ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE ENGINEERING AND TECHNOLOGY

BREAST CANCER DATA CLASSIFICATION USING SVM, NB AND KNN ALGORITHMS

M.Sc. THESIS

Burcu MERAL

Department of Mathematical Engineering

Mathematical Engineering Programme

MAY 2019

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE ENGINEERING AND TECHNOLOGY

BREAST CANCER DATA CLASSIFICATION USING SVM, NB AND KNN ALGORITHMS

M.Sc. THESIS

Burcu MERAL
(509131053)

Department of Mathematical Engineering

Mathematical Engineering Programme

Thesis Advisor: Prof. Dr. Kamil Oruçoğlu

MAY 2019

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

SVM, NB ve KNN KULLANIMI İLE GÖĞÜS KANSERİ VERİ SINIFLANDIRMASI

YÜKSEK LİSANS TEZİ

Burcu MERAL
(509131053)

Matematik Mühendisliği Anabilim Dalı

Matematik Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Kamil Oruçoğlu

MAYIS 2019

Burcu MERAL, a M.Sc. student of ITU Graduate School of Science Engineering and Technology 509131053 successfully defended the thesis entitled “BREAST CANCER DATA CLASSIFICATION USING SVM, NB AND KNN ALGORITHMS”, which he/she prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Prof. Dr. Kamil Oruçoğlu** Istanbul Technical University	.....
Jury Members :	**Doç. Dr. Reşat Köşker** Yildiz Technical University	.....
	Doç. Dr. Tolga Birkandan Istanbul Technical University	.....
		.....

Date of Submission : **3 May 2019**
Date of Defense : **21 May 2019**

To my family,

FOREWORD

It was not easy to come this far and conclude this thesis with so many obstacles. I would like to express my gratitude and thanks to Prof. Dr. Kamil Oruçoğlu for his kindness, patience and support. I feel so lucky to cross paths with him. And I am very grateful to Assoc. Prof. Atabey Kaygun for sharing his feedback on my study. I would also like to thank all of my colleagues who supported me from the beginning.

Lastly, I would like to thank to my dear mom Sultan Meral and father Adem Meral, my sister Hava Meral and my brothers Aydın Meral, Emirhan Meral for continuing encouragement, devotion and prayers for me from the very beginning. Thanks to their support, I could finish this thesis successfully.

May 2019 Burcu MERAL

TABLE OF CONTENTS

Page

ABBREVIATIONS

IARC	**:** International Agency for Research on Cancer
GLOBOCAN	**:** Global Cancer Incidence, Mortality and Prevalence
BC	**:** Breast Cancer
ML	**:** Machine Learning
SVM	**:** Support Vector Machine
KNN	**:** K Nearest Neighbour
UCI	**:** The University of California at Irvine
WDBC	**:** Wisconsin Breast Cancer Database
AL	**:** Artificial Intelligence
DM	**:** Data Mining
TP	**:** True Positive
FP	**:** False Positive
FN	**:** False Negative
TN	**:** True Negative
AUC	**:** Area Under The Curve
ROC	**:** Receiver Operating Characteristics
FPR	**:** False Positive Rate
TPR	**:** True Positive Rate
RBF	**:** Gaussian Radial Basis Function
FNA	**:** Fine Needle Aspirate
SE	**:** Standard Error
NaN	**:** Not a Number
KDD	**:** Knowledge Discovery in Databases
NaN	**:** Not a Number
MLPNN	**:** Multilayer Perceptron Neural Network
CNN	**:** Combined Neural Network
PNN	**:** Probabilistic Neural Network
RNN	**:** Recurrent Neural Network
IBK	**:** Instance Based for K-Nearest Neighbor
SMO	**:** Sequential Minimal Optimization
NB	**:** Naive Bayes
REP	**:** Reduced Error Pruning
WEKA	**:** Waikato Environment for Knowledge Analysis
MRI	**:** Magnetic Resonance Imaging
Std	**:** Standard Deviation

LIST OF TABLES

Page

LIST OF FIGURES

Page

BREAST CANCER DATA CLASSIFICATION USING SVM, NB AND KNN ALGORITHMS

SUMMARY

After skin cancer, the most common cancer type among women is the breast cancer. However rare, men can also suffer from this disease. Breast cancer also has the highest survival rate among all cancers if it was diagnosed in early stage of the disease. The death rate declines depending on how early it was detected, and discovery of new treatment techniques.

