network are Diagnostic, Prognostic and Treatment selection. The explanation for how each network is used in three fields of medicine sector will be given.

Diagnostic reasoning

In diagnostic reasoning field the Bayesian network field is used for estimation of having certain disease based on the probability that earn by observations of symptoms, signs and test results. The network output will be the probability that the patient with these signs has a certain disease or not, this means that test result never come as an exact probability such 100% or 0% probability.

The routine of diagnostic in Bayesian network is requiring a test for decrease the number of uncertainty of disease. After this step network will be waiting for test result as an input that entering by user. Based on result of test the new possibilities will be calculated.

Prognostic reasoning

Prognostic reasoning is the knowledge on a certain facts for making prediction about what will be happen in the future. This field is less certain compared with diagnostic because prediction depends on the available data and information about patient and there are no facts about the future of disease. The routine of this field is to predict the output of specific patient by collection data and suggest the treatments.

Treatment selection

Bayesian network by itself cannot provide a decision in other word, making process to decide which method is the best; however, it is a combination of first part that it named diagnostic part with calculation the probability of each treatment, the optimal treatment alternatives comes out.

2.5. Previous Work

The goal of this review is to identify articles dealing with the use of Bayesian networks and Support Vector Machine models to diagnosis and prediction of cancer and medical cases. I will note the strengths and methods of articles written to address this problem. The research described later in this thesis will represent an effort to address the gaps that are identified in the body of past research.

After the research about classification on disease, one of the first example articles provided by Dominik Aronksy and Peter J.Haug, in which, they presented the development and the evaluation of a Bayesian network for diagnosis of community acquired pneumonia. For train and test the network they extracted 32,000 emergency room patients from the clinical information system over a period of 2 years (June 1995-June 1997). This network performed well for all kind of disease and they achieved a high sensitivity about 95% and area under the curve (AUC) was 0.98. This study demonstrated that Bayesian network is a proper method for pneumonia patient's detection [26].

Basilio Sierra and Pedro Larra-naga have reported different methods of induction of Bayesian networks. These methods made by Genetic Algorithms and these methodologies are applied to the problem of people who are predicting survival after one, three and five years of being diagnosed as having malignant skin melanoma. The accuracy that obtained from models is from 10-fold cross-validation method and the average of accuracy for one year survival was 93.06, for three years survival was 82.18 and for five years was 73.60[27] .

Nathan Hoot and Dominik Aronsky, they used transplant information to construct a Bayesian network model to predict 90-day survival patients from the United Network for Organ Sharing database. The final model of network achieved accuracy by area under the receiver operating characteristic curve of 0.67 by 10 fold cross-validation and by independent validation set they achieved 0.68. Result for positive predictive value showed 91% and it was useful to clinical experience for patients to have good outputs following liver transplantation [28].

Oliver Gevaert et al. have proposed a strategy based on the Bayesian networks use clinical data and microarray data together. The advantage of using these two kinds of data is that it allows understanding the model structure and parameters and it can be integrated in several ways. Bayesian networks by identifying the dependency and independency relationships between class variable can automatically select features. For integrating clinical data and microarray data three methods have evaluated: decision integration, partial integration and full integration. Data set is publicly available data on breast cancer patients and separated into two groups poor and a good prognosis groups. The partial integration group has 0.84 areas under the curve with independent test set. [29]

Jiakai Li et al. developed Bayesian network classifier for prediction of renal transplantation and period of survival of graft. The dataset contains profile information of patients before the transplantation from the University Of Toledo

Medical Center Hospital. P1228 patients records during 1987 through 2009 to the United Network Organ Sharing. The aim of this paper was to construct a decision making tool for identify the suitable member among recipients. For Bayesian network model they used Weka machine learning software. There were two classifiers for the data set for this study. First one was failed or living status of the graft and second one was predict period of the graft survival. First classifier accuracy for graft status was 97.8% and it performed very well and second classifier accuracy was 68.2% .results indicated that Bayesian belief network is more feasible and more successful than other classifier [30-33]

Jayasurya K et al. had compared two machine learning methods such as support vector machine (SVM) and Bayesian networks (BN) in this study for prediction lung cancer output for two years survival patients. For SVM models all the input variables should be known but in Bayesian network model might handle missing data better. The authors assumed that Bayesian networks predicted more accurate cases when data was incomplete. These two models were trained on 322 patients without operation that treated with radiotherapy from Maastricht University. They separated into 3 data sets of 35, 47 and 33 from Ghent, Leuven, and Toronto. For this data set Bayesian network model had area under the curve of 0.77, 0.72 and 0.70 in the same way SVM area under the curve in this model was 0.71, 0.68, and 0.69 respectively. When the dataset has complete data, the Bayesian Network and Support Vector Machine acts more alike but totally, BN models acts well than SVM at handling missing data for the medical domain. [34, 35]

Terence Ng, et al. had purpose with this study to create a decision support tool for clinics to predict survival patients after chemotherapy after 120 days. Often clinicians make incorrect predictions whereas the accurate prediction can help clinicians decide on the patients care. Data set were obtained 400 ill cancer patients in the National Cancer Centre Singapore (NCCS) from 2008 to 2009. There were some missing data by removing them, 325 patients remaining. For this study three classification methods were used such as, naive Bayes (NB), neural network (NN), and support vector machine (SVM). Results for this study showed that the NN model with 78% accuracy and 0.85 area under the curve had the best prediction. [36]

M. Berkan Sesen et al. had research about prediction of survival and treatment selection in lung cancer care. They used Bayesian networks to help experts by estimates survival and treatment selection recommendations. The database is from the English Lung Cancer Database (LUCADA), they compared the accuracies of various approaches to figure out the best structure from knowledge of experts and data set. The results showed for first task that the BN structure achieved area under the ROC curve of 0.75, however Bayesian network for second tasks can only predict the recorded treatment of 29%.[37]

2.6. WEKA and SAMIAM

- **Waikato Environment for Knowledge Analysis**

Figure 2.5: The WEKA GUI chooser

For Data mining and Machine learning there are many available tools for analyzing data, but in this thesis, we've decided to use WEKA which stands for Waikato Environment for Knowledge Analysis. WEKA is open source software that was produce by the University of Waikato in New Zealand and used under the GNU General Public License. The reason that we selected WEKA was for versatility. We can apply a lot of algorithms directly to huge dataset in WEKA [38].

As seen in Figure 2.5, the Explorer, Experimenter and the knowledge Flow is three different graphical interfaces and also there is text based terminal mode that you can calls different methods by function calls directly. The Explorer, the user gains all equipment and features in Weka for quickly analyze the data. The Experimenter is focused on comparing different algorithms against each other. Lastly, in the knowledge Flow the users can set up complex flows to does the total chain from reading the data to plotting the result in a graph. The main benefit of Weka is easy to try various algorithms and filters and finally find a suitable method.

In following part there are some Weka's features:

- **Data Preprocessing** –to check the instability of the data and to eliminate incorrect values preprocessing data cleaning is used. Conversion of a set of data values from the data format of a source data system into the data format of a destination data system can be done by data transformation. Weka supports popular text files such as CSV, JSON and Matlab ASCII files and ARFF.
- **Data Classification** – various algorithms have been implemented such as Bayesian algorithms, Support Vector Machine, Nearest Neighbor Calculations, Tree-based classifiers.

- **Data Clustering** – K- means method as well as density and hierarchical based clustering algorithms.
- **Attribute Association** – describing relations between data in dataset.
- **Attribute Selection** - methods to recognition which attribute help to predict outcome.
- **Data Visualization** – the plot view of data against suitable variables.

➢ **Sensitivity Analysis, Modeling, Inference and More**

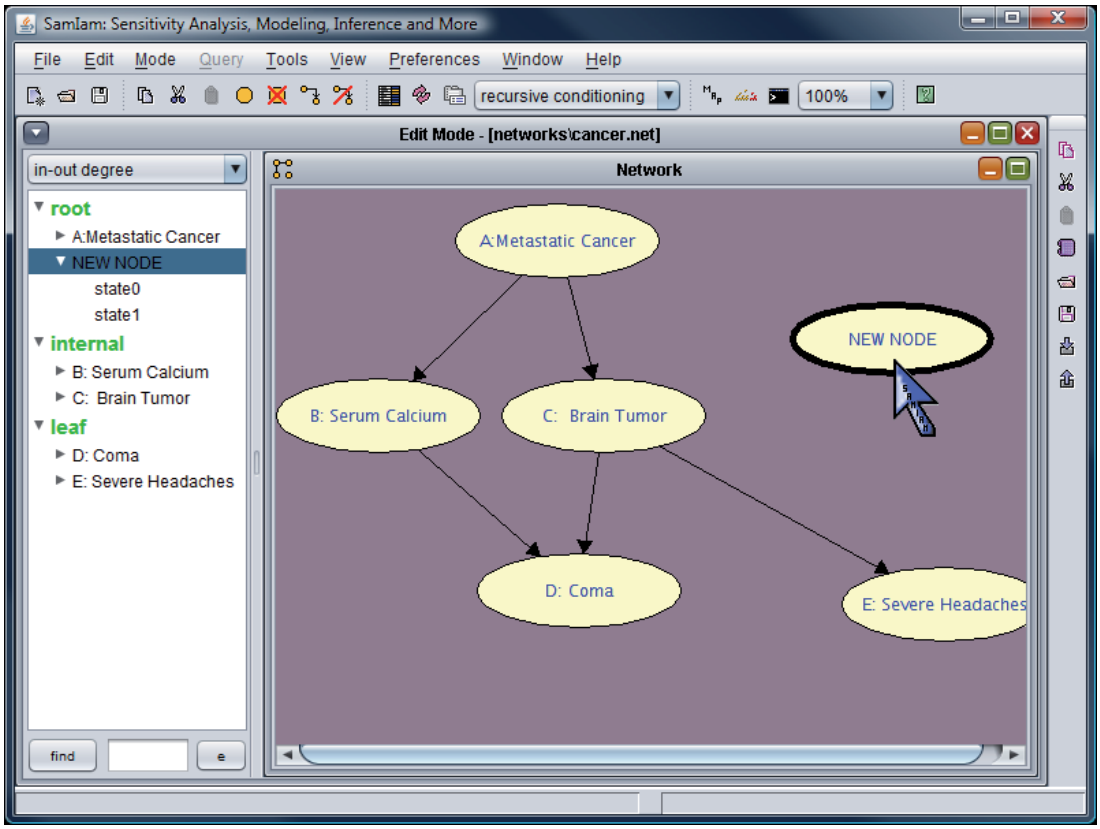

Figure 2.6: SamIam

SamIam is a total solution tool for modeling Bayesian Networks with reasoning. This software developed in java by Professor Darwiche research group at UCLA. Figure 2.6 is screen shot of SamIam, there are two main components in SamIam, first one is graphical user interface (GUI) and second one is reasoning engine.
First component allows users create various Bayesian networks models and save them in a variety formats. Second component includes classical conclusion, parameter estimation, sensitivity analysis, maximum a posteriori (MAP) and most probable explanation (MPE), expectation maximization (EM).
SamIam has a lot of features and tools; here we try to explain some of them [39]

- **Batch Tool** – batch tool allows users to write command and queries programmatically to achieve the output. Input file specified by commands in an XML file, and will write out its results to an XML output file.

- **Code Bandit** – there are sample and ready codes for users based on the user configuration in SamIam. For example for creating Bayesian Network models, code bandit makes it easy and write programs that shows how to run queries.
- **Editing Models –** for editing Bayesian networks, SamIam provides editing tools such as: copy, paste and cut, zooming of networks fragments. Conditional probability table for normalization.
- **EM learning -** Expectation Maximization algorithm is for estimating network's parameter based on the case file that you selected. The data should be form of a Hugin "case file" format. SamIam generate random data from tool-> generate simulated case by given network for sorting the data with case file format.
- **File Formats** – SamIam can read and save Bayesian Networks in different formats because use SMILE (Structural Modeling, Inference, and Learning Engine) library for load and read models.
    - .net of Hugin;
    - .dsl and .xdsl formats of Genie;
    - .dsc format of the Bayesian Network Toolkit in Microsoft;
    - .dne format of Netica;
    - .erg format of Ergo;
- **Inference –** SamIam algorithms which use in Bayesian Networks:
    - the Hugin architecture
    - the Shenoy-Shafer architecture
    - the Recursive Conditioning
- **Maximum á Posteriori (MAP) –** by using evidences it is like to configuration of a specific set based on maximum posterior. For complex networks it is hard to calculate MAP but SamIam solve this complexibility by allowing users to calculate the exact answer.
- **Most Probable Explanation (MPE) –** by using evidences it is like to configuration of all network variables. It is a special case of Maximum a Posteriori probability or MAP,
- **Sensitivity Analysis** – This method is used for debugging models. Besides, through sensitivity analysis, the relationship between parameters in network can be understood.

CHAPTER 3

3. MATERIAL AND METHODS

3.1. DATASET

One hundred and forty-six patients with esophageal carcinoma were treated with esophagectomy between March 2003 and December 2011 at the Thoracic Surgery Clinic, Ankara Numune Teaching and Research Hospital. Clinic pathological variables and survival times were measured and collected by telephone contact and patient's medical records. Out of 146 patients, survival information was available for 119 patients. The preoperative clinical factors including age, sex, type and duration of symptoms, laboratory findings, tumour length, and location of the tumour, histological differentiation, operation type, and pathological TNM staging were all analyzed.

The preoperative survey included physical tests, laboratory tests, esophagogastroduodenoscopy, flexible bronchoscopy, barium esophagography, computed tomography (CT) scan from neck to upper abdomen, ultrasound of the abdomen, and radionuclide bone scans. Some patients received Positron Emission Tomography (PET) CT. Pulmonary and cardiac function studies were done for assessment of surgical tolerance. In total there were 48 features for all patients.

Data strategies were as follows: We identified a total of 119 patients who underwent esophagectomy. Patients age average was 55 years (range 25-80 years). The male patient's number was 76 and female patient's number was 43 and ratio between male and female was 1.7. The in-hospital mortality was 10 out of 119 (8.4 %).

Among the 119 patients, 10 patients died during 30 days, 20 patients died in six months and 11 patients lived more than others but died after five years. All patients were retrospectively restaged histopathologically according to AJCC 7^{th} edition guidelines. This study protocol was approved by the Medical Ethics Committee of Ankara Numune Teaching and Research Hospital.

3.2. METHODS

3.2.1. CLASSIFICATION

The classifier is the important element of the data mining system, which performs the actual recognition. Its first task is to collect and determine the input, and then, extract features from input data and represent by these features to one of the classes. Classification has complexity problem that despite differences between classes, how the feature can separate into classes.

In this thesis PCA method is used for feature extraction that we use for first method. SVM and Bayesian Networks are used for classification. Our target is to chosen suitable classifier from these methods. Methods are explained in the following sections:

The first method we use is Support vector machine, the experiments is construct on the database that we described SVM in detail in previous part. For this method we apply data reduction and trained by using Sequential Minimal Optimization (SMO) [40] algorithm in WEKA. The second method we used Bayesian Networks by using SAMIAM software.

3.2.1.1. The use of PCA

The original data contain 119 samples and 48 features. In another words, we have a 119×42 matrix. By applying PCA in Matlab and Weka we obtain a new matrix that contains 119 samples and reduce features to 11. Now, the new data set is a 119×11 matrix.

In extreme high dimensional dataset for example for gene dataset in bioinformatics approaches, PCA algorithms are limited because of the existence of large data and large covariance matrix. But for normal dataset we can use classical PCA algorithm that often we only need the first two or three principal components to visualize the whole data and these two components give us the all information that we can get from data.

For extracting the first (n) components, PCA Matlab script (princomp.m) is available for users in Statistics Toolbox. Also we can use base Matlab code for calculate coefficient and respective variances of the principal component.

The coefficients for each principal component and diagonal elements that store the variance of the respective principal components are obtained from Matlab. The columns are in order of decreasing component variance, after change it to increasing order the first eleven components are enough for us, because the first principal components has the largest possible variance. These eleven components give us the enough information.

Depending on dataset the filter is select N features (including class attribute) for each of the samples.

3.2.1.2. The use of SVM at this study

We have used Weka software for classification of dataset, as it was outlined in chapter 2. For first step of using Weka, we use the converter for change Excel files to Arff format.

An ARFF (Attribute-Relation File Format) file is an ASCII text file. This file format 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 with the Weka machine learning software [41].

Next step is choose kernel type for classifying data. For this reason the kernel function is important. The two commonly used kernels are linear and RBF kernels.

The case of one dimentional in poly kernel is linear kernel the constant 1 just changing the threshold and the case of two and three are quadratic and cubic kernels.

RBF kernel is similar to mapping the data into an infinite dimensional space and we can not show the radial basis function concretely. This kind of kernel allows to have features that it can be detect by circles (hyperspheres) and the decision boundries become more complex beause of interaction of multiple features.

For our study we use two models of kernels: “Linear Kernel” model and “RBF Kernel” these kernels replaces all missing values and transforms nominal attributes into binary ones and also normalized all features by default.

We used SMO algorithms which is widely used for training support vector machines. For more information on the SMO algorithm, see [42]. Confusion Matrix and Receiver Operating Characteristic (ROC) are used for analyzing outcomes.

3.2.1.3. The use of Bayesian network at this study

Prognosis is the prediction of the outcome of a disease. It classifies patients into several categories according to probability of the network. In this study, we prepare Bayesian network which can predict output of the disease. For achieving this goal we have used SAMIAM and WEKA software which was introduced along chapter 2. In this study we discuss a prognostic prediction of esophageal cancer cases. Figure 3.1 displays a screen-shot of SAMIAM.

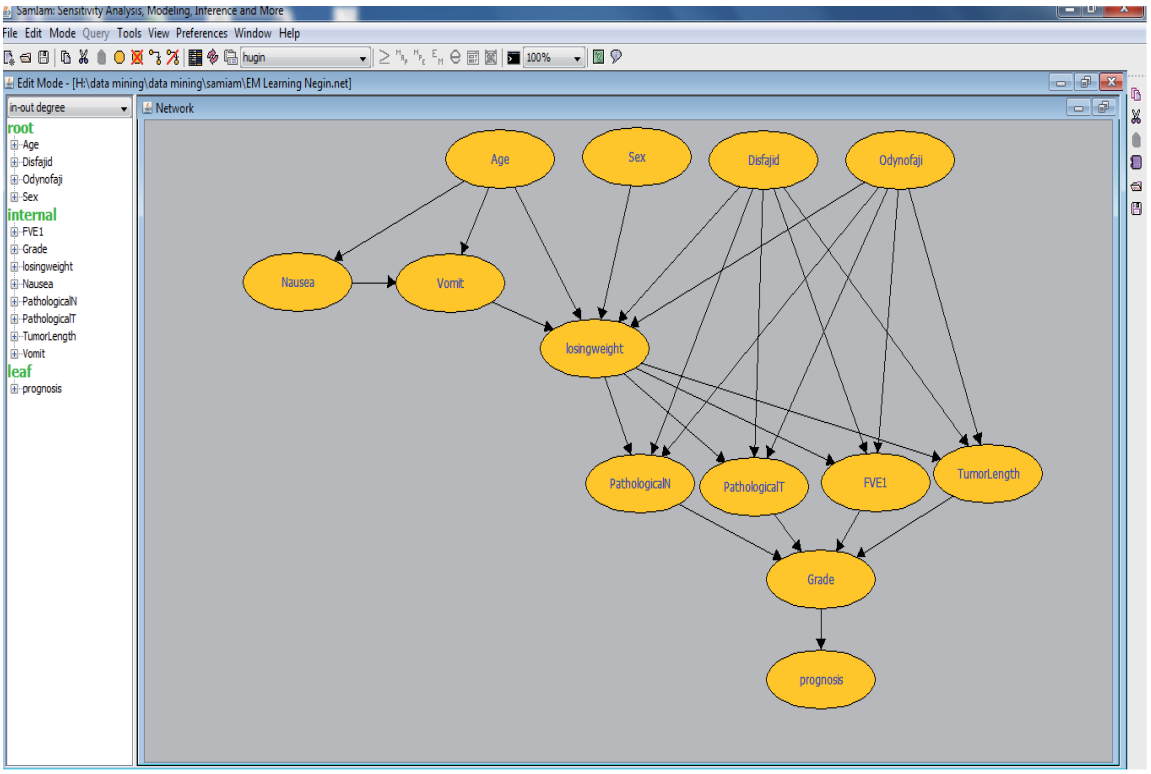

Figure 3.1: Screen-shot of SAMIAM, Esophagus Network.

Bayesian network is created for predicting output of Esophagus cancer. By using knowledge of casual dependences, influences or correlations and knowledge of experts we build the structure of our Bayesian network.

Our Bayesian network has been created by hand. Manual networks are built in several stages and for each of stages required available patient data and related experts in the domain of disease. Most medical experts are familiar with computer applications and software and they tend to use the diagrams and networks so that they can communicate with the engineer. All the available patient data were not used to predict output of network. The main and useful variables and connections between nodes can be identified by using correlation.

Correlation is a measure of linear dependency between two variables and it defines as the covariance of two variables divided by the product of their standard deviations.

Population correlation coefficient formula is:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \qquad (3.2.1.3.1)$$

Where, "cov" is the covariance, $\sigma_X$and σ_Y is the standard deviation of X and Y, μ_X and μ_Y is the mean of X and Y, E is the expectation.

Table 3.1: Correlation of Variables

	Age	*Sex*	*LostWeight*	*Dysphagia*	*Odynophagia*	*Nausea*	*Vomit*	*FEV 1*	*Stage*	*Grade*	*Prognosis*
Age	1										
Sex	-0.094	1									
Lost Weight	-0.231	0.11	1								
Dysphagia	-0.051	0.15	0.074	1							
Odynophagia	-0.103	-0.075	-0.014	0.122	1						
Nausea	-0.219	0.041	0.23	0.081	0.362	1					
Vomit	-0.143	0.019	0.171	-0.328	0.265	0.767	1				
FEV 1	-0.112	0.073	0.074	-0.03	-0.059	0.062	0.063	1			
Stage	0.093	-0.18	-0.052	0.023	-0.146	0.032	-0.002	-0.115	1		
Grade	0.026	-0.042	-0.063	0.175	-0.096	-0.054	-0.067	-0.196	0.412	1	
Prognosis	0.019	-0.033	-0.158	-0.097	-0.225	-0.078	-0.062	0.386	-0.09	-0.281	1

Definition of correlation coefficient is the value between +1 and -1 that explain us about relations between variables. A coefficient +1 indicates a direct relationship or positive correlation. If variable X increases, variable Y increases in the same way variable X decreases, variable Y decreases too. Coefficient -1 indicates indirect relation or negative correlation. As variable X increases, variable Y decreases, variable X decreases, variable Y increases. When coefficient is 0 it means there is no correlation.

After calculating all correlations, our observation is as Table 3.1. Based on the correlation results we can connect edges between nodes. Each disease has its own connections and it all depends on doctor and engineer. The main objective of our research is to assess which clinic pathologic factors influence the mortality and survival in esophagus cancer that underwent curative surgery without preoperative chemotherapy and our results and network by conference with expert and search in websites about prognosis, diagnosis and treatment of disease confirm the importance of 11 factors and connection among nodes may affect prognosis.

A typical clinical question might be who live shorter or longer than 5 years? For answering such questions some factors have to be considered that affects the results of this question. The most related factors to the prognosis is common for all cases, these are age, sex, Dysphagia, Odynophagia, lost weight and vomit and nausea. Other important factors for this kind of disease are the response of pathological N, T, FEV1, tumor grade, tumor length.

The first part of features summarized as Table 3.2:

Table 3.2: Features and Labels

FEATURES	LABELS
Age	25-35, 36-45, 46-55, 56-65, 66-75, 76-85
Sex	Women , Men
Dysphagia, Odynophagia	Yes , No
Vomit and Nausea	Yes , No
Lost weight	1-5, 6-10, 11-15, >16

Tumor Length

Tumor length is one of the most important factors in esophagus cancer, the length's unit is millimeters (mm) and we divide it into 10 with 0.5 gap for each part like as (1-1.5)… (10-10.5).

Forced Expiratory Volume in 1 second (FEV1)

It measures the volume of air in liters that patient give out during the one second, then for calculate the ratio it is converted to the percentage. Average of FEV depends on age and sex of patients.

If FEV1 is greater than 80% it is normal but less than 40% is predicted extreme obstruction. We divided this factor into 5 parts such as (1.1-1.9)… (5.1-5.9).

Cancer Staging

The stage of cancer explains the extension of the cancer in patient's body. It depends on how far the cancer has grown. Is it reached nearby structure or is it spread to the lymph nodes or other organs? The most effective factors in prognosis and treatment options are determining the stage of cancer.

There are 2 types of staging [43]:

- *Clinical stage*, determine by doctor's observations of disease based on the results of physical tests or any imaging scans.
- *Pathological stage*, determine by what is found as a results of the surgery and plus the same signs as the clinical stage.

There are some differences between two stages, in some cases during the surgery the doctors find out accurate tumors place in some areas that imaging test do not able to recognize. So many doctors and patients prefer to pathological stage because it allows the doctor to get a firsthand perception of the extent of the disease.

(TNM) Staging System

TNM staging system is the old system for describing the extent of cancer. This system has been accepted by the American Joint Committee on Cancer (AJCC) and International Cancer Control (UICC) [44].

The TNM is determined by 3 factors that we use in this network: T, N, and M.

T= location and size of Tumor; describes how and where the main tumor has grown in the wall of the esophagus.
N= displaces the tumors nearby lymph nodes; Lymph Nodes are small circle-shaped nodes that resistance in front of infections.
M= shows whether the cancer has leaped (Metastases) to other part of the body.

Numbers and letters after TNM is for more details, in following table shows the grouping stage of TNM system:

Table 3.3: Stage Grouping

Stage	Tumor	Regional Lymph Nodes	Distant Metastases
Stage 0	Tis	N0	M0
Stage IA	T1	N0	M0
Stage IB	T1 T2,T3	N0	M0
Stage IIA	T2,T3	N0	M0
Stage IIB	T2,T3 T1,T2	N0 N1	M0 M0
Stage IIIA	T1,T2 T3 T4a	N2 N1 N0	M0 M0 M0
Stage IIIB	T3	N2	M0
Stage IIIC	T4a T4b Any T	N1,N2 Any N N3	M0 M0 M0
Stage IV	Any T	Any N	M1

Stage grouping is started from stage 0 to stage IV, due to Table 3.3; each letter has its own description by numbers which we will explain below:

T categories:

TX= the early tumor cannot be evaluated.

T0= there is no indication of primary tumor.

Tis= the cancer is in the first level. It can be seen cancer cells in the upper layer of cells lining esophagus.

T1= the cancer cells bring on into the underlying tissue.

T2= the cancer cells bring on into the thick tissue (muscle layer).

T3= the cancer cells bring on into the out cortex of esophagus.

T4= the cancer cells bring out into nearby organs.

- **T4a**= the cancer cells is covering the lung, heart and diaphragm and can be removed by surgery.
- **T4b**= the cancer cells covering main part of aorta, spine and other crucial organs.

N categories:

NX= surrounding lymph nodes cannot be evaluated.

N0= the cancer cells has not bring out to around lymph nodes.

N1= the cancer cells is found nearby 1 or 2 lymph nodes.

N2= the cancer cells is found nearby 3 to 6 lymph nodes.

M categories:

M0= the cancer has not distribute to other organs of body.

M1= the cancer has distribute to other organs of body.

Grade

The description of grade is how cancer cell looks like normal under the microscope. High number of grade is tend to cancer grow faster than lower grade.

GX= the grade cannot be evaluated.

G1= cells are completely separated and they looks like to normal cells.

G2= almost cells are different and it shows more abnormal.

G3= cells are barely distinguishable and it is very abnormal.

Prognosis

The graphic of Bayesian network is for classifying patient based on prognosis of survival after 30 days, 6-12-18 months and 3-5 years.

The prognosis results is the answers to patients who want to know the expectancy of their lives after surgery.

CHAPTER 4

4. RESULTS AND DISCUSSION

4.1. Overview

In this chapter we present classifications performance and quality results of methods were explained in chapter 3. The results of methods reported here were carrying out on the esophagus data set, which was previously described in detail.

4.2. Results of SVM

4.2.1. PCA

We applied principal component analysis feature extraction method to the training data set. The coefficients calculated for data set by using Weka's tool. We show the 2-D plots of the components, which they having the largest eigen values. We can obtain better classification with more principal components but it is not possible to show the separation with more than 2 components. In order to observe the separation of PCA applied training data we presented the 2-D plots of the first two principal components in Figure 4.1.

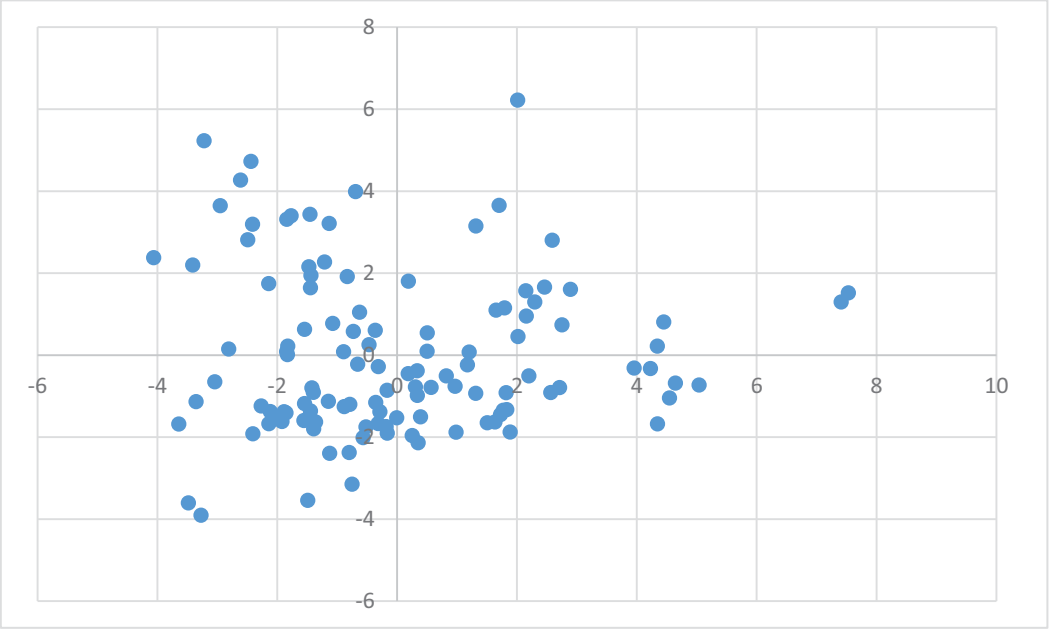

Figure 4.1: PCA reduced components

In Figure 4.2 we can see the distribution of data set. These dataset is defined 119 instances and 11 attributes. Information about the selected attribute is given by histogram depicts the attribute distribution. Labels are depicted by the seven colors in the histogram.

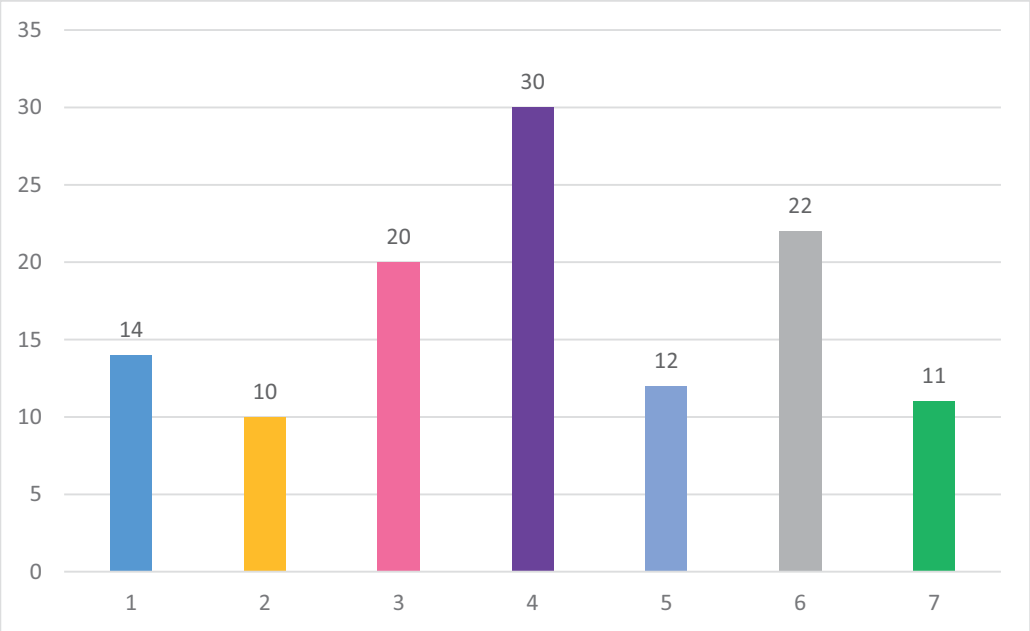

Figure 4.2: Data Set Distribution with seven labels

Figure 4.3 represents distribution of two attributes and their labels. We can see that data is not linearly distribute with two attributes. To solve a nonlinear classification problem with a linear classifier all we have to do is to transfer features to upper dimension but Computations in the upper feature space can be time consuming, due to the fact that working in higher dimension is expensive. By using RBF kernel as mentioned in previous chapter we can solve this problem.

Figure 4.3: Representation of Age and Sex Attributes

- **Linear Kernel with PCA and without PCA 42 Attributes (Multiclass)**

As we mentioned in chapter 3, the goal is to learn the possibilities offered by the Weka software for a model that we have built. The result of applying the chosen classifier will be tested according to the cross validation which has also been set by default 10- fold. Having 10 fold cross validation means 90% of all data set is used for training and 10 % is for test model in each fold test. When we choose 10 fold it is the same as we use training set which means cross validation produce a fair performance of test set.
5-fold cross validation means it contains 80% of dataset and sometimes it works well. Having more than 10 folds produce problems with small dataset.

The larger dataset needed the fewer folds to produce a robust model. As our first observation we choose linear kernel which was tested with and without PCA on data set. The output from model for this part is shown in Table 4.1.

Table 4.1: Classification Results of SVM Classifier by Using Linear Kernel with and without PCA-Total Data Multi Class Classification

Data Set	Multi Class Classifier	Correctly Classified Instances	Incorrectly Classified Instances	Time taken (Seconds)	Recall	False measure	Accuracy
Total Data (PCA)	OvA	41	64	0.39	0.38	0.30	39.04%
	OvO	**48**	**57**	**0.63**	**0.36**	**0.35**	**45.71%**
Total Data (without PCA)	OvA	37	68	0.4	0.35	0.35	35.23%
	OvO	37	68	0.64	0.35	0.35	35.23%

Above table shows the result of SVM classifier after and before using PCA in multi class dataset, as we mentioned in chapter 2, there are two solutions for multi classification problems “one against all” and “one against one”. Here, we try both of them to see which one is good for this type of data set. The best result over all the results is highlighted in bold.

The results show that OvO would be somehow slower than OvA. Though OvO is slower, it classified more correctly instances on the dataset. This results is because of methods structure that is discussed in following paragraph.

OvA select one class as a train set and labeled as positive and all the rest as negative samples for test set. Obviously negative samples are more than positive samples and classifier disregard the class which has fewer samples.
OvO trained first class as positive examples and the second class as negative examples. There is a balance between sample of classes and classifier can detect correctly. OvO classifier seems to have better accuracy than OvA.

We can see the confusion matrix of OvO classifier in Table 4.2 for this model. The classes in confusion matrix indicate the mortality of patients. Classes indicate patients who can live 30 days after surgery, patients who can live after surgery less than 6, 12 and 18 months, and some patients who can live 3 and 5 years after surgery.

Table 4.2: Confusion Matrix of Linear kernel by using OvO without PCA (42 Attributes, Multiclass)

30 Days	6 Month	12 Month	18 Month	3 Years	5 Years	Classified <=as
4	4	2	0	0	0	30 Days
4	11	4	0	1	0	6 Month
2	4	16	1	5	2	12 Month
0	2	2	3	5	0	18 Month
0	1	5	4	12	0	3 Years
0	0	5	1	3	2	5 Years

False measure and Recall are used here but before giving the definitions for these two concepts we have to describe other related concepts:

- **True Positive**: indicate the number of correct prediction that an instance is positive (TP).
- **True Negative**: indicate the number of correct predictions that an instance is negative (TN).

- **False Positive**: indicate the number of incorrect predictions that an instance is negative (FP).
- **False Negative**: indicate the number of incorrect of predictions that an instance is positive (FN).

By the definitions above, Recall is:
Measures the number of correctly classified candidates out of the number of all candidates and is given by [45]:

$$\text{Sensitivity} = \frac{TP}{TP+FP} \quad (4.2.1)$$

False Alarm Rate is specificity, which is given by [45]:

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (4.2.2)$$

In the light of this definition we can say that the result should be equal to 1 for Recall and 0 for False Alarm Rate.

As we can see it can be seen from Confusion Matrix that all compounds have been classified. It is clear that such obvious compounds are useless and it cannot be realized the correct class. However, there are 43 correctly classifieds instances and 76 Incorrectly Classified Instances. We can find the number of false positives and false negatives In the Confusion Matrix. The accuracy of the model is 36.13%. Based on this accuracy rate upon initial analysis, this is not a very good model.
The accuracy for unbalanced datasets cannot be used for the assessing usefulness of classification models. For this purpose there is a useful statistical characteristic that it is called **Receiver Operating Characteristic** (ROC), for which the value near 0.5 means the lack of any statistical dependence. As we can see in figure 4.3 the ROC curve is presented in the plot frame. The axis X indicates false positive rate and axis Y indicates true positive rate and the color displays the value of the threshold, in this way that the color that it is closer to blue corresponds to the lower threshold value. The results of ROC curve must always be convex. For visualizing the ROC curve we should click the right mouse button on the result of model type in the Result list frame and selecting the menu item Visualize threshold curve / Class.

By using the term “Area under Curve” we can measure the accuracy. If an area is 1 it represents a perfect test, but an area 0 .5 represents a worthless test.
By analyzing ROC curve and calculate area under curves we recognize that area under the curve for 2 classes in this classifier is higher than other classes. Figure 4.4 represent patients who are alive for 30 days and 6 months and both AUC are higher than baseline. We select baseline higher than 70-80 based on following arrangement [46]:

- 90-1 = excellent
- .80-.90 = good
- .70-.80 = fair
- .60-.70 = poor
- .50-.60 = fail

As we see from Figure 4.5 the accuracy of a prognostic test in academic system is fair because four classes area under curve is around 0.8.

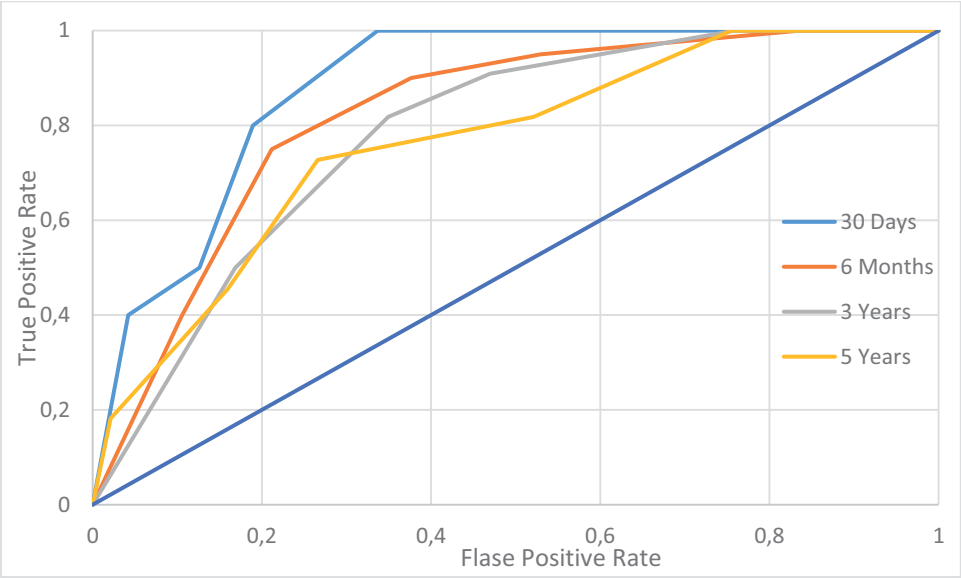

Figure 4.4: Receiver Operating Characteristic curve for SVM Classifier by Using Linear Kernel without using PCA for 30 days and 6 months and 3, 5 years alive patients.

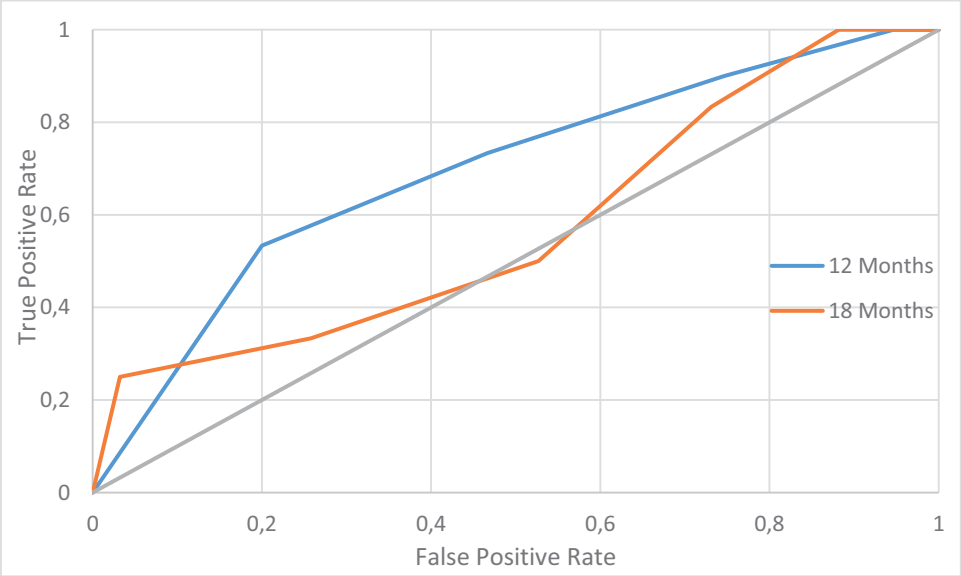

Figure 4.5: Receiver Operating Characteristic curve for SVM Classifier by Using Linear Kernel without using PCA for 12, 18 months alive patients.

- **RBF kernel with and without PCA 42 Attributes (Multiclass)**

We use RBF kernel that we described before in chapter 3 as second kernel. Here, we again try two classifiers to solve multi classification problem. The highest accuracy of this model is for OvO classifier with using PCA. The accuracy is 45.71% and this percentage shows that PCA affects results when RBF is used as a kernel; this is because of PCA that reduce the input parameters and simplify the classification for RBF kernel.