The most important symptoms of breast cancer is a lump or area of thickened breast tissue. Most lumps are not cancerous, but deciding the type of tumor has vital importance. There are two types of breast tumors: malignant and benign. Malignant cells are the one that cause death.

There are many risk factors that leads to breast cancer such as the family history, age, being diagnosed with certain benign (non-cancerous) breast tumors previously, genetic risk factors, alcohol consumption, radiation exposure. Mammograms are useful for breast cancer screening. Nevertheless they have limitations and it is possible that the cancer could not yet developed at the time of the mammography.

Breast cancer has 5 stages. Stage 0 is the beginning phase in which the cells have not yet invaded surrounding tissues. In Stage 1, cancer cells spread in small area. In Stage 2 and 3, cancer cells increase in size. In the final stage 4, the cancer spread the vital organs like brain, liver, lungs.

Data mining is a collection of algorithms and techniques that aim make meaningful conclusions from large amounts of data. It is a multidisciplinary discipline that uses statistics, artificial intelligence, database technology and machine learning.

Machine learning uses two types of learning methods: supervised learning and unsupervised learning. Supervised learning aims to improve a model based on both input and output data. Classification and regression algorithms are the among the supervised learning models. Supervised learning algorithms search for relationships between input attributes and target attributes. In this thesis we investigate few classification algorithms. In unsupervised learning, the machine is not trained so machine has to find hidden structure in unlabelled data. Clustering and association is the type of unsupervised learning.

Using technology in health care system provides crucial benefits. In recent years, machine learning algorithms have been immensely used in the medical field from disease diagnosis to improving patient care. Algorithms identifying patterns in medical images had huge impact and opened a different avenues of diagnosis and treatment in health care. Examples include detecting tuberculosis, brain aneurysms or Alzheimer.

This thesis is focused on the application of several data mining techniques on Wisconsin breast cancer dataset. In order to improve breast cancer cell recognition,

the proposed system includes comparisons of three different, commonly used machine learning algorithms: K-nearest-neighbor (KNN), naive Bayes (NB) and support vector machines (SVM). The aim of this dissertation is to employ these three classification algorithms on Wisconsin Breast Cancer Database dataset and compare three classification algorithms in terms of their performance and accuracy rate. According to the findings SVM algorithm beat the NB and KNN algorithms.

SVM, NB ve KNN KULLANIMI İLE GÖĞÜS KANSERİ VERİ SINIFLANDIRMASI

ÖZET

Deri kanserinden sonra, kadınlar arasında en sık görülen kanser türü göğüs kanseridir. Sadece kadınlar değil, erkekler de bu hastalıktan nadir bir şekilde de olsa muzdarip olabilir. Eğer erken teşhis edilirse yaşam oranı da çok yüksektir ve buna bağlı olarak ölüm oranında azalma görülür.

Göğüs kanserinin en önemli belirtileri şişkinleşmiş veya kalınlaşmış dokunun alanıdır. Kötü huylu ve iyi huylu olarak adlandırılan iki tıp tümör vardır. Tümörün tipine göre kanser hücresinin iyi huylu ya da kötü huylu olup olmadığına karar vermek gerekmektedir. Kötü huylu hücreler ölüme neden olan hücrelerdir. Göğüs kanserine neden olan birçok risk faktörü vardır; aile geçmişi, yaş, daha önceden iyi huylu (kanserli olmayan) hücre bulgusu teşhisi konması, genetik risk faktörü, alkol tüketimi, radyasyona maruz kalma.

Mamogramlar Göğüs kanseri taraması için kullanılır, ancak yine de mamogramlarında kanseri teşhis etmekte yanılma payı vardır. Göğüs kanseri 5 evrelidir. Başlangıç aşaması olan aşama 0'da, hücreler etrafındaki dokuları istila etmemiştir. Aşama 1'de kanser hücreleri küçük bir alana yayılır. 2. ve 3. aşamada ise kanser hücreleri büyümeye başlar. En son evre olan 4. aşama ise, kanser beyin, karaciğer, akciğerler gibi hayatı organlara yayılmaya başlamıştır. Göğüs kanseri genellikle son aşamalarda teşhis edildiği için ölüm oranları bu kadar yüksektir.