Table 4.3: Classification Results of SVM Classifier by Using RBF Kernel with and without PCA-Total Data Multi Class Classification

Data Set	Multi Class Classifier	Correctly Classified Instances	Incorrectly Classified Instances	Time taken (Seconds)	Recall	False measure	Accuracy
Total Data (PCA)	OvA	46	59	0.14	0.43	0.42	43.80%
	OvO	**48**	**57**	**0.09**	**0.45**	**0.44**	**45.71%**
Total Data without (PCA)	OvA	38	67	0.24	0.36	0.36	36.19%
	OvO	35	70	0.48	0.33	0.33	33.33%

One against One multi class classifier has K (K-1)/2 binary models such that k is the number of classes. There are two ways for select parameter:

1. First, for any two classes a parameter selection process is conducted. Finally each decision function finds its own optimal parameters.
2. For all K (K-1)/2 binary models the same parameters are selected and used for classification problems. The parameters which achieve the highest overall accuracies are selected.

Each way has its own advantages and disadvantages; often these two methods shown similar performance but first method is not always good choice, one parameter set maybe is not good for all K (K-1)/2 and may lead to over-fitting. For this purpose software usually use second approach by considering the same parameter.

As we can see in Table 4.3, this problem may be occurring in data without PCA. OvO accuracy is lower than OvA while in total observation OvO shows better

accuracy. In this problem OvO classifier cannot set optimal parameter for classification.

In Table 4.4 shows confusion matrix that the results still is not good because it shows that this model cannot detect and classify true classes.

Table 4.4: Confusion Matrix of RBF kernel by using OvO without PCA (42 Attributes, Multiclass)

30 Days	6 Month	12 Month	18 Month	3 Years	5 Years	Classified as
4	4	2	0	0	0	30 Days
3	9	6	1	1	0	6 Month
3	2	18	1	4	2	12 Month
0	2	2	3	4	1	18 Month
0	1	5	4	12	0	3 Years
0	0	4	1	4	2	5 Years

Based on the accuracy and correct/ incorrect classified instances we can say that this model is better than linear kernels model. Now we can draw Roc curve again and see the performance of model.

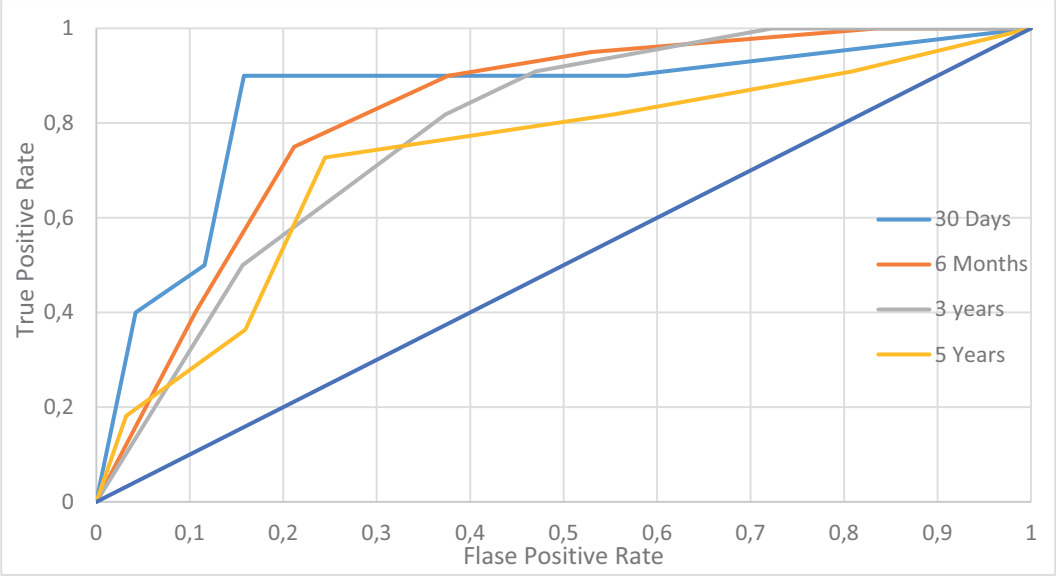

Figure 4.6: Receiver Operating Characteristic curve for SVM Classifier by Using RBF Kernel with PCA for 30 days, 6 months and 3, 5 years live patients.

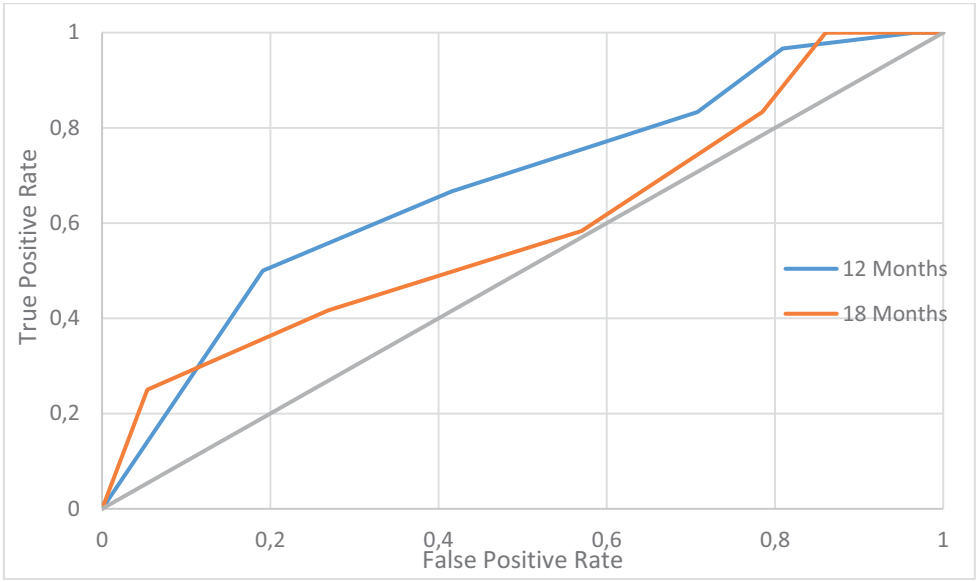

Figure 4.7: Receiver Operating Characteristic curve for SVM Classifier by Using RBF Kernel with PCA for 12, 18 months patients.

By calculating the area under curves we recognize that area under the curve for four classes is higher than other classes. Figure 4.6 represent patients who are alive for 30 days and 6 months and 3, 5 years and all classes AUC are higher than baseline. Figure 4.7 shows that for more true positive answers, classifier detects wrong classes but this classifier can detect more classes when the number of true positive increase instead of linear kernel.

In order to gain higher accuracy, we select 9 attributes from total data by expert opinion and now we try to classify 105 instances with 9 attributes and 6 classes.

“Age, Sex, Odynophagia, Vomit, and Lose weight, FEV1, Stage, and Tumor Length” selected as our features and “Prognosis” selected as a label. The same as previous method we use two kernels Linear and RBF with multiclass classifiers.

- **Linear kernel with and without PCA with 9 Attributes(Multiclass)**

By using 9 attributes we have obtained 40 correct classified instances and 79 incorrect classified instances with PCA and OvA classifier. The accuracy of this model is 35.23%. As it is shown in Table 4.5 by decreasing number of attributes, the information of data decrease and in the same manner, accuracy of the models also decreases. The dependency of accuracy to these data can be considered as the reason of these reductions. The over-fitting problem in OvO occurs again. Considering the training time, it can be concluded that with low attributes, OvA acts better than OvO.

Table 4.5: Classification Results of SVM Classifier by Using Linear Kernel with and without using PCA- 9 Attributes Multiclass

Data Set	Multi Class Classifier	Correctly Classified Instances	Incorrectly Classified Instances	Time taken (Seconds)	Recall	False measure	Accuracy
9 Attributes (PCA)	**OvA**	**37**	**68**	**0.06**	**0.35**	**0.34**	**35.23%**
	OvO	34	71	0.18	0.32	0.31	32.38%
9 Attributes (Without PCA)	OvA	37	82	0.07	0.30	0.30	30.47%
	OvO	34	71	0.11	0.32	0.32	32.38%

In this case, for reducing the dimensionality of our data, PCA is used before classification. It is very obvious that using PCA shortened the time. In order to reduce computational time, the usage of PCA is suggested. We could see from the classification results of linear SVM, when all data are selected, the classification results are not better than the results of usage of PCA.

Using PCA with Linear kernel is better for classification, one possible reason is that the unimportant features such as noise were discarded while the informative ones kept and linear kernel can detect more instances as a correct classified label in this case.

Table 4.6 shows confusion matrix of linear kernel after using PCA with OvA classifier.

Table 4.6: Confusion Matrix of Linear kernel by using OvA with PCA (9 Attributes, Multiclass)

30 Days	6 Month	12 Month	18 Month	3 Years	5 Years	Classified as
6	1	0	0	2	0	30 Days
2	4	10	1	2	1	6 Month
1	3	16	2	6	2	12 Month
0	0	3	3	6	0	18 Month
1	0	8	3	6	4	3 Years
0	1	2	0	6	2	5 Years

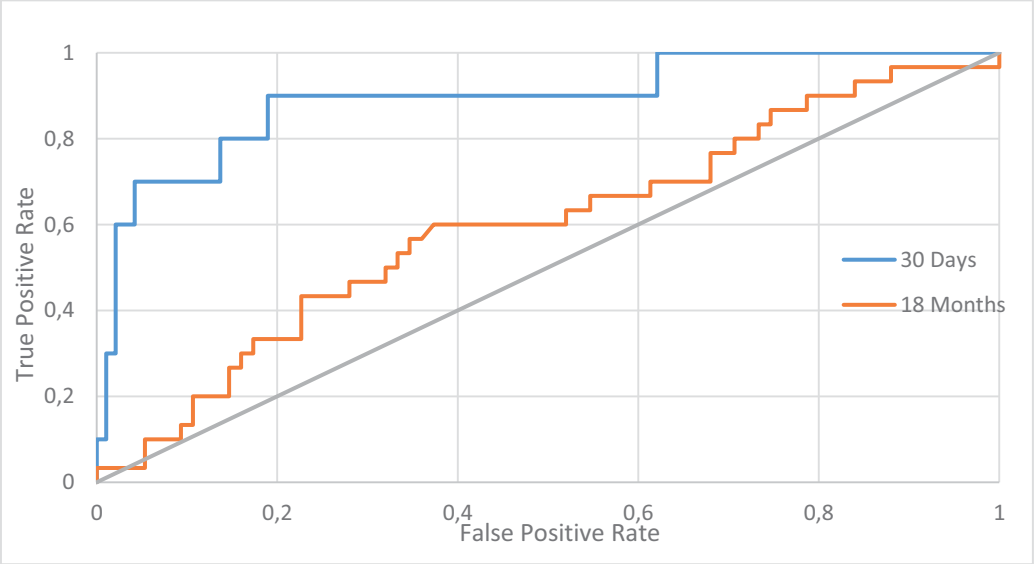

Figure 4.8: Receiver Operating Characteristic curve for SVM Classifier by Using Linear Kernel with PCA for 30 days, 18 months live patients.

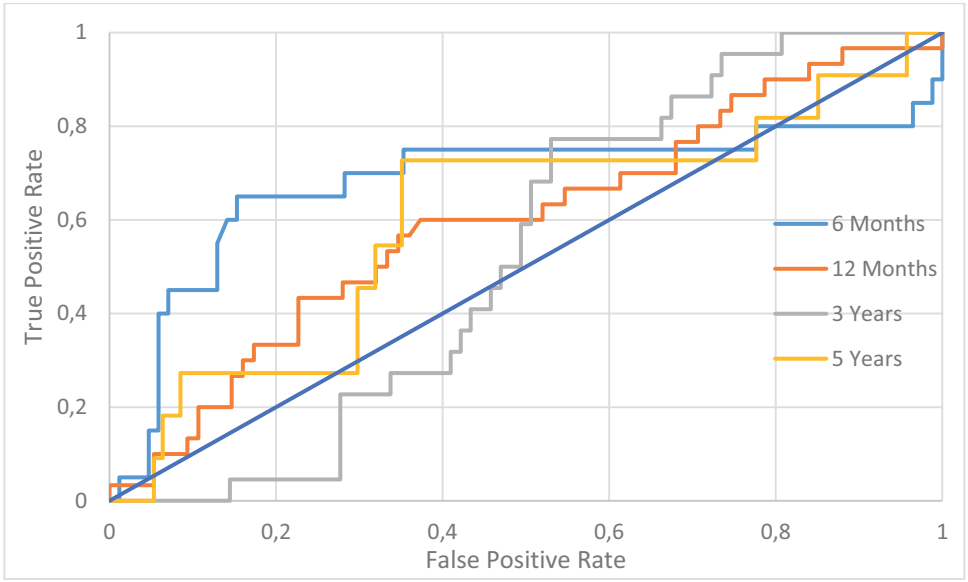

Figure 4.9: Receiver Operating Characteristic curve for SVM Classifier by Using Linear Kernel with PCA for 6 and 12 months and 3, 5 years live patients.

Despite the fact that linear classifier with 9 attributes can classify two classes with highest area under the curve, it is still not the optimal classifier. As we can see from Figure 4.8, for both classes, while true classes are increasing, the false positive cases increase too. For these two classes, linear classifier acts better than other classes. Figure 4.9 shows that five classes are very close to baseline and it means their area under curves are around 50-60 and this classifier has a poor performance.

- **RBF Kernel With and Without PCA with 9 Attributes(Multiclass)**

RBF kernel is used for this model with and without PCA, as we can recognize that in this data set with 9 attributes, all information reduce because of reduction of data into 9 attributes, for this reason the range of accuracy reduces from all data. By using PCA based on our expectation the information should scale down into lesser attributes size; but contrary to expectations, PCA increases the size of attributes to 15 which is not less than 9. This paradox occurs because 9 attributes were selected by the expert and they are highly correlated and it can face the risk of over fitting. Degradation of prediction accuracy with usage of PCA in less attributes is because of this problem. As we can see in Table 4.7 the best model of all observation is RBF without using PCA, with OvA classifier.

Table 4.7: Classification Results of SVM Classifier by Using RBF Kernel with and without using PCA- 9 Attributes Multiclass

Data Set	Multi Class Classifier	Correctly Classified Instances	Incorrectly Classified Instances	Time taken (Seconds)	Recall	False measure	Accuracy
9 Attributes (PCA)	OvA	38	67	0.09	0.36	0.35	36.19%
	OvO	35	70	0.2	0.33	0.32	33.33%
9 Attributes (Without PCA)	**OvA**	**40**	**65**	**0.23**	**0.38**	**0.38**	**38.09%**
	OvO	36	69	0.1	0.33	0.33	34.28%

We have obtained 40 Correctly Classified Instances and 65 Incorrectly Classified Instances. The accuracy of this model is 38.09%. For this kind of data set we recognize that RBF kernel is better than linear kernel. RBF kernel in comparison with linear kernel can be flexible with the data. This flexibility is due to the fact that there are parameters in RBF, which can change the width of kernel while in linear kernel has no extra parameters to change.

Table 4.8 shows confusion matrix for RBF Kernel without using PCA with 9 attributes multiclass and Figure 4.10 and Figure 4.11 are ROC curves of this model. By comparison between linear and RBF kernel we can recognize that RBF can classify more classes with higher area under curves. RBF kernel is more flexible than linear kernel.

Table 4.8: Confusion Matrix of RBF kernel by using OvA without PCA (9 Attributes, Multiclass)

30 Days	6 Month	12 Month	18 Month	3 Years	5 Years	Classıfied as
6	1	1	0	1	1	30 Days
2	7	6	1	2	2	6 Month
0	4	9	5	5	7	12 Month
0	0	2	6	2	2	18 Month
1	0	4	3	10	4	3 Years
0	1	1	0	7	2	5 Years

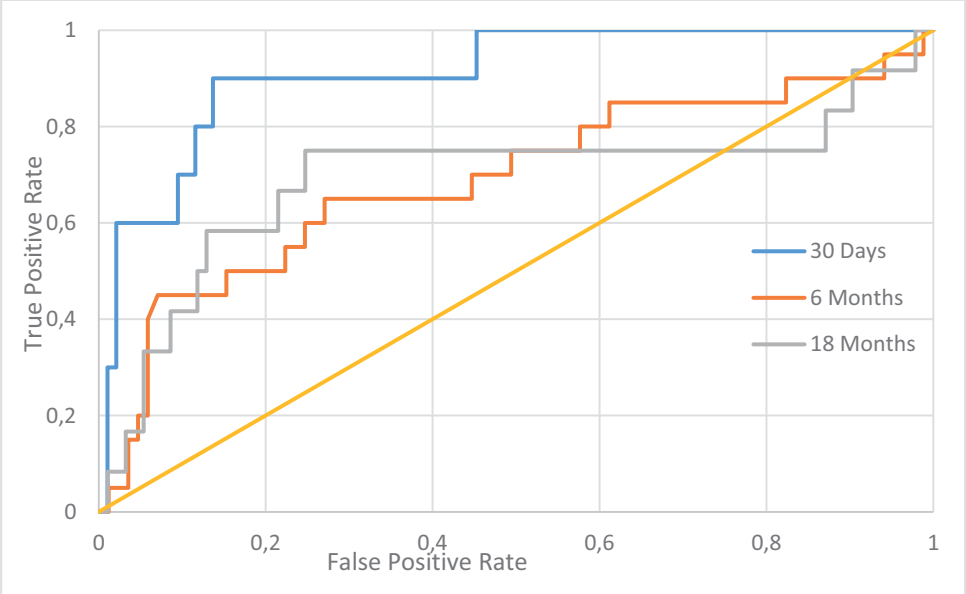

Figure 4.10: Receiver Operating Characteristic curve for SVM Classifier by Using RBF Kernel without PCA for 30 days and 6, 18 months live patients.

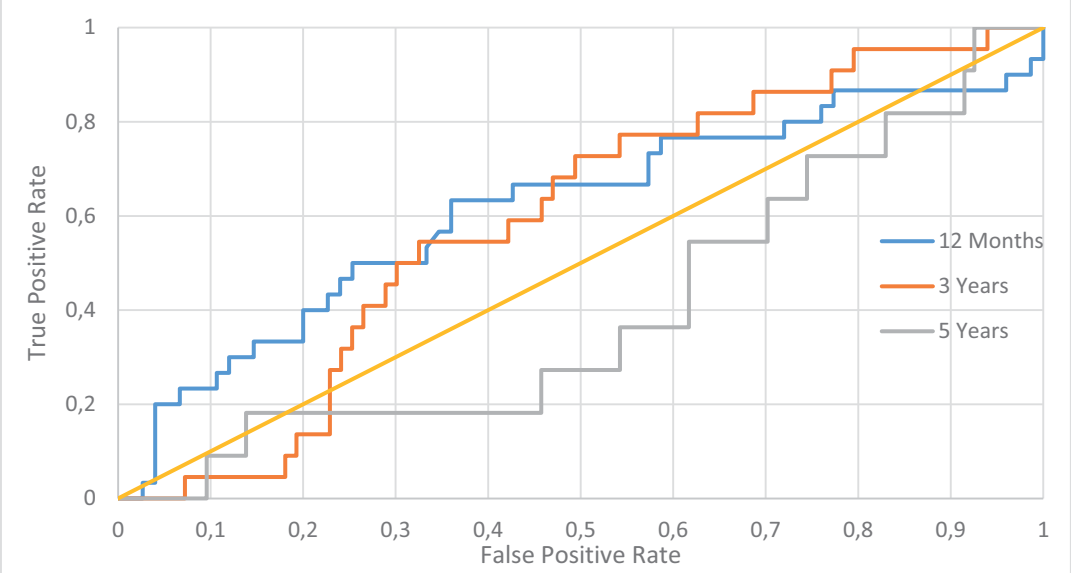

Figure 4.11: Receiver Operating Characteristic curve for SVM Classifier by Using RBF Kernel without PCA for 12 months and 3, 5 years live patients.

Up to now in total observations the average of accuracies are not good. For increase the accuracy of classification we decrease number of labels to two. It means for 5 and 6 labeled as “good” and 1, 2, 3, 4 classes are labeled as “bad”.

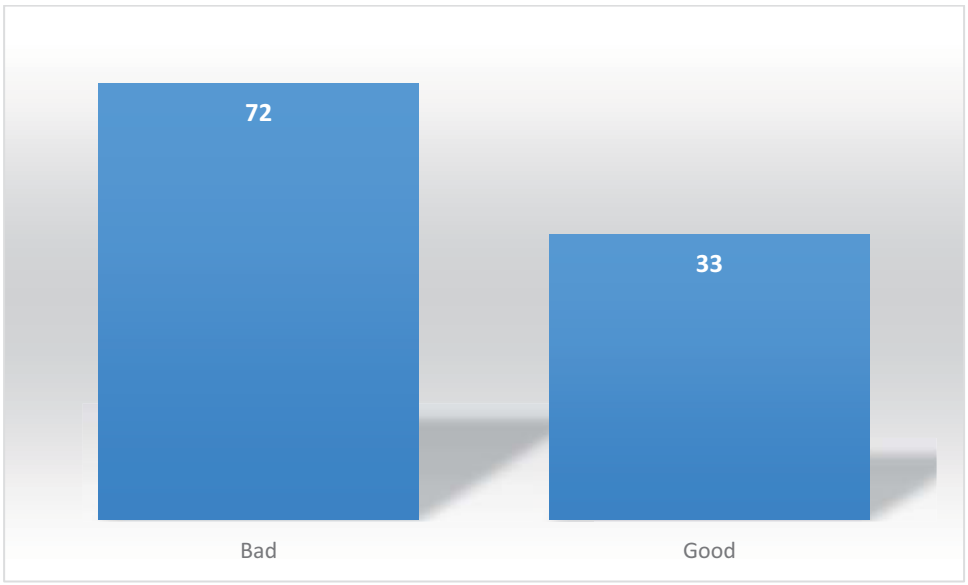

Figure 4.12: Binary Data Set Distributions

- **Linear Kernel With and Without PCA with 9 Attributes(Binary class)**

As a result, by division of all data into two groups through the usage of binary classification ("good and bad") class, SVM classifier classifies the data better. This good results via usage of SVM is because of existence of more data in each group.
Table 4.9 shows results of 9 Attributes with binary class "good and bad". As we can see the accuracy of model increases suddenly and the same amount correct classified instances and incorrect classified instances are increases.
By change Multi class to Binary class, it is normal to have high accuracy because SVM classifier uses linear separator which is optimal for two classes mentioning in chapter 2. SVM can detect two classes easier than six different classes by multi classification classifiers.
As a result, by division of all data into two groups through the usage of binary classification ("good and bad") class, SVM classifier classifies the data better. This good results via usage of SVM is because of existence of more data in each group.