Veri madenciliği, büyük miktarda veriden anlamlı sonuçlar çıkarmak için gerekli süreçlerin bütünüdür. Gizli kalıpları keşfetmeye yardımcı olacak birçok teknik vardır. İstatistik, yapay zeka, veri tabanı teknolojisi ve makine öğrenmesi kullanılan çok disiplinli bir beceridir. Artan ham veri nedeniyle, verilerin analiz edilmesi ve bu analizler baz alınarak çıkarımların yapılması hayatı öneme sahiptir. Bu yüzden veri madenciliğinin değeri gün geçtikçe artmaktadır.

Makine öğrenmesi, denetlenen öğrenmeyi ve denetimsiz öğrenmeyi sağlayan iki tür öğrenme yöntemi kullanır. Denetimli öğrenme, hem girdi hem de çıktı verilerine dayanan bir model geliştirmektir. Yani elimizdeki veri kümesinin ne olduğunu ve bu verilerden ne gibi sonuçlar çıkması gerektiğini biliriz. Sınıflandırma algoritmaları ve regresyon algoritmaları, denetlenen öğrenme modelinin türüdür. Temelde, kaynak veri ile hedef veri arasında bir ilişki arar ve bulur. Bazı sınıflandırma algoritmaları bu tezin de konusudur ve denetimli öğrenmenin örneklerindendir. Denetimsiz öğrenmede, verilerden elde edilmek istenen çıktının ne olduğunun daha önceden bilinmemesi ile oluşturulan modeldir. Kümeleme ve ilişkilendirme algoritmaları, denetimsiz öğrenmenin türüdür.

Son yıllarda, makine öğrenme algoritmaları tıbbi alanda yaygın olarak kullanılmaya başlamıştır. Hedef sağlık alanında gelişmelere katkıda bulunmak ve hastalıkların

teşhisinde yardımcı olmaktır. Örneğin, röntgen filmlerini kullanarak hastalığın teşhisinin koyulması ile sağlık alanında önemli adımların temelleri atılmıştır. Tüberkülozu bulmak, beyin kanamalarını veya Alzheimer hastalığını tespit etmek, röntgen filmlerine bakarak tanıları teşhis ve tedavi etmek veri bilimi sayesinde mümkün olmuştur. Bu nedenle, teknolojiyi sağlık alanında kullanmak önemli faydalar sağlamıştır.

Makine algoritmaları kullanılarak sağlık alanında önemli gelişmeler ve yenilikler kaydedilmektedir. Bu tezin konusu da yine sağlık alanında en çok görülen kanser tiplerinden biri olan göğüs kanserinin önceden veya daha erken evrelerde keşfedilmesini sağlayacak bir algoritma oluşturmaktır. Oluşturulan bu algoritmaların doğruluk oranının en yüksek olması hedeflenmektedir.

Göğüs kanseri hücresinin iyi huylu ya da kötü huylu olup olmadığını tespit etmek için kullanılan birçok sınıflandırma algoritması mevcuttur. Bunlardan k en yakın komşuluk, naive bayes ve destek vektör makinesi algoritmaları bu tezin de konusunu oluşturur. Göğüs kanseri, özellikle kadınlarda erken teşhis edilmedikçe ölüme neden olan başlıca kanser türüdür. Bunu önlemek ve erken teşhise yardımcı olmak için bu üç algoritmaların uygulanması ve sonuçlara göre doğruluk oranlarının karşılaştırılması yapılmıştır.

K en yakın komşuluk sınıflandırma algoritması temel olarak bilinmeyen noktaya en yakın k tane komşunun seçilmesi ve ordaki çoğunluğun oyuna göre bilinmeyen noktanın çoğunluğun komşuluğuna atanmasını ifade eder. Bu yüzden tembel öğrenme algortiması olarak adlandırılır. Asıl hesaplama sınıflandırma adımına kadar ertelenir. Çok büyük veri setlerinde kullanılması tavsiye edilmez.

Naif bayes algoritması, olasılık ilkelerinde bayes theoremini kullanarak verinin hangi kategoriden olduğunu tespit etmek için kullanılır.

Destek vektör makine algoritmasının amacı, sınıflandırılacak veri için doğru hiper düzlemi sağlamak yani iki ayrı sınıfa sahip veri için bu verileri birbirinden ayıracak çizgiyi yada sınırı bulmaktır. Doğru hiper düzlem kıstası iki sınıfı en iyi ayıran, en uzak çizgiyi bulmaktır. İki sınıf arasındaki uzaklık marjın olarak adlandırılır. Marjın en yüksek değere sahipken algoritma en doğru hiper düzlemdedir denilebilir.