Table 4.9: Classification Results of SVM Classifier by Using Linear Kernel with and without using PCA- 9 Attributes Binary class

Data Set	Correctly Classified Instances	Incorrectly Classified Instances	Time taken (Seconds)	Recall	False measure	Accuracy
9 Attributes Without PCA	75	30	0.18	0.71	0.69	71.42%
9 Attributes PCA	**76**	**29**	**0.01**	**0.72**	**0.71**	**72.38%**

We have obtained 76 correctly classified instances and 29 incorrectly classified instances. The accuracy of this model is 72.38% and it proves that this model is enough good for classification patients based on their period of being alive. There are not significant differences between using PCA or without PCA but the same problem is occur here. Attributes over fitting after applied PCA does not affect results in binary model. In Table 4.10 shows confusion matrix of model and we can easily calculate True Positive and False positive or True Negative and False Negative from below table.

Table 4.10: Confusion Matrix of Linear Kernel with PCA (9 Attributes, Binary class)

Bad	Good	Classified as
62	10	Bad
19	14	Good

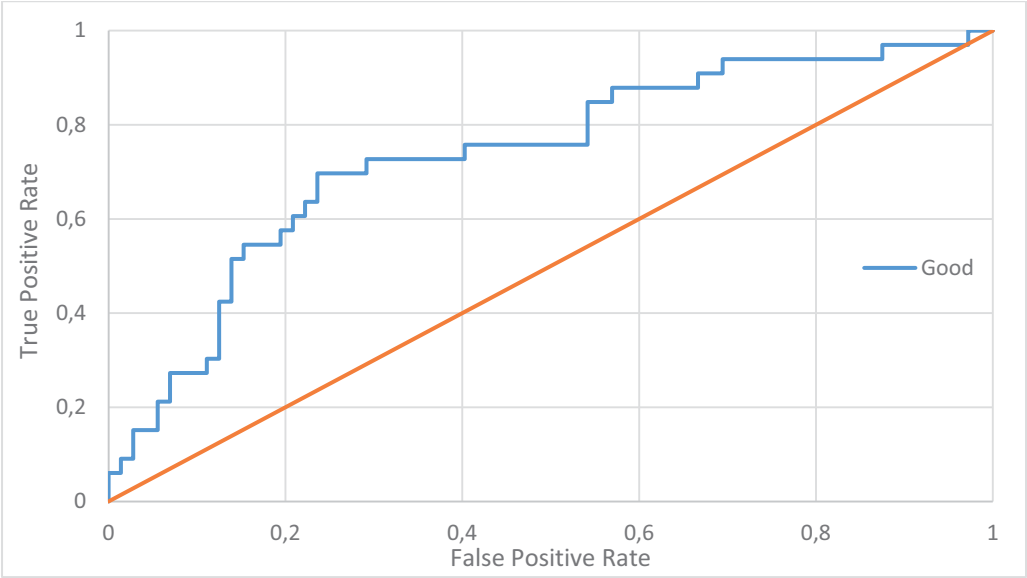

Figure 4.13: Receiver Operating Characteristic curve for SVM Classifier with PCA by Using Linear Kernel

Figure 4.13 represent ROC curve and the area under curve. These curves combine the ROC information from the two classes, results display the test comparing the areas under the two ROC curves. The results indicate that the two areas are not significantly different (AUC= 0.737) and it is good enough for this model.

- **RBF Kernel Without PCA with 9 Attributes(Binary class)**

Radial Bases kernel can work better in this data set because as we mentioned in Figure 4.3 data is nonlinearly distributed and this kernel can maps samples into a higher dimensional space and can handle it when the relation between classes and attributes are nonlinear.

Table 4.11: Classification Results of SVM Classifier by Using RBF Kernel without using PCA- 9 Attributes Binary class

Data Set	Correctly Classified Instances	Incorrectly Classified Instances	Time taken (Seconds)	Recall	False measure	Accuracy
9 Attributes Without PCA	73	32	0.04	0.69	0.67	69.74%
9 Attributes PCA	**76**	**29**	**0.02**	**0.72**	**0.70**	**72.38%**

We have obtained higher classification model accuracy after test binary data set. As we can see from Table 4.11, correctly classified instances are 76 and incorrectly classified instances are 29. The accuracy of this model is 72.38% and we can

understand that this model is good for classification because it is very easy to implement by experts. In Table 4.12 true positive and false positive or true negative and false negative is good factors for analyzing the test. Figure 4.12 represent ROC curve and the area under curve is 0.747 and it is good enough against all result so far. In this situation the best model for classification is using RBF Kernel with or without PCA for 9 Attributes Binary class dataset.

Table 4.12: Confusion Matrix of RBF Kernel with PCA (9 Attributes, Binary class)

Bad	Good	Classified as
63	9	Bad
20	13	Good

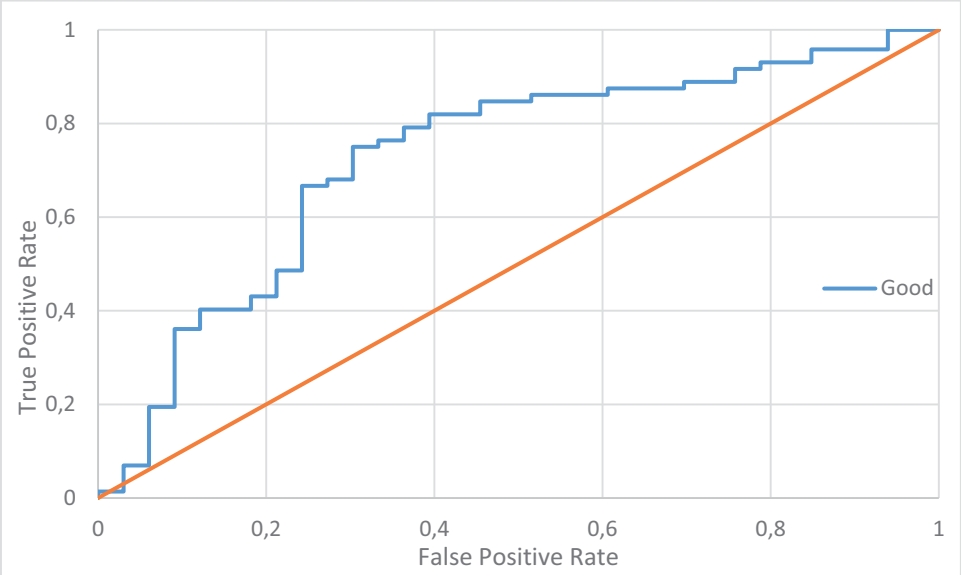

Figure 4.14: Receiver Operating Characteristic curve for SVM Classifier with PCA by Using Linear Kernel

Performance comparison of classification models when the classes are binary is simpler for evaluating. This observation predicts two classes (bad and good), the model is correlated because they occurred form multiple measurements on the same individual. Better prognostic test will be closer to top left corner than poorer test. As values on the X axis true positive rate increase the chance of incorrectly diagnosing negative cases (as positive) increases. In this case the proportion between TP and FP is better than others because when TP is in max range 0.8 the FP rate is in 0.5 in this plot.

Table 4.13: Comparison of the different Kernels and Datasets in SVM

Data Set	Accuracy OvO/OvA	Kernel Type	PCA W/O PCA
Total Data Multi Class	35.23/35.23	Linear	W/O PCA
	33.33/36.19	RBF	W/O PCA
	45.71/39.04	Linear	PCA
	45.71/43.80	**RBF**	**PCA**
9 Attributes Multi Class	32.38/30.47	Linear	W/O PCA
	34.28/38.09	**RBF**	**W/O PCA**
	32.38/35.23	Linear	PCA
	33.33/36.19	RBF	PCA
Total Data Binary Class	73.94	Linear	W/O PCA
	73.94	RBF	W/O PCA
	74.78	**Linear**	**PCA**
	73.10	RBF	PCA
9 Attributes Binary Class	71.42	Linear	W/O PCA
	69.74	RBF	W/O PCA
	72.38	Linear	PCA
	72.38	**RBF**	**PCA**

Table 4.13 shows Comparison of the different kernels and datasets with Support Vector Machine classifier as we use for our first method.

Linear kernel is one method used in SVM to find a linear line for separating two or more classes. Since the Esophagus data may not be linearly separable, we also considered non-linear SVM such as RBF kernel. It mapped the samples' features to the higher dimensional space where non-linear features become linearly separable. There are some reasons for selecting RBF as a best choice for classification:

- First reason is ability of RBF kernel in handling data when there is a non-linear relation between class labels and attributes. In addition, the linear kernel is the special case of RBF kernel.
- As second reason, though RBF has hyper parameter which can affect the complexity of model selection, it is the best choice against linear (the discussed result is shown in Table 4.13).

"One against all" and "One against One" is used to perform multi classification. The accuracy of SVM on this data by using OvO and OvA is largely dependent on number of data and attributes. In all data with multi classes, OvO performed better results because of its structure. This data set has fewer samples in each class; so by using OvO the balance between classes is created. In 9 attributes with multi classes OvA performed better than OvO. In OvO the single parameter selected for calculate decision function, however one parameter set for one decision function may lead to over fitting, so OvO classifier cannot set the best parameter specially in case of 9 attributes that the information of all data decrease.

By testing the Support Vector Machine for all data multi class with and without PCA we can see from results, (PCA+SVM) classification accuracy and (PCA-SVM) classification accuracy are very close together. Also increasing the number of principal components decreases classification performance but it depends on used kernel. From results we can conclude that using (PCA-SVM) is a good choice for our aim for classification system but using PCA for a data set which has more features toward sample size is useful.

We have 119 patients with 42 attributes but for classification problem and for true results we need more data for each class. For example in first data set for each class we have almost 10-20 instances and this range of number is low for doing multi classification. For this purpose we divided classes into two classes.

To summarize, the first method provides the result of linear and RBF kernels of SVM in classification of Esophagus data combined with before and after using PCA with multi classifiers. Both linear and RBF, SVM performed good classification accuracy under appropriate feature selection. Based on the RBF kernel properties we can recognize that the best choice is RBF kernel after using PCA with 9 attributes and binary class data set. Since the highest accuracy in Figure 4.13 belongs to liner kernel with 9 attributes and binary class in all data set and with having higher dimension, linear kernel can separate easily but if in all data one of the data get lost or there were missing data we cannot use this model. By decreasing the number of attributes we also decrease the number of missing data. In addition to missing data problem in all data there is over fitting problem that generally occurs when a model is more than enough complex, such as having too many parameters or attributes.

4.3. Results of Bayesian Network

In previous chapters we explained the Bayesian network methods and dataset which is for our Bayesian network model. On Figure 3.1 we try to show the general view of the network, in this section the results are shown and we discuss about the output of network in detail.

The general methodology of Bayesian networks is described in chapter 2, Bayesian networks works on relationships of conditional probability and dependence between its variables [47, 48]. All the data preprocessing and performance analysis was done with SamIam and Weka. All continuous variables which were represented in the network were discretized into equal range of numbers. We have 12 nodes and each node has its own conditional probability table. To begin with, for each feature like Age, we try to estimate the probability of distribution for per interval for example in range 25-45. For each node of network we try to calculate probability and all these probability is collected into node's properties we can change and arrange nodes properties based on their conditionally (dependent and independent) probabilities and their relationships. Vomit node probability is change based on the nodes which they are related to vomit and it changes when related nodes change. The related nodes as it shown in Figure 4.15 are Nausea and Age.

All these calculations are done for all nodes based on the relations in their networks. For better analyzing we create three kinds of networks to see the accuracy with more complicated networks. Figure 4.14, 4.15 and 4.16 are three different networks. First Network is created based on correlation and expert opinion, for second Network we try to decrease number of edges. This helps us to better understanding the Network and it has less calculation. Third Network is just based on correlation between nodes and it is complicated than two other networks.

We employed a five-fold cross-validated experimental design for train and test set. We have 119 data and spilt them into 5 sets. Each set contains train and test set which is used for performance analysis. In these sets the order of features are looks like the networks nodes order. Model accuracy was determined using the averaging of results of this stage which was obtained by using the Python programming language for tested sets for five times and the general accuracy is network accuracy and network performance. Table 4.14 shows the results.

Table 4.14: Comparison of the different Networks' Accuracies with SamIam

Training and Test Set	Network 1	Network 2	Network 3	Network 4
Average	12.64	19.34	11.81	15.94

As we have seen in chapter 2, common accuracy amounts are more than 85% but our results here do not seem to be reasonable. The problem may be because of network structure or lack of data. For solving these problems we have referred to an expert for improving the network and request for more data and we have decided to omit some extra nodes because as it can be seen that the network which is less complicated has higher accuracy rate. So “Nausea”, “Dysphagia”, “Grade” are eliminated because features such as “Vomit”, “Odynophagia”, “Stage” are respectively used instead of them. As a result the size of the network is reduced from 12 to 9 nodes.

Based on the expert’s advice we have changed the connection of nodes in this way: all of the nodes are connected with “Prognosis” because all other features affect “Prognosis” in some way.

The network is drawn in SamIam and CPTs are calculated again with new conditional probability. In this way we have encountered with three kinds of problems:

1. Entering the network in SamIam made software slow
2. Calculation of CPTs appeared to be complicated
3. Data does not seem to be sufficient

In order to deal with first and second problems the structure of expert’s network has changed a little, figure 4.16 indicates final version of our network.

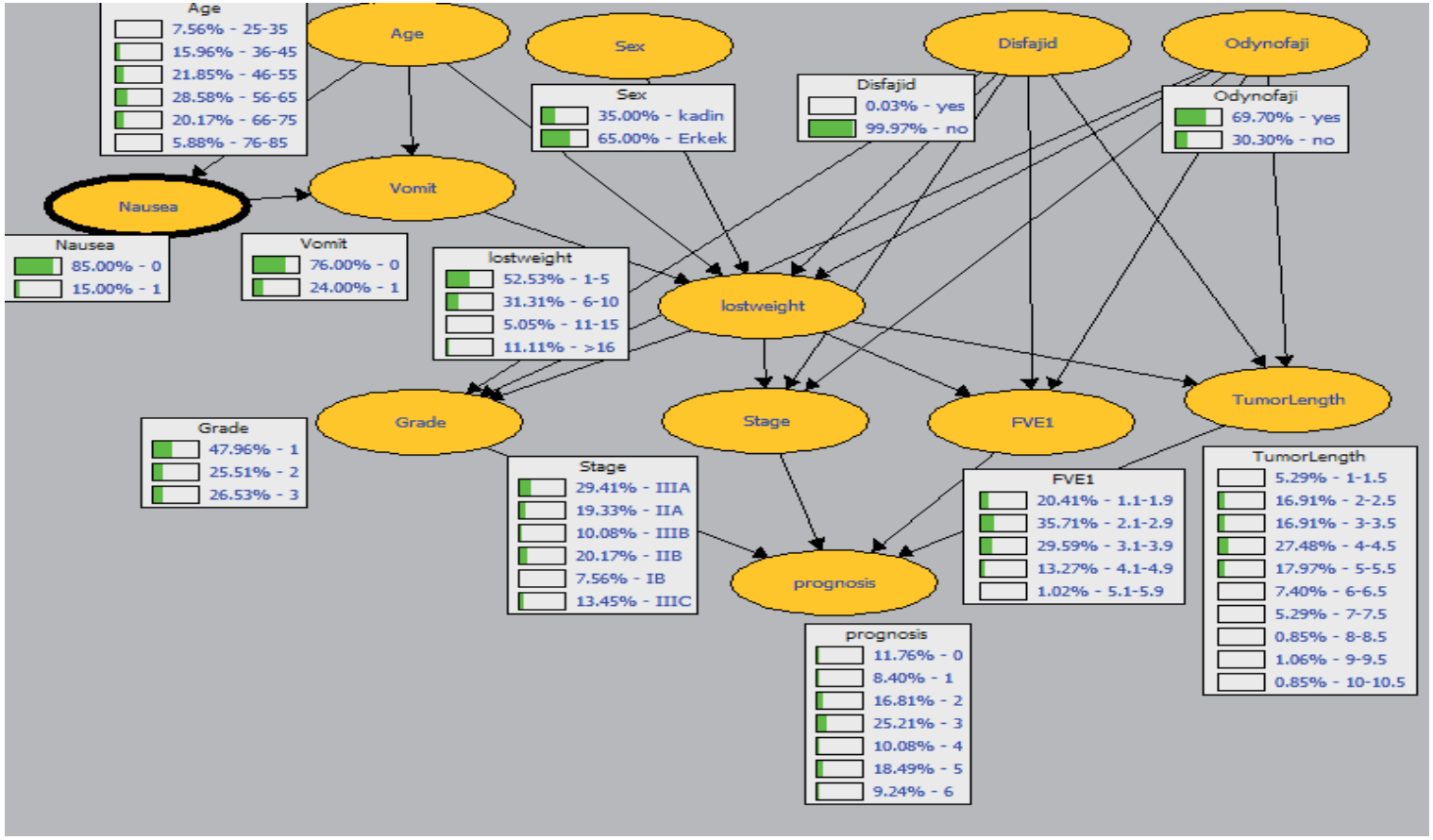

Figure 4.15: First Bayesian Network and Conditional Probability Tables in Esophagus Networks

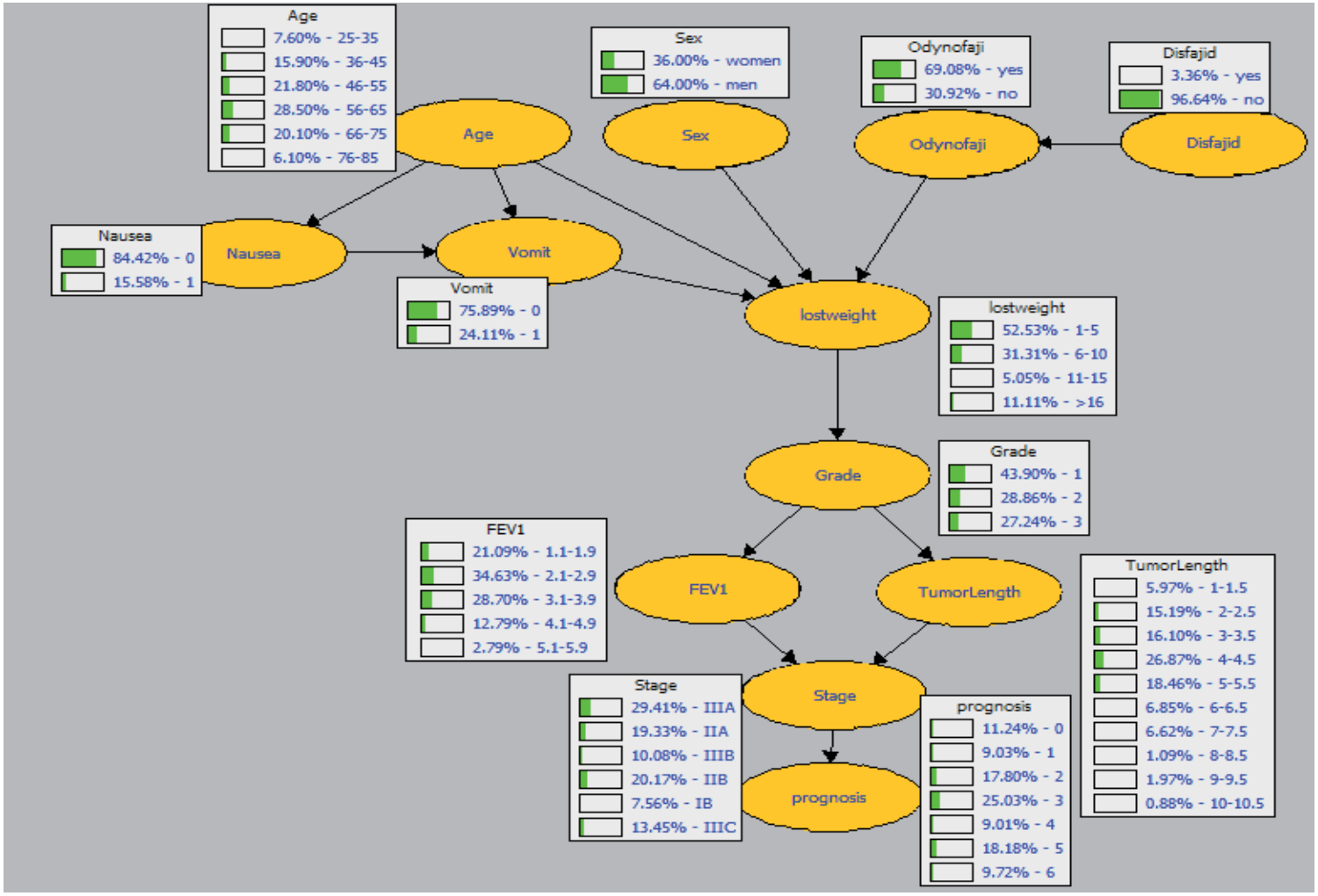

Figure 4.16: Second Bayesian Networks and Conditional Probability Tables in Esophagus Networks

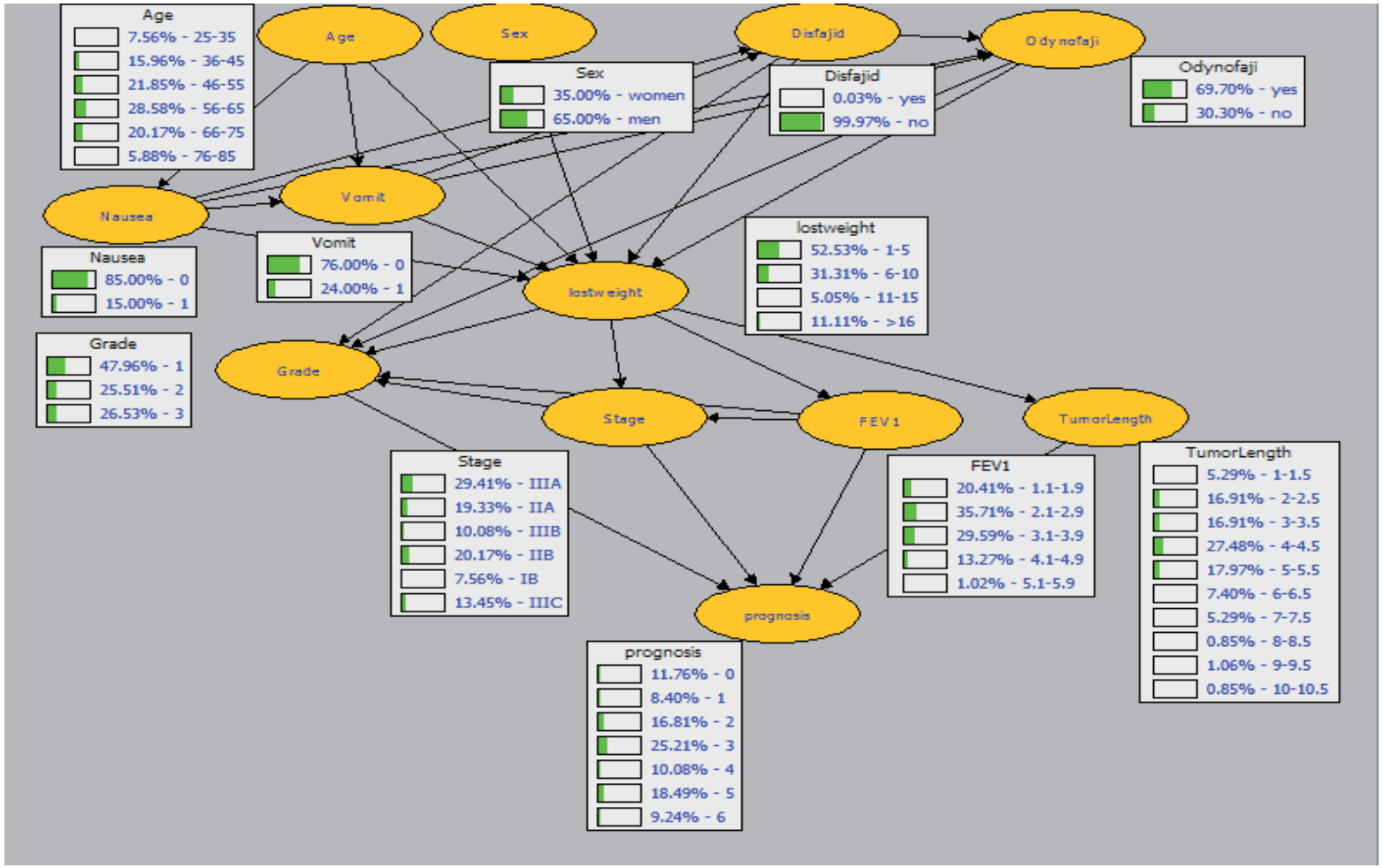

Figure 4.17: Third Bayesian Networks and Conditional Probability Tables in Esophagus Networks

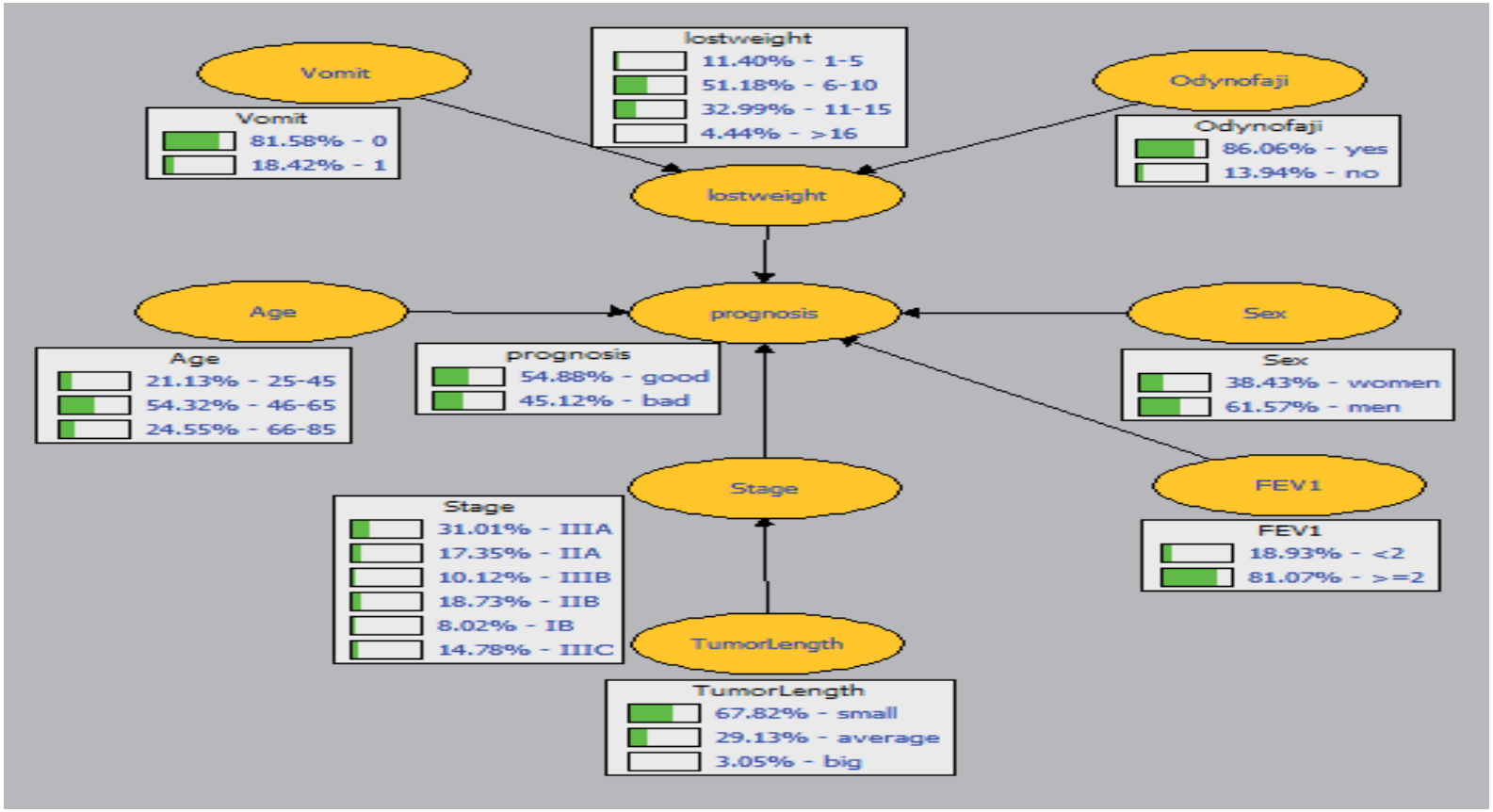

Figure 4.18: Final Bayesian Network schema and Conditional Probability Tables in Esophagus Networks

Esophagus Networks

Weka can also show statistics and class distribution over the values of each attribute area like Figure 4.19. The screenshot below shows these diagrams.

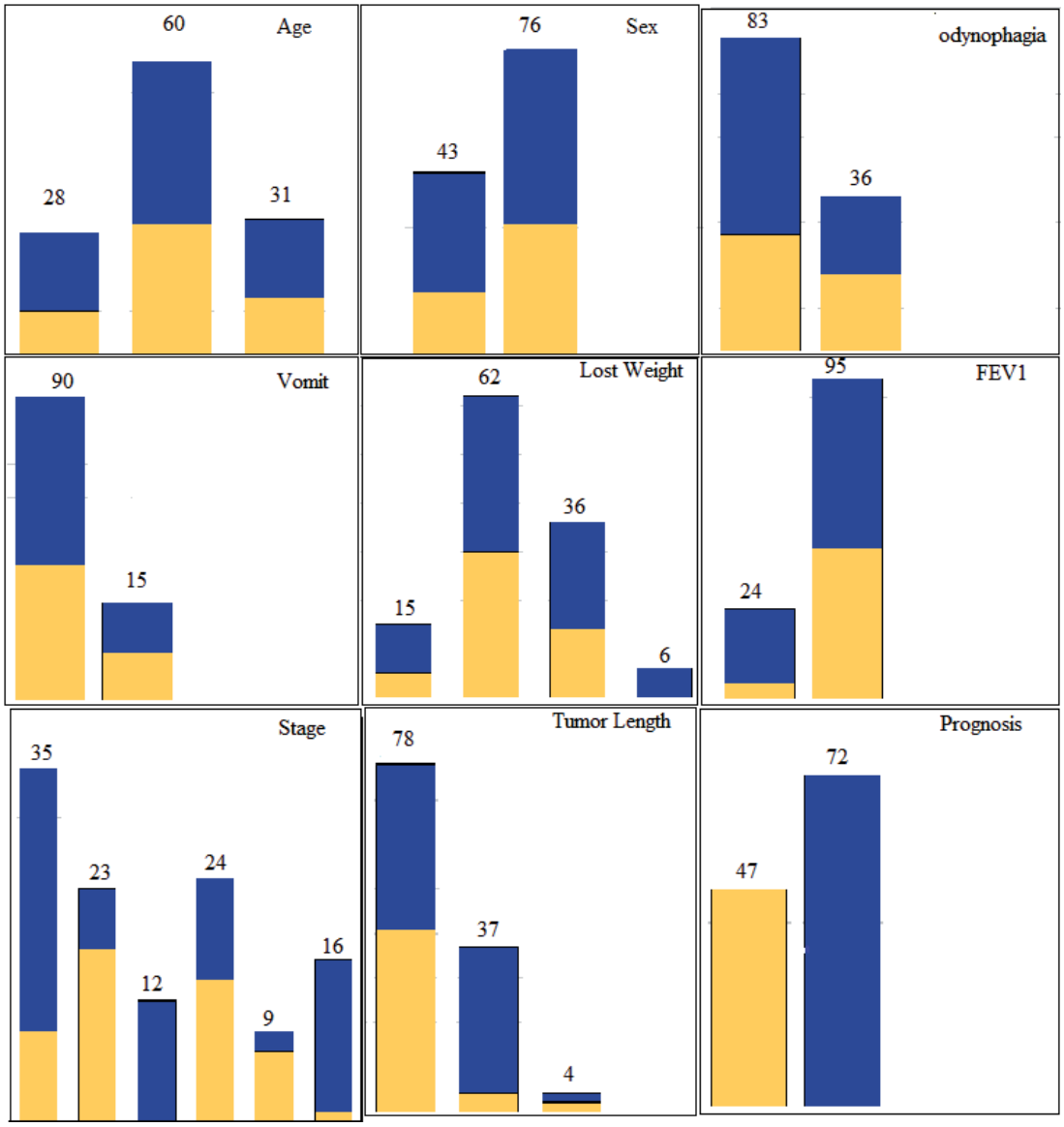

Figure 4.19: Class Distributions over the Values of Each Attribute Area

- **First Bayesian Network Model**

Figure 4.20 is shown graphical representation of the first network, which is created based on correlation between attributes. Running this network with 10-fold cross validation produces the following output:

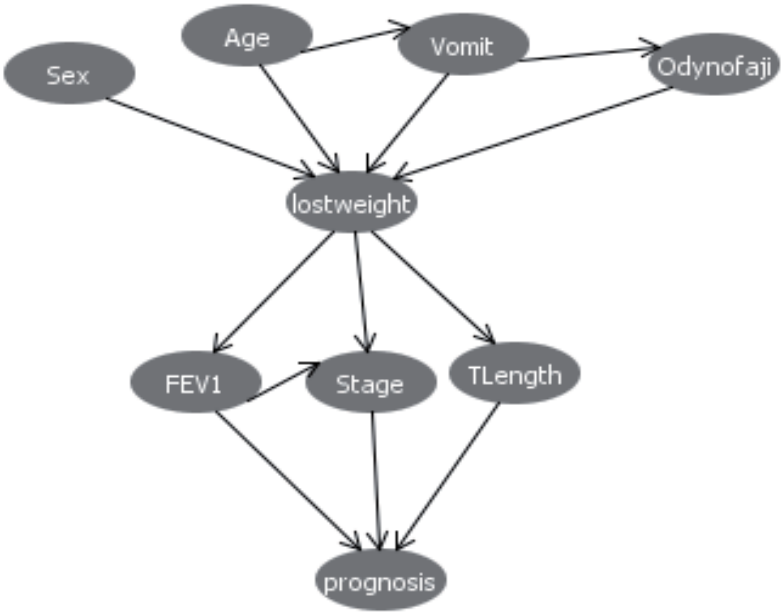

Figure 4.20: Graphical Representation of First Network Model

Table 4.15 shows this Network has 87 correctly classified instances and 32 incorrectly instances.

Table 4.16 confusion matrix indicates that 22 instances from class "B" are wrongly classified as class "A" and 10 instances from class "A" are wrongly classified as class "B". The degree of accuracy of a Bayesian network, with related to a data set, depends on how well the Bayesian network, and is determined by how well the joint probability distribution represented by the Bayesian network matches the joint probability distribution described by the given data set. In this network due to relation between attributes, accuracy is high.

Table 4.15: Classification Results of Bayes Net Classifier by Using First Network-

9 Attributes Binary class

BAYES NET	Correctly Classified Instances	Incorrectly Classified Instances	Time taken (Seconds)	Recall	False measure	Accuracy
First Network	87	32	0.01	0.731	0.734	73.10%

Table 4.16: Confusion Matrix of First Network

Good	Bad	<= Classified as
37	10	Good
22	50	Bad

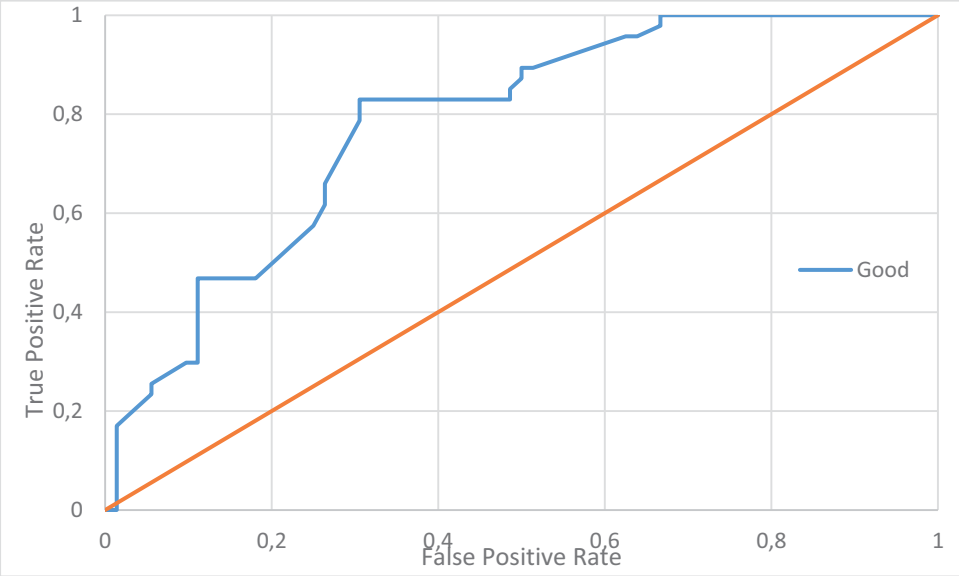

Figure 4.21: Receiver Operating Characteristic curve for Bayes Net Classifier by Using First Network

The Figure 4.21 shows ROC curve representing good test plotted on the graph. The accuracy of the test depends on how well the test separates the group. Accuracy of this model that is measured by the area under the ROC curve represents a good performance based on the arrangement that it mentioned before.

- **Second Bayesian Network Model**

Figure 4.22 is shown graphical representation of second network which is suggested by expert but based on the first problem which is lack of data resulting unwanted random outcomes, we try to change some edges. The outputs with 10-fold cross validation are as follows:

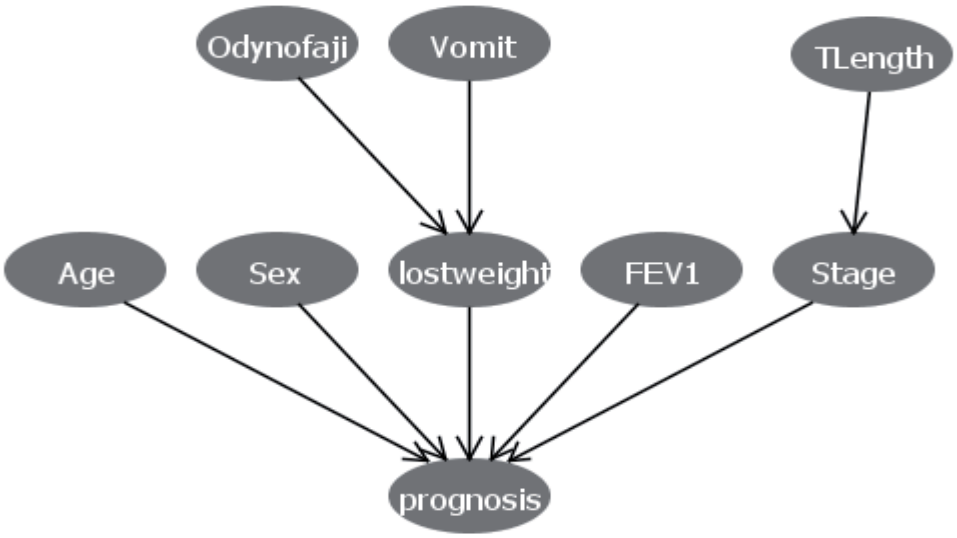

Figure 4.22: Graphical Representation of Second Network Model

Table 4.17 shows this Network has 53 correctly classified instances and 66 incorrect instances. The second network accuracy is 44.53% and it is better than first network because we decrease number of edges. This solution effect the accuracy of network but is still low.

Table 4.17: Classification Results of Bayes Net Classifier by Using Second Network- 9 Attributes Binary class

BAYES NET	Correctly Classified Instances	Incorrectly Classified Instances	Time taken (Seconds)	Recall	False measure	Accuracy
Second Network	53	66	0.02	0.445	0.421	44.53%

As we can see from Table 4.18 confusion matrix indicates that 34 instances from class B and 19 instances from class A are classified correctly but 13 instances are wrongly classified as class B and 53 instances are wrongly classified as class A. Again correctly classified numbers are low and we have to improve the network by change the edges.

Table 4.18: Confusion Matrix of Second Network

Good	Bad	<= Classified as
34	13	Good
53	19	Bad

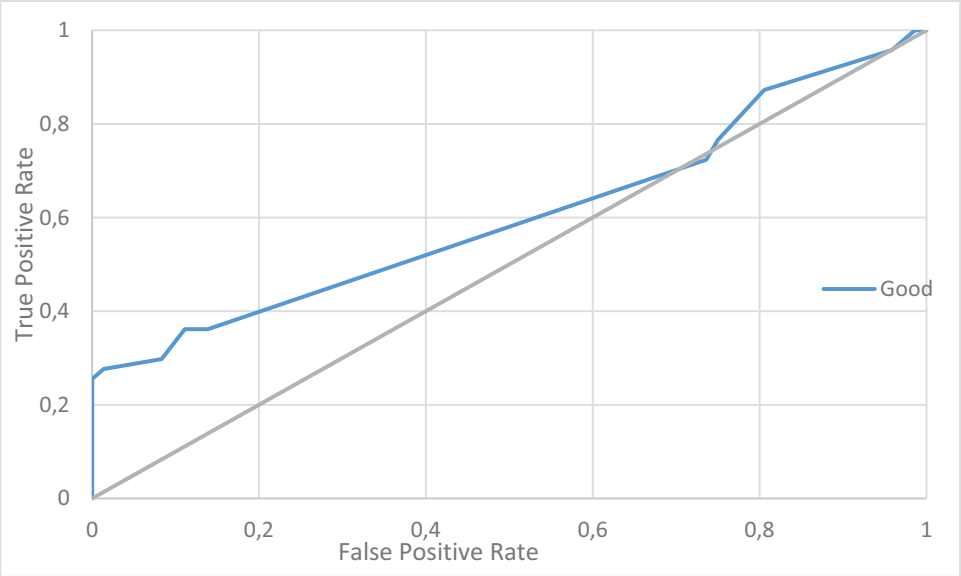

Figure 4.23: Receiver Operating Characteristic curve for Bayes Net Classifier by Using Second Network

Figure 4.23 can be shown that the area under curve is equal to the probability that the model will correctly identify the positive case when presented with a randomly chosen pair of cases in which one case is positive and one case is negative. This model is poor for classification problems because AUC for this model is very low and it is very close to the baseline.