Fakat veriler her zaman birbirinden keskin bir çizgi ile ayrılmayabilir. Bu tarz durumlarda çekirdek hilesi adı verilen bir yöntem kullanılır. Bu yöntem ile veri bir üst boyuta taşınır ve ayırma işlemi burada yapılır. En yaygın olarak kullanılan çekirdek yöntemleri; doğrusal, polinomiyal ve radial tabanlı fonksiyondur. Bu tezde de bu 3 çekirdek yöntemi göğüs kanseri verisetine uygulanmıştır ve radial tabanlı fonksiyonun diğerlerine göre daha etkili ve yüksek doğruluk ürettiği gözlenmiştir.

Destek vektörleri genel olarak nesne tanıma; yüz ve parmak izi tanıma, el yazısı tanıma, yazı karakteri tanıma, tıbbi tahminler, kanser verileri vb. alanlarda sıklıkla kullanılır. Hem doğrusal hem doğrusal olmayan verilere uygulanabildiği için yaygın olarak kullanılan bir algoritmadır. Aynı zamanda yüksek doğruluk oranına sahiptir. Dezavantaj olarak çekirdek fonksiyonlarının pozitif tanımlı sürekli fonksiyonlar olması sayılabilir.

Bu tezin amacı, Wisconsin Göğüs Kanseri Veri Tabanı veri seti kullanılarak bu üç sınıflandırma algoritmasının performanslarını ve doğruluk oranlarını karşılaştırmaktır. Bu veriseti hücrenin boyutu, yapısı, çapı gibi bilgilerden oluşmaktadır.

Sınıflandırma algoritmalarını uygulamak için Wisconsin Göğüs Kanseri Veri Tabanı verileri test ve eğitim olmak üzere iki gruba ayrıldı. Buna göre 569 verinin bulunduğu WDBC veri setinin 381 kaydı eğitim için 188 kaydı test için kullanılmıştır. Yani veri kümesi $\frac{2}{3}$ ü eğitim datası $\frac{1}{3}$ ü test datası olmak üzere iki gruba ayrılmıştır. Bulgulara göre SVM algoritması NB ve KNN algoritmasından daha yüksek doğruluk oranına sahiptir. SVM algoritması 3 tane kernel fonksiyonu için uyguladığımızı daha önce söylemiştik. Buna göre lineer çekirdek için %93.1, polinomiyal için %85.1 ve radiyal tabanlı fonksiyon için %98.4 luk bir doğruluk oranı elde edilmiştir.

1. INTRODUCTION

Cancer is a term for a group of diseases that can occur because of mutations and abnormal changes in the genes. Breast cancer refers to the abnormal changes in the cells of the breast tissue. This process may form a lump or mass of extra tissue called tumor. There are two types of tumor called cancerous (malignant) or non-cancerous (benign). This classification depends on whether or not tumors can spread by invasion or metastasis. Benign tumors can not spread by invasion and metastasis, but they grow locally. However malignant tumors can spread. Breast cancer occurs because of malignant tumors developed in the breast [8–10].

According to International Agency for Research on Cancer (IARC) using the Global Cancer Incidence, Mortality and Prevalence (GLOBOCAN) 2018 female breast cancer (BC) is the second most commonly diagnosed cancer type in the world and the leading cause of cancer death among women [11]. Breast cancer is the reason the 11.6% of the total cases of the death in worldwide cancer incidences. It is estimated that in 2018 about 2.1 million new diagnoses will be made [12].

According to a research made by American Cancer Society (ACS) 252,710 women suffered from invasive breast cancer in America in 2017 alone. It is a rare disease for men with less than 1% of all breast cancer cases. Approximately 41,000 people are expected to die from breast cancer in 2017 in America [9].

The mortality rate of the breast cancer (BC) is decreased with the help of improvements in treatment and early detection [9]. Finding BC in early stages is the most important issue for preventing deaths. So, health care systems can take advantage from data mining algorithms that can help to detect diseases in early stages.