- **Third Bayesian Network Model**

Figure 4.24 is shown third network which is tested for showing that some networks works better than others but it is not enough to accept as an optimal network. Here, network's structure is the same network with 9 nodes. Running this network with 10-fold cross validation produces the following output:

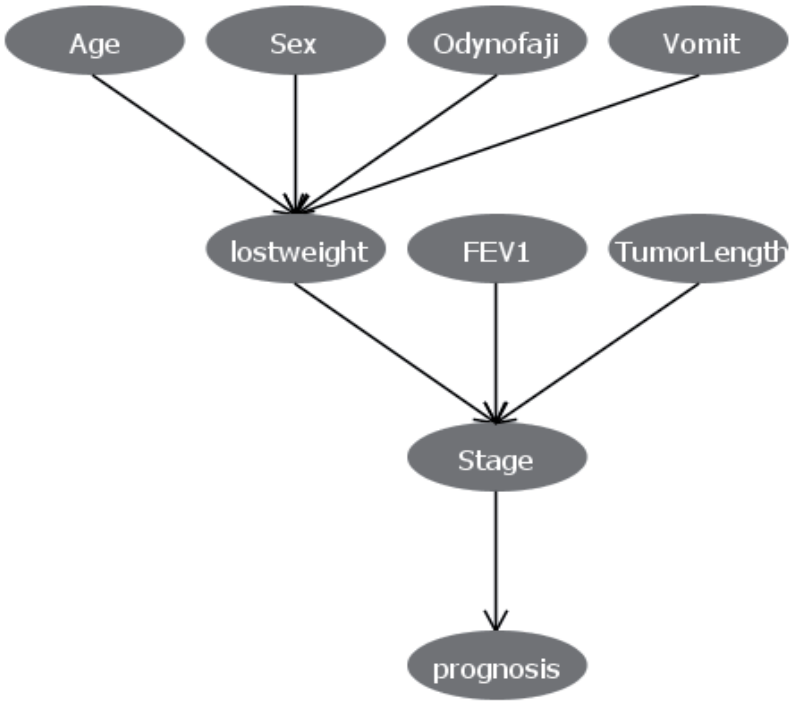

Figure 4.24: Graphical Representation of Third Network Model

Table 4.19 shows the best results over other networks. 83 correctly classified instances with 34 incorrectly classified instances and the result shows that this model of Bayesian network reach to 71.42% accuracy. As we can see

Table 4.20 True Positive and True Negative number is high for this model. Figure 4.25 ROC curve representation is better than other models because of False Positive and True Positive Rate. However the accuracy, there is a problem with network's structure. In this structure all the nodes connect to the Stage node and this means Stage decide instead of all nodes and Stage is the only node that it affects prognosis and it is independent from all other nodes. Obviously this kind of structure with the highest accuracy is meaningless and useless for Classification but it is useful for clinical approaches when the staging level is not complete and doctors have to decide without having stage information of patients.

Table 4.19: Classification Results of Bayes Net Classifier by Using Third Network-

9 Attributes Binary class

BAYES NET	Correctly Classified Instances	Incorrectly Classified Instances	Time taken (Seconds)	Recall	False measure	Accuracy
9 Attributes (Binary class)	85	34	0.07	0.717	0.744	71.42%

Table 4.20: Confusion Matrix of Third Network

Good	Bad	<= Classified as
33	14	Good
20	52	Bad

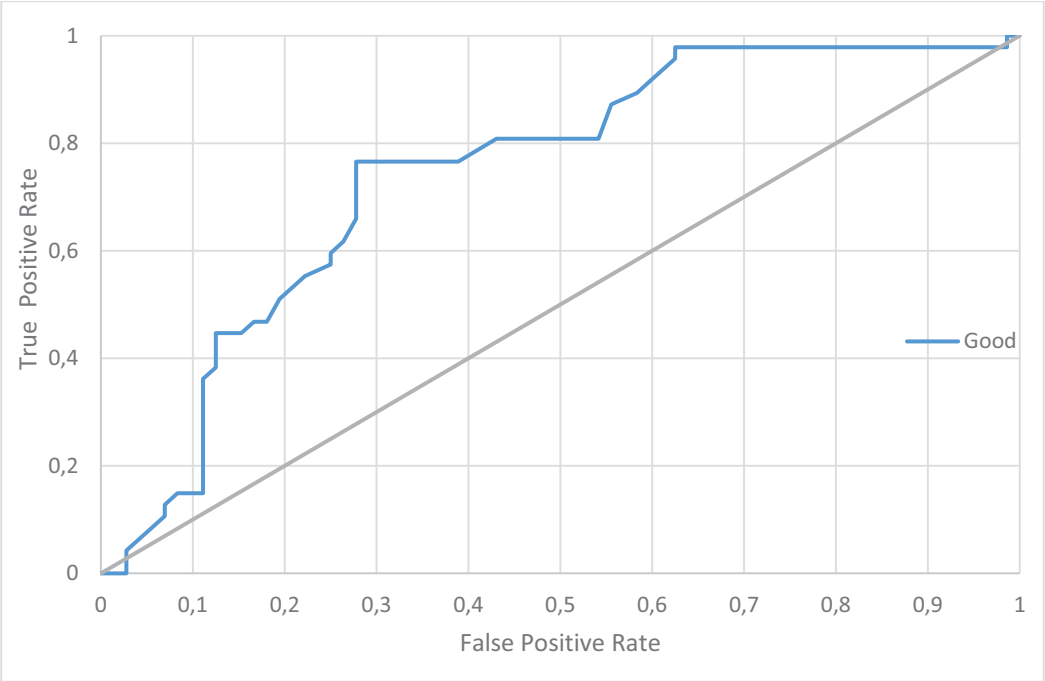

Figure 4.25: Receiver Operating Characteristic curve for Bayes Net Classifier by Using Third Network

As an overall summary of diagnostic accuracy, the AUC can be discuss. When the ROC curve corresponds to perfect accuracy the AUC equals 1 and 0.5 for random chance. Rarely the evaluated AUC is less than 0.5. This estimated among AUC explains the test does worse than chance. As we can see here the AUC of Good class is higher than 0.5 and this model can be select as good model for classification in special case when we have not enough information from some attributes.

- **Fourth Bayesian Network Model**

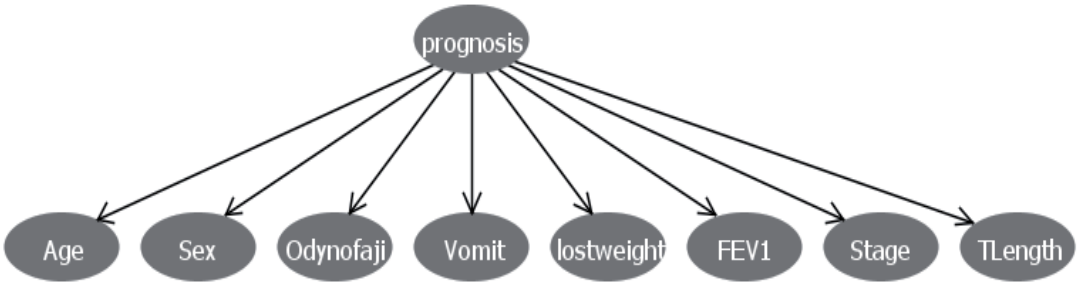

Figure 4.26: Graphical Representation of Forth Network Model

Figure 4.26 is shown graphical representation of Fourth network which it is Naïve Bayes. Here, attributes of class are independent so create Naïve Bayes network is logical and this structure is near to first Network which is suggested by expert. The accuracy that Table 4.21 shows is 67.22% that it is high from first network and confusion matrix in Table 4.22 shows that True Positive and True Negative number is higher than first network. The ROC curve of this model as we can see in figure 4.26 is better than other Networks because of False Positive and True Positive Rate.

Comparison between Third and Fourth network shows us, third network has higher accuracy than forth network but for doctors the structure of network is more important because the one attribute, for example "Age" singly influence "Prognosis" in this type of cancer, but in Third network "Age" just influence "LostWeight" and based on the expert opinion Age is one of the effective factor for prognosis. We can say the same things for all other nodes so we can accept Forth network as our final model.

Table 4.21: Classification Results of Bayes Net Classifier by Using Third Network-9 Attributes Binary class

BAYES NET	Correctly Classified Instances	Incorrectly Classified Instances	Time taken (Seconds)	Recall	False measure	Accuracy
9 Attributes (Binary class)	80	39	0.03	0.672	0.676	67.22%

Table 4.22: Confusion Matrix of Forth Network

Good	Bad	<= Classified as
32	15	Good
24	48	Bad

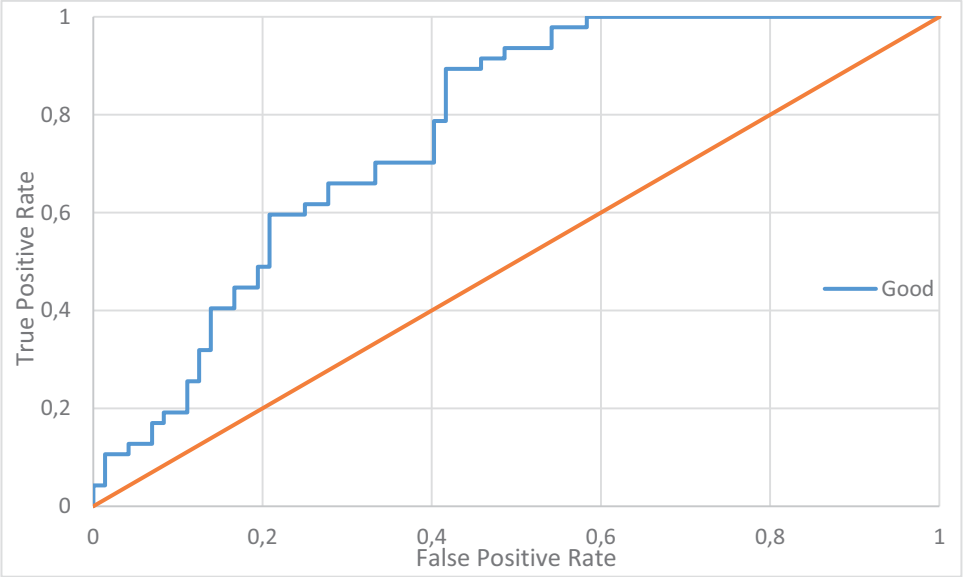

Figure 4.27: Receiver Operating Characteristic curve for Bayes Net Classifier by Using Forth Network

Selection the optimal threshold under a diversity of clinical situations, can be done through ROC analysis which could balance the inherent tradeoffs among sensitivity and specificity. In this case we can see from Figure 4.27 that this model act as an optimal case for clinical usage. The area of Good class is higher than baseline and the highest point of good class occurs when false positive is equal to 0.4.

Figure 4.28 shows the comparison of four networks. For the analysis of discussed Networks the Esophagus data set has been used with 119 instances and 9 attributes. From Figure 4.28, it is clear that accuracy of first network is the highest but the analysis of another parameters and network structure formed by Naïve Bayes is the best among these. It is also seen that accuracy of Naïve Bayes is better than others when the dataset is small.

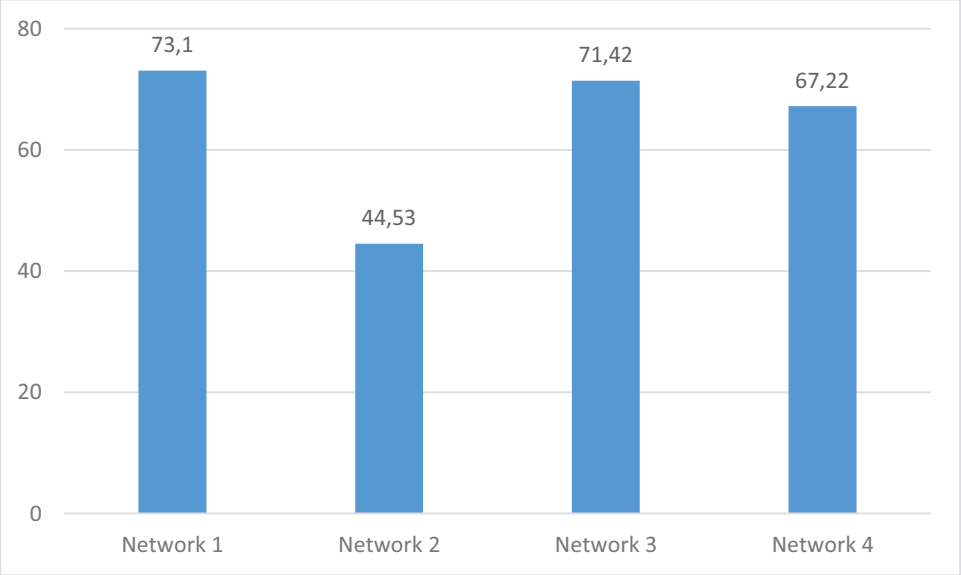

Figure 4.28: Comparison of Networks Accuracy

In this thesis we presented Support Vector Machine and Bayesian network classification methods developed for medical decision making problems to find the most appropriate structure of the model. Comparing the first and second approach shows that the results are comparable in terms of performance.

Firstly, it can be observed that the Bayesian networks performed better than Linear Support Vector Machine but Support Vector Machine with RBF kernel obtain higher prediction accuracy than Bayesian network. It has to be mentioned, that although better accuracy is gained on SVM, it doesn't not allow interpretation of prediction, while BNs do, and interpretation in fourth network constructed as above, can be done efficiently.

Subsequently, increasing the number of data doesn't effect on the SVM performance because it just creates boundaries among data regarding their classes, but while Bayesian networks are involved with probability of data so they demonstrate better accuracy when the number of data is increased.

The most striking advantage of Bayesian networks in comparison with SVM is that Bayesian networks can continue to classify even if there is missing data is occurred, while the standard formulation of Support Vector Machine does not allow for missing values for any attributes to classified for solving this problem there are some preprocessing methods to prepare data sets for use by an SVM.

CHAPTER 5

5. CONCLUSION AND FUTURE STUDIES

5.1. Conclusion

This study analyzed clinical pathological data of patients with esophageal carcinoma who underwent esophagectomy and investigated prognostic factors on mortality and survival.

In this thesis, we tested two classification methods for the prognostic prediction, the objective of this study is to develop and analyze the performance of a number of classification models to classify patients. Our data set was collected at the "Ankara Numune Teaching and Research Hospital" between 2003 and 2011. Data contains 119 patients with preoperative clinical factors and other attributes including age, sex, type and duration of symptoms and post-operative prognostic results for all patients. The Support Vector Machine method was used to analyze prognosis and classify the patients based on their life interval. Weka was used to compare the models and confusion matrices and Receiver Operating Characteristic (ROC) curves were used for statistical analysis.

By our observations, Support Vector Machine has shown %70.58 accuracy with using "RBF kernel" with using Principal component analysis for data reduction. We obtain from comparing result with or without PCA that preprocessing the features using PCA did not significantly increase classification accuracy of the SVM. Support Vector Machine can be helpful for obscure analysis when data are irregularly distributed and produce very accurate classifiers.

The prognostic prediction model will help to enhanced clinical practice for many complex diseases and other problems. Converting this knowledge into daily applied diagnostics will be a challenge in health domain.

For this purpose we decided to use Bayesian Networks for prognostic prediction, Because Bayesian Networks are easy to use and understand. Bayesian Networks with decision making tasks can help physicians and other health professionals.

The main objective of our study was to predict life span of the patients with the esophageal carcinoma that underwent curative surgery without preoperative chemotherapy or radiotherapy.

We identified a total of 119 patients who underwent esophagectomy and obtained 73.10% accuracy.

As for a summary of results it has to be mentioned that the Bayesian networks are better than Support vector machine for prognosis cases and for doctors in clinical approaches because it is easy to understand how network and classifier works.

We evaluated the effectiveness of dimension reduction and normal distribution of real data in improving the classification accuracy. The dimension reduction method significantly improved classification accuracy of the Bayesian network and SVM. The accuracy of the two classifiers after applying PCA was very similar, with no statistical differences in the area under the ROC curve.

One major disadvantage of SVM method over Bayesian Networks is missing data, missing data is a given in the medical domain, so machine learning models should have satisfactory performance even when missing data occurs. Bayesian networks (BN) can handle missing data better than support vector machines (SVM). To test the hypothesis, we trained a BN and SVM model for 119 patients and compared their accuracy in four separate datasets which are 42 and 9 attributes (Multiclass and Binary class). Bayesian networks can continue to classify even if missing data is occurred, while SVM don't continue to work.

5.2. Future Work

In this study we focus on Bayesian Networks in post-operative prognostic prediction in future research can be focus on Bayesian Networks in pre-operative prediction. By incorporating both pre-operative and post-operative information we can obtain useful results that it can help to doctors for getting correct decisions and may help improve power of prediction.

We have carefully validated performance from the literature and characterized which predictors are important and the results shows us Decision tree algorithm and Artificial Neural Network approaches are also useful in data mining. For using these algorithms for our data we have some limitation such as, number of data. We need more data to obtain reliable result so Future works have to focus on improving the BN performance by including more patients, more variables, and more diversity in Networks models, besides another option for improving our work can be finding more effective features.

We can examine methods for handling data sets with missing features in SVM. Some of these methods are simple such as ignoring the missing data by discarding examples with a missing attribute value or discarding an attribute that has missing values. A second possible solution is the *imputation* method by which a value is generated for the attribute. There are more robust methods that are likely to supply a value closer to the one that is missing: for example, a K-Nearest neighbors (KNN) approach or replacing missing data with mean and mode values are methods that we can use for handling missing data. Another technique is the use of a separate SVM to determine the likely value. Our future work focus on improving SVM accuracy by using above approaches for handling missing data.

6. REFERENCES

[1] Lavrač, Nada. "Selected techniques for data mining in medicine." *Artificial intelligence in medicine* 16, no. 1 (1999): 3-23.

[2] French, Nick, and Simon French. "Decision theory and real estate investment."*Journal of Property Valuation and Investment* 15, no. 3 (1997): 226-232.

[3] Lucas, Peter. "Logic engineering in medicine." *Knowledge Engineering Review*10 (1995): 153-180.

[4] Cooper, Gregory F. "The computational complexity of probabilistic inference using Bayesian belief networks." *Artificial intelligence* 42, no. 2 (1990): 393-405.

[5] Aha, David W., Dennis Kibler, and Marc K. Albert. "Instance-based learning algorithms." *Machine learning* 6, no. 1 (1991): 37-66.

[6] Denoeux, Thierry. "A k-nearest neighbor classification rule based on Dempster-Shafer theory." *Systems, Man and Cybernetics, IEEE Transactions on* 25, no. 5 (1995): 804-813.

[7] Specht, Donald F. "Probabilistic neural networks." *Neural networks* 3, no. 1 (1990): 109-118.

[8] Quinlan, J. Ross. "Induction of decision trees." *Machine learning* 1, no. 1 (1986): 81-106.

[9] Bratko, Ivan, and Igor Kononenko. "Learning diagnostic rules from incomplete and noisy data." *Interactions in Artificial Intelligence and Statistical Methods* (1987): 142-153.

[10] Rice, Thomas W., Eugene H. Blackstone, and Valerie W. Rusch. "of the AJCC Cancer Staging Manual: esophagus and esophagogastric junction." *Annals of surgical oncology* 17, no. 7 (2010): 1721-1724.

[11] Schnall, S. F. "Carcinoma of the esophagus." *Current Opinion in Gastroenterology* 5, no. 4 (1989): 542-548.

[12] Mannell, Aylwyn. "Carcinoma of the esophagus." *Current problems in surgery*19, no. 10 (1982): 556-647.

[13] Perez, Carlos A., Srinivasan Vijayakumar, J. D. Bradley, and S. Mutic. "Carcinoma of the Esophagus." *Technical Basis of Radiation Therapy: Practical Clinical Applications* (2006): 511-524.

[14] Boyd, David P., Herbert D. Adams, and Ferdinand A. Salzman. "Carcinoma of the esophagus." *New England Journal of Medicine* 258, no. 6 (1958): 271-274

[15] Ellis Jr, F. Henry, Robert C. Jackson, Julius T. Krueger Jr, Herman J. Moersch, O. Theron Clagett, and Robert P. Gage. "Carcinoma of the esophagus and cardia: results of treatment, 1946 to 1956." *New England Journal of Medicine*260, no. 8 (1959): 351-358.

[16] Boo-Chai, Khoo. "Carcinoma of the esophagus: An epidemiological study."*Plastic and Reconstructive Surgery* 66, no. 5 (1980): 799.

[17] Sherman, Julius. "Carcinoma of the Esophagus: A Report of Fifty Cases." *The American Journal of the Medical Sciences* 175, no. 1 (1928): 79-83.

[18] Mimi, C. Yu, David H. Garabrant, John M. Peters, and Thomas M. Mack. "Tobacco, alcohol, diet, occupation, and carcinoma of the esophagus." *Cancer research* 48, no. 13 (1988): 3843-3848.

[19] Sadjadi, Alireza, Reza Malekzadeh, Mohammad H. Derakhshan, Alireza Sepehr, Mehdi Nouraie, Masoud Sotoudeh, Abbas Yazdanbod et al. "Cancer occurrence in Ardabil: Results of a population-based Cancer Registry from Iran."*International journal of cancer* 107, no. 1 (2003): 113-118.