The use of machine learning (ML) algorithms in medical field has been increasing gradually. With the advent of new technologies in medicine, large amounts of cancer data have been collected and are now available to the medical research community. This also creates a challenge for medical researchers to identify the right pattern from

complex data sets and find the techniques to diagnose cancer based on these patterns. However, examination of samples taken from patients are the key part of finding the right diagnosis. A variety of ML techniques including Support Vector Machines (SVMs), Naive Bayes (NB), K-Nearest Neighbors (KNN) are used in this thesis. These techniques have been widely applied in cancer research to develop predictive models that aims effective and accurate decision making.

The main objective of ML algorithms is to create models that can be used to perform classification, prediction and estimation. A good classification model should accurately classify the testing data thus the data set is divided into two sets called training set and testing set. By using training set, model is trained and then it is used to predict the unlabelled testing data.

The main aim of this study is to find accurate data mining algorithms with higher classification success rates that can help in detecting breast cancer in early stages. Since the mortality rate can be decreased by diagnosing the disease in early stages, this study is expected to improve the quality of treatment. By comparing the performance of applied methods and their classification accuracy, we measure and determine the best algorithms that perform better than the other methods implemented.

In this study, we employ KNN, NB and SVM algorithms to diagnose BC. We used Wisconsin Breast Cancer Dataset [1] from The University of California at Irvine (UCI) Machine Learning Repository [13] for training and testing experiments.

This thesis consists of four Chapters. In Chapter 1, we give general information about breast cancer and mention other studies related to diagnosing breast cancer. We also explain the basic concepts of data mining and why it has growing importance in health care. In Chapter 2, we explain the classification methods that are applied in this study. These are Support Vector Machines (SVM), Naive Bayes (NB) and the k-Nearest Neighbor (KNN) algorithms. In Chapter 3, we present our experiments on the Wisconsin Breast Cancer Database (WDBC) data set that we used in this thesis. Our experiments and selected classification methods are implemented in Python. The comparison of the selected algorithm by using confusion matrix is given in the last chapter which is devoted to conclusions, discussions about experimental results and further research.

1.1 General Information About Data Mining

Machine Learning (ML) can be described as the intersection of Computer Science and Statistics. Machine learning is an application of artificial intelligence (AI) that examines how computers can learn from data. It uses various types of algorithms to identify compound patterns and estimate outcomes [14, 15]

Definition 1.1.1. A computer program is said to learn from an experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T as measured by P, improves with experience [16].

Taking into consideration of Definition 1.1.1, to build a learning system it is crucial to identify the following three features:

1. Task T,
2. Performance measure P,
3. Training experience E.

In 1950, in the paper [17] Alan Turing proposed to consider the question "Can machines think?" and tried to chase the idea of a machine that could learn. Also he claimed a game called the imitation game that often referred as 'Turing Test'. It is a game played with two players and a human interrogator. The aim of the game for interrogator is to decide which of the players is human and which is the computer according to responses they gave to the questions posed by the interrogator [17]. So this test is an objective way to determine whether machines can think. From this point of view, it can be said that Alan Turing's study had influence on the birth of artificial intelligence.

Artificial intelligence (AI) is a field of computer science that aims to create machines that behave like humans. Over last years, AI takes advantage of machine learning, pattern recognition, neural networks. In daily life self-driving cars, chess-playing computers, voice to text features can be the example of AI systems [18].

Data Mining (DM) is a sub-field of Machine Learning (See Figure 1.1) that deals with analyzing and extracting relevant and useful information from vast databases and data

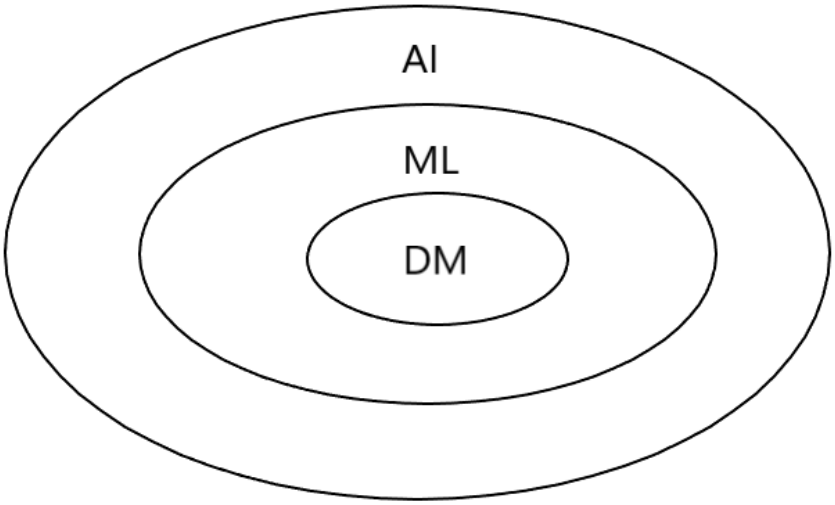


Figure 1.1 : Visual distinction between AL-ML-DM

warehouses using statistics and AI. It has methods, models, algorithms and techniques to acquire the meaningful information. DM is also referred as Knowledge Discovery in Databases (KDD) [19].