[20] Carolyn, Anderson. UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN, Accessed January 21, 2014.

http://courses.education.illinois.edu/EdPsy584/lectures/PrincipalComponents-online.pdf.

[21] FISHER, R. A., 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, 111–132.

[22] VAPNIK, V., and A. CHERVONENKIS, 1964. A note on one class of perceptrons. *Automation and Remote Control*, **25**

[23] BOSWELL, DUSTIN. "Introduction to support vector machines." (2002).

[24] MADZAROV, GJORGJI, DEJAN GJORGJEVIKJ, AND IVAN CHORBEV. "A Multi-class SVM Classifier Utilizing Binary Decision Tree." Informatica (Slovenia) 33, no. 2 (2009): 225-233.

[25] J. Pearl. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufman, San Mateo, California, 1988.

[26]: Aronsky, Dominik, and Peter J. Haug. "Diagnosing community-acquired pneumonia with a Bayesian network." In *Proceedings of the AMIA Symposium*, p. 632. American Medical Informatics Association, 1998.

[27]: Sierra, Basilio, and Pedro Larranaga. "Predicting survival in malignant skin melanoma using Bayesian networks automatically induced by genetic algorithms. An empirical comparison between different approaches." *Artificial Intelligence in Medicine* 14, no. 1 (1998): 215-230.

[28]: Kamath, Patrick S., Russell H. Wiesner, Michael Malinchoc, Walter Kremers, Terry M. Therneau, Catherine L. Kosberg, Gennaro D'Amico, E. Rolland Dickson, and W. Kim. "A model to predict survival in patients with end-stage liver disease." *Hepatology* 33, no. 2 (2001): 464-470.

[29] Gevaert, Olivier, Frank De Smet, Dirk Timmerman, Yves Moreau, and Bart De Moor. "Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks." *Bioinformatics* 22, no. 14 (2006): e184-e190.

[30]Li, Jiakai, Gursel Serpen, Steven Selman, Matt Franchetti, Mike Riesen, and Cynthia Schneider. "Bayes Net Classifiers for Prediction of Renal Graft Status and Survival Period." *International Journal of Biological and Biomedical Sciences* 2, no. 4 (2010).

[31] N. Hoot, "Models to Predict Survival after Liver Transplantation," M.S. thesis, Vanderbilt University, Nashville, Tennessee, USA, 2005.
[32] J.-H. Ahn, J.-W. Kwon and Y.-S. Lee, "Prediction of 1-year Graft Survival Rates in Kidney Transplantation: A Bayesian Network Model," in Proc. INFORMS & KORMS, Seoul, Korea, 2000, pp. 505-513.

[33] R. Bouckaert, Bayesian Network Classifiers in Weka, Technical Report, Department of Computer Science, Waikato University, Hamilton, NZ, 2005.

[34] : Jayasurya, K., G. Fung, S. Yu, C. Dehing-Oberije, Dirk De Ruysscher, A. Hope, Wilfried De Neve, Yolande Lievens, P. Lambin, and A. L. A. J. Dekker. "Comparison of Bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy."*Medical physics* 37 (2010): 1401.

[35] Dekker, Andre, Cary Dehing-Oberije, Dirk De Ruysscher, Philippe Lambin, A. Hope, K. Komati, G. Fung, Shipeng Yu, Wilfried De Neve, and Yolande Lievens. "Survival Prediction in Lung Cancer Treated with Radiotherapy: Bayesian Networks vs. Support Vector Machines in Handling Missing Data." In *Machine Learning and Applications, 2009. ICMLA'09. International Conference on*, pp. 494-497. IEEE, 2009.

[36] Ng, Terence, Lita Chew, and Chun Wei Yap. "A Clinical Decision Support Tool To Predict Survival in Cancer Patients beyond 120 Days after Palliative Chemotherapy." *Journal of Palliative Medicine* 15, no. 8 (2012): 863-869.

[37] Sesen, M. Berkan, Ann E. Nicholson, Rene Banares-Alcantara, Timor Kadir, and Michael Brady. "Bayesian Networks for Clinical Decision Support in Lung Cancer Care." *PloS one* 8, no. 12 (2013): e82349.

[38] Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. "The WEKA data mining software: an update."*ACM SIGKDD Explorations Newsletter* 11, no. 1 (2009): 10-18.

[39] Darwiche, Adnan. Computer Science Department University of California Los Angeles (UCLA), "Sensitivity Analysis, Modeling, Inference and More (SamIam)."

Retrieved from: http://g6g-softwaredirectory.com/ai/Bayesian-networks/20740-UCLA-SamIam.php.

[40] Platt, John C. "12 FAST TRAINING OF SUPPORT VECTOR MACHINES USING SEQUENTIAL MINIMAL OPTIMIZATION." (1999).

[41] WITTEN, IAN H., AND EIBE FRANK. *DATA MINING: PRACTICAL MACHINE LEARNING TOOLS AND TECHNIQUES*. MORGAN KAUFMANN, 2005.

[42] Platt, John. "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods." *Advances in large margin classifiers* 10, no. 3 (1999): 61-74.

[43] Edge, Stephen B., and Carolyn C. Compton. "The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM." *Annals of surgical oncology* 17, no. 6 (2010): 1471-1474..

[44] National Cancer Inistitute, "Cancer Staging." Accessed January 21, 2014. http://www.cancer.gov/cancertopics/factsheet/detection/staging.

[45] R.O. Duda, P.E. Hart, and D.G. Stork, *"Pattern Classication",* WileyInterscience, 2002.

[46] University of Nebraska. Medical Center. Department of Internal Medicine, "Using the Receiver Operating Characteristic (ROC) curve to analyze a classification model ." Accessed January 14, 2014. http://www.math.utah.edu/~gamez/files/ROC-Curves.pdf.

[47] Jensen, Finn Verner, and Thomas Dyhre Nielsen. *Bayesian networks and decision graphs*. Springer, 2007.

[48] Mitchell, Tom M. "Machine learning and data mining." *Communications of the ACM* 42, no. 11 (1999): 30-36.

7. APPENDIX A: DATA SET

	Age	SEX	Lost Weight	Tumour Location	OP Type	Tumour Length (cm)	Complaint Duration	dysphagia	odynophagia	regurgitation	Nausea	Vomit	stomach ache	Stuck	Halitosis	Calcium	CA-ionize	Potassium	Chloride	Sodium	Sedim	Hb	Hct
1	42	1	1	2	4	4.5	180	1	1	0	0	1	0	0	0	9.61	4.6	4.65	103	141	60	11	33.6
2	30	2	1	2	4	5.5	60	1	0	0	0	0	0	0	0	9.83	4.8	4.41	102	139	18	16	50
3	62	2	0	2	4	2.5	30	1	0	0	0	0	0	0	0	9.97	4.5	5.23	100	141	54	14	43.2
4	68	1	1	3	2	7	180	1	1	0	0	0	0	0	0	9.03	4.2	4.23	100	136	40	14	43.3
5	57	2	1	2	4	3.5	180	1	0	0	0	0	0	1	0	9.28	4.2	4.46	103	142	83	12	36.3
6	46	2	0	1	5	4	210	1	1	0	0	0	0	0	0	8.1	4.3	4.03	107	143	64	6.7	23.8
7	27	2	2	2	4	5	60	1	0	0	0	0	0	1	0	10.09	4.7	6.09	100	140	16	17	50.1
8	78	1	1	2	4	3	45	1	0	0	0	0	0	0	0	8.47	4.4	4.18	108	140	22	11	34.3
9	69	2	0	3	2	3.5	60	1	1	0	0	0	0	0	0	8.44	4.2	3.98	108	141	50	12	36.4
10	51	2	0	2	4	5	30	1	0	0	0	0	0	0	0	9.67	4.4	4.18	97	136	64	14	42.3
11	56	2	1	3	2	2	90	1	0	0	0	0	0	1	0	9.62	4.6	5.09	107	142	37	9.5	30.8
12	67	1	1	2	2	5.5	60	1	1	0	0	0	0	1	0	11.5	4.9	4.4	112	140	13	14	42.2
13	59	2	2	2	4	4	90	1	0	0	0	0	0	0	0	9.09	4.6	4.55	101	135	3	16	46.7
14	72	1	1	1	5	2	240	1	0	0	0	0	0	0	0	8.5	4.5	3.49	107	141	35	13	37.3
15	60	2	2	2	4	3.5	30	1	0	0	0	0	0	0	0	9.85	4.9	4.21	106	147	18	15	47.7
16	71	1	1	2	4	6	365	1	0	0	0	0	0	0	0	9.76	4.8	4.23	101	139	46	13	36.5
17	35	2	2	2	4	4	60	1	0	0	0	0	0	0	0	10.05	4.9	4.8	107	143	40	14	45
18	36	1	0	1	5	2.5	15	0	0	0	0	1	0	0	0	8.33	4.8	3.35	103	137	59	9.6	28.5
19	53	1	1	3	2	5	120	1	1	0	1	1	1	1	0	9.3	4.7	3.9	104	138	38	11	32.4
20	44	1	2	2	4	4	90	1	0	0	0	0	0	0	0	8.4	4.5	3.47	105	141	11	12	33.7
21	59	2	1	2	4	3	60	1	0	1	0	0	0	0	0	9.79	4.6	5.38	103	138	26	14	42.3
22	76	2	0	2	4	7	60	1	0	0	0	0	0	0	0	9.63	4.5	4.55	107	145	1	18	55.3
23	51	1	0	1	5	5	365	1	0	0	0	0	0	0	0	8.75	4.5	4.19	100	141	83	12	36.2
24	59	2	1	3	2	3	60	1	0	0	0	0	0	0	0	9.06	4.7	4.2	98	135	43	11	33.4
25	49	2	1	2	4	7	365	1	0	0	0	0	1	0	0	9.18	4.4	3.93	99	133	68	13	38.3
26	74	2	2	3	2	8	90	1	0	0	0	1	0	0	0	7.1	4.2	2.71	100	143	64	11	32.4

	Age	SEX	Lost Weight	Tumour Location	OP Type	Tumour Length (cm)	Complaint Duration	dysphagia	odynophagia	regurgitation	Nausea	Vomit	stomach ache	Stuck	Halitosis	Calcium	CA-ionize	Potassium	Chloride	Sodium	Sedim	Hb	Hct
27	70	1	1	3	2	5.5	60	1	0	0	0	0	0	0	0	9.01	4.5	4.8	103	143	28	13	39.3
28	47	1	1	3	2	6	180	1	0	0	0	0	0	0	0	8.8	4.2	2.75	104	139	106	10	30.3
29	76	2	1	2	4	4	365	1	0	0	0	0	0	0	0	8.64	4.7	3.74	99	135	17	13	41.1
30	69	2	2	3	2	5	365	1	0	0	0	0	0	0	0	9.9	4.7	4.95	100	141	19	15	45
31	30	1	3	1	5	4.5	365	1	0	0	0	0	0	0	0	9	4	4.89	108	143	10	13	40.3
32	56	2	1	2	2	3	120	1	0	0	0	0	0	0	0	8	4.2	4.3	104	139	8	13	36
33	34	1	2	2	4	4	150	1	0	0	0	0	0	0	0	9.06	4.7	3.88	106	141	18	13	38.4
34	73	1	1	3	2	6	30	1	0	0	0	0	0	0	0	9.05	4.7	3.4	98	137	35	14	42.3
35	65	2	2	3	2	5	15	0	0	0	0	1	1	0	0	9.19	4.8	5.38	106	144	61	10	32.8
36	60	2	1	3	2	4.5	120	1	0	0	0	0	0	0	0	9.3	4.7	4.2	100	137	12	12	36.6
37	59	2	1	3	2	3	120	1	0	0	0	0	0	0	0	8.8	4.2	3.69	105	144	40	13	36.8
38	45	1	1	1	5	4	60	1	0	0	0	0	0	0	0	8.3	4.7	3.97	104	142	46	14	38.7
39	69	1	2	2	3	2.5	45	1	0	0	0	0	0	0	0	8.7	4.6	4.45	100	140	55	14	40
40	56	2	1	3	4	6	60	1	0	0	0	0	1	0	0	9.3	4.8	4.4	105	138	22	15	42.2
41	55	1	1	1	5	5	120	1	0	0	0	0	1	0	0	8.6	4.3	4.02	109	145	66	11	35.3
42	63	2	1	3	1	3	30	1	0	0	0	0	0	0	0	9.02	4.4	4.9	102	140	38	16	48.5
43	77	2	1	2	4	5	180	1	0	0	0	1	0	0	0	8.7	4.1	3.98	105	141	48	13	39.4
44	37	2	2	1	5	1.5	365	1	0	0	1	1	0	0	0	8.6	4.1	4.25	102	151	19	16	48.2
45	75	2	1	2	2	2	60	1	0	0	0	0	0	0	0	9.1	4.4	3.88	102	143	54	13	38.9
46	41	2	1	2	4	2.5	120	1	0	0	0	0	0	0	0	7.1	3.1	5.49	99	137	29	13	36.5
47	40	2	1	1	5	9	180	1	0	0	0	0	0	0	0	8.9	4.3	4.27	101	141	29	14	41
48	38	2	1	2	4	5	60	1	0	0	0	1	0	0	0	9.01	4.1	3.8	97	135	37	12	31.8
49	61	2	1	2	2	1.5	30	1	0	0	1	1	0	0	0	8.9	4	4.79	98	143	2	16	43.5
50	35	2	2	2	4	4	30	1	0	1	1	1	0	1	0	9.6	4.6	4.4	107	144	25	16	47.1
51	63	2	1	2	3	3.5	30	1	1	0	0	0	0	1	1	9.7	4.8	3.8	105	143	15	13	37.9
52	46	2	2	2	4	6	45	1	1	0	1	1	0	1	0	9.3	4.5	3.8	101	136	18	13	40.1
53	25	2	1	2	4	7	60	1	1	1	1	1	1	1	1	9	4.8	4	104	139	15	12	35.8
54	55	2	2	3	2	9	40	1	0	0	0	0	0	1	0	8.5	4.4	4.5	102	132	17	10	32.5
55	43	2	2	2	3	4.5	60	1	1	1	0	0	0	1	0	9.3	4.6	3.78	97	136	16	14	40
56	58	2	1	2	4	4.5	90	1	1	1	0	0	0	0	0	9.4	5.1	3.8	103	139	22	17	49.3

	Age	SEX	Lost Weight	Tumor Location	OP Type	Tumor Length (cm)	Complaint Duration	dysphagia	odynophagia	regurgitation	Nausea	Vomit	stomach ache	Stuck	Halitosis	Calcium	CA-ionize	Potassium	Chloride	Sodium	Sedim	Hb	Hct
57	44	2	3	3	2	0	45	1	1	1	1	1	1	1	1	9.8	4.7	4.3	100	135	12	12	37.1
58	72	2	2	3	2	2.5	30	1	0	0	0	0	0	0	0	9.4	5.2	4.5	106	139	18	13	38.5
59	71	2	1	2	4	3.5	45	1	0	1	1	1	0	0	0	9.26	4.2	3.9	101	140	26	14	40.3
60	51	2	2	2	4	4	90	1	1	1	0	0	1	0	0	9	4.8	4.1	104	139	10	13	37.1
61	55	2	1	2	4	5	45	1	1	0	0	0	0	0	0	9.1	4.8	4.5	103	139	33	13	39.3
62	37	2	3	3	4	7	30	1	0	1	1	1	0	0	0	9	4.9	4	105	136	13	13	38.5
63	59	1	2	3	4	6.5	90	1	1	1	0	0	0	1	1	9.3	4.9	4.2	106	139	12	13	38.6
64	51	1	1	2	4	5	120	1	1	1	1	1	0	0	0	10	5.3	4.1	104	140	16	16	45.8
65	55	1	1	2	4	1.5	45	1	1	1	1	1	1	0	0	9.2	4.9	4.3	104	140	23	11	33.7
66	70	2	2	3	4	4	60	1	0	0	0	0	0	0	0	9.52	4.4	4.8	100	137	33	14	41.8
67	67	1	1	3	2	4	45	1	1	1	0	0	0	0	0	9.4	4.6	4.6	110	142	24	14	42.7
68	70	1	1	1	1	2.5	30	1	1	0	0	0	0	0	0	9.7	4.9	4.1	107	141	18	14	41
69	55	1	2	1	5	5.5	30	1	1	1	1	1	0	1	0	9.5	4.7	3.84	106	140	28	13	37.3
70	60	2	3	2	4	4	45	1	1	0	0	0	0	1	1	9.2	4.4	4.3	107	138	68	15	43.1
71	40	2	2	2	4	5	30	1	1	1	1	1	0	0	1	9.15	4.6	4.1	107	141	42	15	45
72	47	2	1	2	4	4	90	1	1	0	0	0	0	1	1	9.5	4.7	3.9	104	141	15	11	34
73	65	1	1	3	4	2	30	1	1	1	1	1	0	1	0	9.14	4.4	4.75	97	138	44	13	41
74	52	1	1	2	4	4	45	1	1	0	0	0	0	1	1	10.13	4.7	3.77	106	140	21	12	38.9
75	55	2	2	3	2	5	45	1	1	0	0	0	1	1	0	9.65	4.6	5.22	100	135	27	15	43.2
76	51	1	2	2	4	2.5	180	1	0	0	0	0	1	1	0	9.84	4.8	4.32	104	139	25	14	41.1
77	45	2	2	2	4	3.5	45	1	0	0	0	0	0	1	1	9.37	4.6	4.03	105	139	51	14	39.1
78	42	2	2	2	4	4	30	1	1	0	0	0	0	1	0	9.37	4.2	4.52	109	144	42	15	43.4
79	68	2	2	3	2	10.5	45	1	1	1	1	1	0	1	1	8.41	4.5	3.52	101	135	15	9.1	30
80	33	2	2	2	4	4.5	30	1	1	1	1	1	0	1	0	8.81	4.5	3.94	100	138	5	15	42.7
81	56	1	1	3	2	3.5	60	1	1	1	1	1	1	1	0	10	4.7	4.03	106	142	12	12	34.9
82	53	2	2	3	2	4	60	1	0	0	0	0	0	1	0	10.13	4.7	4.53	95	142	32	14	40.8
83	60	2	3	3	2	5.5	45	1	0	1	1	1	0	0	1	9.12	4.6	3.76	103	143	22	10	30.1
84	57	2	1	1	5	4.5	180	1	1	0	0	1	0	0	0	9.61	4.6	4.65	103	141	60	11	33.6
85	39	2	1	3	2	7	60	1	0	0	0	0	0	0	0	9.83	4.8	4.41	102	139	18	16	50
86	62	1	0	2	4	1.5	30	1	0	0	0	0	0	0	0	9.97	4.5	5.23	100	141	54	14	43.2

	Age	SEX	Lost Weight	Tumor Location	OP Type	Tumor Length (cm)	Complaint Duration	dysphagia	odynophagia	regurgitation	Nausea	Vomit	stomach ache	Stuck	Halitosis	Calcium	CA-ionize	Potassium	Chloride	Sodium	Sedim	Hb	Hct
87	59	1	1	2	3	3	180	1	1	0	0	0	0	0	0	9.03	4.2	4.23	100	136	40	14	43.3
88	49	2	1	3	4	2.5	180	1	0	0	0	0	0	1	0	9.28	4.2	4.46	103	142	83	12	36.3
89	62	2	0	3	3	4.5	210	1	1	0	0	0	0	0	0	8.1	4.3	4.03	107	143	64	6.7	23.8
90	52	1	2	3	4	7	60	1	0	0	0	0	0	1	0	10.09	4.7	6.09	100	140	16	17	50.1
91	30	2	1	1	5	1.5	45	1	0	0	0	0	0	0	0	8.47	4.4	4.18	108	140	22	11	34.3
92	58	1	0	2	4	2	60	1	1	0	0	0	0	0	0	8.44	4.2	3.98	108	141	50	12	36.4
93	43	2	0	2	4	4	30	1	0	0	0	0	0	0	0	9.67	4.4	4.18	97	136	64	14	42.3
94	44	1	1	2	4	1	90	1	0	0	0	0	0	1	0	9.62	4.6	5.09	107	142	37	9.5	30.8
95	53	2	1	1	5	3.5	60	1	1	0	0	0	0	1	0	11.5	4.9	4.4	112	140	13	14	42.2
96	60	2	2	2	3	2	90	1	0	0	0	0	0	0	0	9.09	4.6	4.55	101	135	3	16	46.7
97	76	2	1	1	5	6	240	1	0	0	0	0	0	0	0	8.5	4.5	3.49	107	141	35	13	37.3
98	67	2	2	1	5	3	30	1	0	0	0	0	0	0	0	9.85	4.9	4.21	106	147	18	15	47.7
99	62	1	1	2	4	5	365	1	0	0	0	0	0	0	0	9.76	4.8	4.23	101	139	46	13	36.5
100	53	1	2	2	4	4	60	1	0	0	0	0	0	0	0	10.05	4.9	4.8	107	143	40	14	45
101	80	1	0	3	1	2	15	0	0	0	0	1	0	0	0	8.33	4.8	3.35	103	137	59	9.6	28.5
102	49	2	1	2	3	4	120	1	1	0	1	1	1	1	0	9.3	4.7	3.9	104	138	38	11	32.4
103	42	1	2	3	4	4.5	90	1	0	0	0	0	0	0	0	8.4	4.5	3.47	105	141	11	12	33.7
104	50	1	1	3	3	2.5	60	1	0	1	0	0	0	0	0	9.79	4.6	5.38	103	138	26	14	42.3
105	75	2	0	3	4	3	60	1	0	0	0	0	0	0	0	9.63	4.5	4.55	107	145	1	18	55.3
106	65	2	0	3	1	5	365	1	0	0	0	0	0	0	0	8.75	4.5	4.19	100	141	83	12	36.2
107	76	1	1	2	4	2.5	60	1	0	0	0	0	0	0	0	9.06	4.7	4.2	98	135	43	11	33.4
108	66	1	1	3	2	6	365	1	0	0	0	0	1	0	0	9.18	4.4	3.93	99	133	68	13	38.3
109	63	2	2	1	5	3	90	1	0	0	0	1	0	0	0	7.1	4.2	2.71	100	143	64	11	32.4
110	75	2	1	3	4	4	60	1	0	0	0	0	0	0	0	9.01	4.5	4.8	103	143	28	13	39.3
111	66	2	1	2	4	3.5	180	1	0	0	0	0	0	0	0	8.8	4.2	2.75	104	139	106	10	30.3
112	62	2	1	3	2	2.5	365	1	0	0	0	0	0	0	0	8.64	4.7	3.74	99	135	17	13	41.1
113	68	1	2	1	5	2	365	1	0	0	0	0	0	0	0	9.9	4.7	4.95	100	141	19	15	45
114	52	2	3	2	3	3.5	365	1	0	0	0	0	0	0	0	9	4	4.89	108	143	10	13	40.3
115	68	2	1	3	4	4	120	1	0	0	0	0	0	0	0	8	4.2	4.3	104	139	8	13	36
116	62	2	2	3	2	6	150	1	0	0	0	0	0	0	0	9.06	4.7	3.88	106	141	18	13	38.4

	Age	SEX	Lost Weight	Tumour Location	OP Type	Tumour Length (cm)	Complaint Duration	dysphagia	odynophagia	regurgitation	Nausea	Vomit	stomach ache	Stuck	Halitosis	Calcium	CA-ionize	Potassium	Chloride	Sodium	Sedim	Hb	Hct
117	62	1	1	2	4	5	30	1	0	0	0	0	0	0	0	9.05	4.7	3.4	98	137	35	14	42.3
118	57	1	2	2	4	4	15	0	0	0	0	1	1	0	0	9.19	4.8	5.38	106	144	61	10	32.8
119	43	2	1	3	4	2.5	120	1	0	0	0	0	0	0	0	9.3	4.7	4.2	100	137	12	12	36.6

	MCV	MCH	MCHC	leukocyte	platelet	fibrinogen	RDW	T. BIL.	Glucose	LDH	FEV 1 lt	pT	pN	M	Stage (TNM)	Grade	Patoloji	MOR-30	MOR-6 month	MOR-12 month	MOR-18 month	MOR-3 year	MOR-5 year	Follow-up time (days)	ADJ CT/RT
1	85	28	32.7	6.6	465	620	14	0.51	88	328	2.1	3	1	0	1	1	1	0	0	1	0	0	0	345	0
2	100	31	31.1	5.74	234	451.04	17	0.53	82	237	2.3	3	2	0	3	3	1	0	1	0	0	0	0	155	1
3	91	29	31.9	6.17	275	426	15	0.67	118	373	3.4	2	1	0	4	1	1	0	0	0	0	0	1	1560	0
4	88	29	33.3	7.43	256	569	16	0.71	184	421	1.9	5	2	0	6	3	2	1	0	0	0	0	0	16	0
5	87	28	32.7	7.03	202	516	14	1.02	106	299	3.70	2	0	0	2	1	3	0	0	0	0	1	0	980	0
6	61	17	28.3	8.39	430	563	22	0.34	63	291	2.7	3	1	0	1	2	1	0	0	1	0	0	0	348	1
7	80	27	33.7	6.65	247	506	14	0.77	82	505	3.8	3	2	0	3	2	1	0	0	1	0	0	0	359	1
8	82	26	32.3	6.24	246	276	17	0.62	78	346	3	3	1	0	1	1	1	0	0	0	0	1	0	1010	0
9	99	32	32.1	5.8	276	442	13	0.37	107	280	3.2	1	2	0	1	2	2	0	0	0	1	0	0	513	1
10	86	28	32.5	16.3	455	582	16	0.56	101	242	2.7	3	2	0	3	3	1	0	0	1	0	0	0	354	1
11	63	19	30.9	8.29	375	575	21	0.42	74	428	4.1	2	0	0	2	3	2	0	0	0	0	0	1	1513	1
12	85	29	34.2	5.64	321	361	14	0.4	93	400	2	3	2	0	3	2	1	0	1	0	0	0	0	161	0
13	88	30	33.6	5.99	169	313	17	0.93	75	281	2	3	1	0	1	1	1	0	0	0	0	1	0	867	1
14	89	30	33.9	13.4	303	659	17	0.42	89	245	2	1	0	0	5	3	1	0	0	1	0	0	0	295	0
15	91	29	32.1	6.1	243	529	13	0.97	84	336	2	5	1	0	6	3	1	0	1	0	0	0	0	151	0
16	78	28	35.2	9.75	378	464	17	0.7	93	400	2.1	5	1	0	6	3	1	0	1	0	0	0	0	177	0
17	80	25	31.4	8.49	314	319	13	0.92	102	352	3.6	3	0	0	2	1	1	0	0	0	1	0	0	444	0
18	80	27	33.7	3.66	141	725.92	17	0.33	128	500	2.8	2	0	0	4	1	1	0	0	0	0	1	0	810	0
19	77	25	32.9	10.5	420	625	12	0.5	105	254	1.8	5	2	0	6	3	1	1	0	0	0	0	0	11	0
20	85	30	35.2	4.49	227	327	13	0.41	90	154	2.6	2	1	0	4	2	1	0	0	0	1	0	0	461	0
21	90	31	34	11.9	340	434	13	1.19	73	668	4.5	2	0	0	2	1	1	0	0	0	0	1	0	801	0
22	91	30	33.3	6.93	142	358	14	1.49	85	420	2.1	5	2	0	6	3	1	0	1	0	0	0	0	169	1
23	74	23	31.8	7.39	430	537	21	0.49	110	298	2	3	1	0	1	2	1	0	0	1	0	0	0	219	1
24	62	21	33.3	9.33	558	721	15	0.6	70	473	3.1	2	1	0	4	1	1	0	0	0	0	1	0	912	0

	MCV	MCH	MCHC	leukocyte	platelet	fibrinogen	RDW	T. BIL.	Glucose	LDH	FEV 1 lt	pT	pN	M	Stage (TNM)	Grade	Patoloji	MOR-30	MOR-6 month	MOR-12 month	MOR-18 month	MOR-3 year	MOR-5 year	Follow-up time (days)	ADJ CT/RT
25	91	32	35.1	7.08	256	242	16	0.5	98	646	1.9	3	2	0	3	3	1	0	1	0	0	0	0	158	1
26	80	26	32.3	4.12	165	412	17	0.67	86	126	4.1	3	2	0	3	3	1	0	1	0	0	0	0	171	0
27	85	28	32.4	9.06	249	331	19	0.67	65	651	2.7	5	1	0	6	2	1	0	0	1	0	0	0	300	1
28	74	25	33.1	10.3	437	842	17	0.83	83	431	2.4	5	1	0	6	3	1	0	0	1	0	0	0	338	1
29	92	29	32	7.25	167	289	17	2.34	135	584	3.8	2	0	0	2	1	1	0	0	0	0	0	1	1611	0
30	84	29	33.9	6.1	281	529	15	1.19	67	716	1.8	5	2	0	6	3	2	1	0	0	0	0	0	8	0
31	72	24	32.4	7.39	392	350	29	0.8	82	414	3.9	3	1	0	1	2	1	0	0	0	1	0	0	175	1
32	95	33	34.8	5.26	191	230	15	0.9	109	502	2.1	2	0	0	2	3	2	0	0	0	0	1	0	804	0
33	88	30	33.9	7.99	149	230	13	0.94	94	358	2	3	2	0	3	2	1	0	1	0	0	0	0	162	1
34	85	29	33.6	6.4	249	352	14	0.44	81	388	2.6	4	1	0	6	2	1	0	1	0	0	0	0	178	0
35	84	26	31.3	9.92	398	512	13	0.4	85	284	3.5	3	1	0	1	1	2	0	0	1	0	0	0	300	0
36	88	30	33.9	6.4	177	306	13	0.97	78	365	4.2	3	1	0	1	1	2	0	0	0	1	0	0	479	1
37	87	30	34.2	5.6	371	380	14	0.5	90	315	2.5	5	0	0	1	2	2	0	0	1	0	0	0	284	1
38	88	31	35.1	4.7	214	380	14	0.6	75	132	1.9	2	0	0	2	1	1	0	0	0	0	0	1	1712	0
39	87	30	34	8.2	290	590	14	0.6	88	119	3.2	3	1	0	1	1	1	0	0	0	0	1	0	873	1
40	85	29	34.6	8.2	170	300	15	0.33	92	283	3.7	3	2	0	3	2	2	0	0	1	0	0	0	328	1
41	81	25	31.2	6.9	252	430	16	0.4	75	580	2	5	2	0	6	3	1	1	0	0	0	0	0	15	0
42	87	29	33.6	11	414	530	14	0.6	101	289	2.2	3	0	0	4	1	2	0	0	0	0	1	0	901	0
43	88	29	32.5	13.2	363	520	13	0.77	91	315	3.1	3	2	0	3	2	1	0	1	0	0	0	0	177	0
44	76	25	32.7	7.85	294	520	12	0.9	91	354	4.3	2	0	0	4	3	1	0	0	0	0	0	0	1900	1
45	83	27	32.7	8.43	301	710	17	0.8	96	251	2.9	2	0	0	2	1	2	0	0	0	0	1	0	877	1
46	78	27	34.9	8.2	329	512	13	0.8	83	130	4.1	2	0	0	2	2	1	0	0	0	1	0	0	435	0
47	86	30	34.9	5.3	256	320	12	0.78	74	149	3.1	3	2	0	3	3	1	0	1	0	0	0	0	85	1
48	76	27	36.2	6.7	243	330	14	0.9	101	198	2.1	3	0	0	4	2	1	0	0	1	0	0	0	259	1
49	91	34	37.2	9.5	219	330	15	2.5	123	146	3.8	1	1	0	4	1	1	0	0	0	0	0	1	1612	1
50	83	29	34.8	5.7	254	284.5	13	1.3	87	329	3.2	3	1	0	1	1	1	0	0	1	0	0	0	259	0
51	86	29	33.1	4.7	240	344.2	13	0.5	88	155	2.8	3	0	0	2	1	1	0	0	0	0	0	1	1578	0
52	83	27	32.9	8.2	275	395	13	0.7	85	280	2.9	5	0	0	1	1	1	0	0	0	0	0	0	30	0
53	98	33	33.3	6.3	225	342.4	14	0.5	67	278	1.8	5	1	0	6	1	1	0	1	0	0	0	0	59	0
54	84	27	32.2	13.6	517	527.3	15	0.6	131	85	3.3	3	0	0	4	1	2	0	0	0	0	0	0	998	0

	MCV	MCH	MCHC	leukocyte	platelet	fibrinogen	RDW	T. BIL.	Glucose	LDH	FEV 1 lt	pT	pN	M	Stage (TNM)	Grade	Patoloji	MOR-30	MOR-6 month	MOR-12 month	MOR-18 month	MOR-3 year	MOR-5 year	Follow-up time (days)	ADJ CT/RT
55	88	30	34.1	6.8	210	330.8	14	0.9	111	321	4.1	3	1	0	1	1	1	0	0	0	0	0	0	355	0
56	86	30	34.6	9.9	190	375	13	0.8	56	226	3.1	2	0	0	2	1	1	0	0	0	0	0	0	28	0
57	86	27	32	6.3	304	295	12	0.9	87	281	2.4	5	1	0	6	3	1	0	0	1	0	0	0	279	1
58	94	32	34.3	4.5	176	293.2	14	2.5	72	310	3.1	5	1	0	6	1	2	0	0	0	0	0	0	26	0
59	86	29	34.1	5.8	260	366	15	0.8	66	210	2.5	3	0	0	2	1	1	0	0	0	0	0	0	879	0
60	87	29	34	5.9	263	397	13	0.5	92	129	2	3	1	0	1	1	1	0	0	0	0	0	0	19	0
61	89	30	33.8	7	257	606	14	1.3	96	150	3.5	3	0	0	2	1	1	0	0	0	0	0	0	311	0
62	89	30	34.2	8	260	308	14	0.3	74	154	2.1	3	1	0	1	1	1	0	0	1	0	0	0	300	0
63	86	29	34	7.2	283	398	13	0.8	66	119	1.9	5	1	0	6	1	1	1	0	0	0	0	0	25	0
64	92	31	33.9	4.6	162	319	13	0.9	82	168	2.1	3	0	0	2	1	1	0	0	0	0	0	0	237	0
65	81	27	33.2	4.3	221	314	17	0.7	68	151	3.10	3	1	0	1	1	1	0	0	0	0	1	0	900	0
66	94	30	32.2	11.1	311	568	15	0.6	75	124	2.9	3	0	0	4	3	3	0	0	1	0	0	0	342	0
67	93	31	33.8	7.1	206	308	13	1.4	78	186	1.9	3	0	0	4	3	3	0	0	0	0	0	1	1731	0
68	94	32	33.8	9.1	233	277.5	14	0.8	95	136	4.1	2	0	0	2	1	1	0	0	1	0	0	0	201	0
69	93	31	33.8	3.6	200	354	12	0.7	93	114	4.7	5	0	0	1	1	1	0	0	0	0	1	0	874	1
70	83	28	34.3	7.3	276	456	15	0.6	85	121	1.9	3	1	0	1	1	1	0	1	0	0	0	0	97	0
71	90	30	33.8	7.2	198	304	12	0.5	96	255	2.8	3	0	0	2	1	1	0	0	0	0	1	0	943	1
72	72	24	32.8	9.1	426	321	33	0.5	78	101	2.1	3	0	0	2	1	1	0	0	0	0	0	0	30	0
73	88	29	32.7	7.2	202	429	14	0.36	88	324	1.8	2	0	0	5	2	3	1	0	0	0	0	0	18	0
74	79	25	31.1	7.2	306	298	18	1.58	92	261	3.7	2	0	0	4	2	1	0	0	1	0	0	0	205	0
75	92	32	34.3	13.8	412	777	13	0.43	93	242	2.1	5	0	0	1	1	1	0	0	1	0	0	0	355	0
76	84	28	33.6	5.46	293	452	13	0.61	84	368	3.7	3	0	0	2	1	1	0	0	1	0	0	0	249	0
77	90	31	34.5	7.36	276	487	15	0.49	96	347	3.9	2	0	0	2	1	1	0	0	0	0	1	0	951	1
78	89	30	33.6	8.2	174	282	13	0.5	89	287	4.5	3	1	0	1	2	1	0	0	1	0	0	0	317	0
79	68	21	30.3	8.7	298	350	19	0.53	181	260	3.9	3	1	0	1	1	1	0	0	0	0	0	1	1733	0
80	80	28	34.7	8.59	269	410	14	0.52	116	289	4.8	3	0	0	4	2	1	0	0	0	0	1	0	794	1
81	86	29	33.8	3.58	124	368	15	0.81	104	448	5.1	1	0	0	5	3	1	0	0	0	1	0	0	179	1
82	86	29	33.3	8.98	496	650	14	0.53	88	306	1.8	3	1	0	1	2	1	1	0	0	0	0	0	9	0
83	87	29	33.2	4.76	237	354	14	0.38	129	311	2.1	5	1	0	6	3	2	0	0	1	0	0	0	351	0
84	85	28	32.7	6.6	465	620	14	0.51	88	328	2.3	3	1	0	1	3	1	0	0	1	0	0	0	278	1

	MCV	MCH	MCHC	leukocyte	platelet	fibrinogen	RDW	T. BIL.	Glucose	LDH	FEV 1 lt	pT	pN	M	Stage (TNM)	Grade	Patoloji	MOR-30	MOR-6 month	MOR-12 month	MOR-18 month	MOR-3 year	MOR-5 year	Follow-up time (days)	ADJ CT/RT
85	100	31	31.1	5.74	234	451.04	17	0.53	82	237	1.9	3	1	0	1	2	2	0	1	0	0	0	0	157	1
86	91	29	31.9	6.17	275	426	15	0.67	118	373	3.9	2	0	0	2	1	1	0	0	0	0	0	1	1649	0
87	88	29	33.3	7.43	256	569	16	0.71	184	421	3.1	2	1	0	4	3	1	0	0	0	0	0	0	30	1
88	87	28	32.7	7.03	202	516	14	1.02	106	299	4.1	3	0	0	4	1	2	0	0	0	0	1	0	877	0
89	61	17	28.3	8.39	430	563	22	0.34	63	291	2.9	3	1	0	1	2	1	0	0	1	0	0	0	299	1
90	80	27	33.7	6.65	247	506	14	0.77	82	505	1.8	5	1	0	6	3	2	1	0	0	0	0	0	10	0
91	82	26	32.3	6.24	246	276	17	0.62	78	346	4.5	1	0	0	5	2	1	0	0	0	0	1	0	812	0
92	99	32	32.1	5.8	276	442	13	0.37	107	280	2.8	2	0	0	2	1	1	0	0	0	1	0	0	170	0
93	86	28	32.5	16.3	455	582	16	0.56	101	242	3.1	3	0	0	4	3	1	0	0	1	0	0	0	361	0
94	63	19	30.9	8.29	375	575	21	0.42	74	428	5.2	1	0	0	5	2	1	0	0	0	0	0	1	1512	0
95	85	29	34.2	5.64	321	361	14	0.4	93	400	1.9	3	2	0	3	2	1	1	0	0	0	0	0	7	0
96	88	30	33.6	5.99	169	313	17	0.93	75	281	4.2	3	0	0	5	1	1	0	0	0	0	1	0	911	0
97	89	30	33.9	13.4	303	659	17	0.42	89	245	2.6	1	2	0	1	1	1	0	0	1	0	0	0	201	1
98	91	29	32.1	6.1	243	529	13	0.97	84	336	2.1	5	0	0	1	3	1	1	0	0	0	0	0	11	0
99	78	28	35.2	9.75	378	464	17	0.7	93	400	1.9	2	2	0	1	1	1	0	1	0	0	0	0	177	1
100	80	25	31.4	8.49	314	319	13	0.92	102	352	3.2	2	0	0	4	2	1	0	0	0	1	0	0	501	0
101	80	27	33.7	3.66	141	725.92	17	0.33	128	500	3.5	2	0	0	5	1	2	0	0	0	0	1	0	879	0
102	77	25	32.9	10.5	420	625	12	0.5	105	254	2.9	3	1	0	1	2	1	0	0	0	0	0	0	30	0
103	85	30	35.2	4.49	227	327	13	0.41	90	154	4.1	3	1	0	1	1	1	0	0	0	1	0	0	467	1
104	90	31	34	11.9	340	434	13	1.19	73	668	3.6	2	0	0	4	2	1	0	0	0	0	1	0	934	0
105	91	30	33.3	6.93	142	358	14	1.49	85	420	2.3	3	2	0	3	3	2	0	1	0	0	0	0	17	1
106	74	23	31.8	7.39	430	537	21	0.49	110	298	2	2	2	0	1	2	2	0	0	1	0	0	0	162	1
107	62	21	33.3	9.33	558	721	15	0.6	70	473	2.9	2	0	0	2	1	1	0	0	0	0	1	0	811	0
108	91	32	35.1	7.08	256	242	16	0.5	98	646	3.2	5	0	0	1	1	1	0	1	0	0	0	0	148	0
109	80	26	32.3	4.12	165	412	17	0.67	86	126	3.8	3	0	0	4	2	1	0	1	0	0	0	0	139	0
110	85	28	32.4	9.06	249	331	19	0.67	65	651	3.9	3	0	0	4	3	2	0	0	1	0	0	0	257	0
111	74	25	33.1	10.3	437	842	17	0.83	83	431	2.3	2	1	0	4	1	1	0	0	1	0	0	0	278	1
112	92	29	32	7.25	167	289	17	2.34	135	584	3.2	3	0	0	5	1	1	0	0	0	0	0	1	1498	0
113	84	29	33.9	6.1	281	529	15	1.19	67	716	2.8	2	0	0	4	3	1	0	0	0	0	0	0	30	0
114	72	24	32.4	7.39	392	350	29	0.8	82	414	4.1	3	0	0	2	1	1	0	0	0	1	0	0	481	1

	MCV	MCH	MCHC	leukocyte	platelet	fibrinogen	RDW	T. BIL.	Glucose	LDH	FEV 1 lt	pT	pN	M	Stage (TNM)	Grade	Patoloji	MOR-30	MOR-6 month	MOR-12 month	MOR-18 month	MOR-3 year	MOR-5 year	Follow-up time (days)	ADJ CT/RT
115	95	33	34.8	5.26	191	230	15	0.9	109	502	2.5	2	0	0	5	1	1	0	0	0	0	1	0	893	1
116	88	30	33.9	7.99	149	230	13	0.94	94	358	2	2	2	0	1	3	2	0	1	0	0	0	0	101	1
117	85	29	33.6	6.4	249	352	14	0.44	81	388	3.1	2	1	0	4	3	1	0	1	0	0	0	0	145	1
118	84	26	31.3	9.92	398	512	13	0.4	85	284	2.4	3	1	0	1	1	1	0	0	1	0	0	0	312	1
119	88	30	33.9	6.4	177	306	13	0.97	78	365	2	3	0	0	4	1	3	0	0	0	1	0	0	448	0