In ML, there are two types of techniques: supervised learning and unsupervised learning. Supervised learning is used when we have both input and output data. Unsupervised learning is used when we have only input data but no corresponding output data. So, unsupervised algorithms have to discover the internal structure of the data on their own [14, 20, 21].

Supervised learning is mainly used in classification models and regression models. The main object of a classifier is to place observations into distinct discrete categories rather than predicting quantities. If we have only two possible categories, then such a classifier is called a binary classifier. Some of the most used classification algorithms are k-Nearest-Neighbor, Decision Trees, Naive Bayes, Support Vector Machines. The main goal of regression models is to predict output values by taking into consideration of the relationship between two or more variables. Some of the

commonly used Regression models are Linear Regression, Logistic Regression, and Polynomial Regression [14, 21].

Data mining is also referred as Knowledge Discovery in Databases (KDD) which is a process that composed of data preparation, data selection, data cleansing, incorporating prior knowledge on data sets and interpreting accurate solutions from the observed results as shown in the Figure 1.2 [22]. According to [2], some of basic steps of KDD can be summarized as follows.

1. Developing an understanding of the application domain
2. Creating a target data set
3. Data cleaning and preprocessing
4. Data reduction and projection
5. Matching tile goals of the KDD process (step 1) to particular data mining method
6. Choosing the data mining algorithm(s)
7. Data mining
8. Interpreting mined patterns
9. Consolidating discovered knowledge

Data cleaning identifies noisy and inaccurate data from data set and removes the unwanted records. And one of the important task in data cleaning is to handle with missing values. Also choosing the right data mining algorithms is crucial step. This step includes finding right pattern and determining appropriate parameters depending used algorithm. Data mining step deals the data by employing the needed data mining technique and tools. Data mining is the main step of KDD process and most of works are done in this step. Interpreting mined patterns step includes visualization of methods. [2].

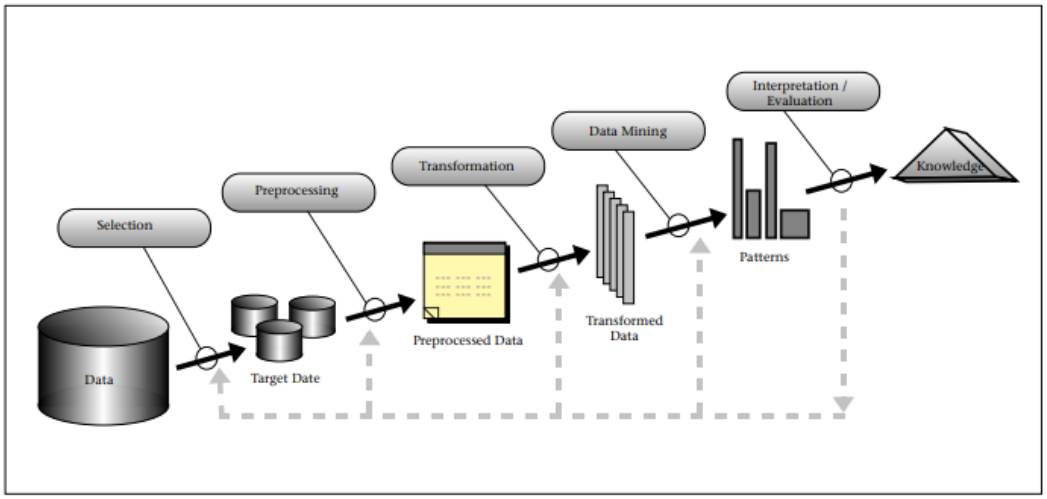


Figure 1.2 : An overview of the steps comprising tile KDD process [2]

1.2 General Information About Breast Cancer

Cancer is a general term for abnormal cells that grows in a certain types of tissue in the body, and then spread to other parts of the body. According to WHO, in 2018 9.6 million people died from cancer. And WHO also claims 18.1 million new cases was seen all over the world in 2018 [23].

Lung and breast cancer are the leading types of cancer in terms of the number of diagnoses, but breast cancer is the most common cancer type among women. It constitutes 11.6% of the total cases which means approximately 2.1 million diagnoses were made in 2018. Death rate of breast cancer is approximately 6.6% of all diagnosed cases. This means approximately 627.000 people did not survive from breast cancer. Also it is estimated that 29.5 million new cancer cases will occur in 2040 if the current survival rates continue to prevail [23, 24].

There are two types of tumors called malignant and benign. Malignant cells are the one that causes cancer.

The risk of being cancer increases with the age. Apart from that risk factors that leads breast cancer are family history, being diagnosed with certain benign (non-cancerous) breast previously, genetic risk factor, alcohol consumption, radiation exposure [25].

There are five stages of breast cancer [26].

- **Stage 0:** The beginning phase and cancer cells do not spread to nearby tissue

- **Stage 1:** It is the early stage of cancer and cancer cells spread in a small area.
- **Stage 2:** A tumor can be 20 to 50 millimeters and some of the lymph nodes can be affected
- **Stage 3:** A tumor is larger than 50 millimeters and lymph nodes are affected
- **Stage 4:** It is the stage that cancer has spread the other parts of the body

Early detection of cancer is the key to decrease the mortality rate. It depends on two strategies screening and early diagnosis. Early diagnosis is possible with awareness and improvement in health care. Especially in low income countries, resources in medical area are limited and healthcare are insufficient so breast cancer is diagnosed in late stages. Thus, survival rate is between 10-40% [27].

Mammography is a very common screening method to detect cancer. Basically, it is an X-ray picture of the breast. It is used to find lumps or other signs of breast cancer to determine the stage of cancer if it is detected [8, 25]. So, especially among women ages 40 to 70, it can be helpful to decrease the number of deaths. But it has some drawbacks. Mammograms sometimes can not find cancer when it is there and patients are exposed radiation to take mammography [28].

1.3 Literature Review

In this section, we discuss previous studies on classification algorithms that were applied in diagnosing breast cancer.

A study by Alam et al. evaluated the performance of Naive Bayes, support vector machine, decision tree and logistic regression on breast cancer data set [13]. In that study, researchers calculated the rate of accuracy without feature selection and accuracy with feature selection with using genetic algorithms. It is observed that logical regression, linear regression and SVM showed better accuracy rates 98.24%, 98.24% and 98.07% respectively [29].

In a research conducted by Saygılı random forest method gave 98.77% classification accuracy in detecting breast cancer [30].

In another work carried out by Polat and Gunes, researchers examined the least square SVM classifier algorithm. They obtained a classification accuracy of 98.53% [31].

Übeyli et al. presented a comparative study using multilayer perceptron neural network (MLPNN), combined neural network (CNN), probabilistic neural network (PNN), recurrent neural network (RNN) and support vector machine (SVM) applied on WDBC data set. The classification accuracy of SVM obtained by their study was better than the other algorithms with the accuracy of 99.54% [32].

Othman and Yau applied and compared many classification and clustering algorithms on breast cancer dataset by using WEKA data mining tool. The algorithms applied in this study include Bayes Network, SVM with Radial Basis Function, Pruned Tree, Single Conjunctive Rule Learner and Nearest Neighbors Algorithm. The 75% of data is used for training and the remaining part is used for testing. The highest accuracy rate belongs to the Bayes network classifier with an accuracy of 89.71%, and it is followed by Radial Basis Function with a percentage of 87.43% and subsequently decision tree with pruning and single conjunctive rule learner. Nearest neighbor has the lowest rate with percentage around 84.57% [33].

Salama et al. presented a comparative study of different data mining techniques applied on different healthcare datasets including WDBC database. In their study, they implemented various data mining methods which included Naive Bayes (NB), Decisions Tree (J48), Instance Based for K-Nearest Neighbor (IBK) and Sequential Minimal Optimization (SMO). They used Waikato Environment for Knowledge Analysis (WEKA) for their analyses. The comparison of methods in their study showed an encouraging level of accuracy. SMO is more accurate than other classifiers with 97.7% accuracy rate. Also the combinations of SMO and MLP, and SMO and IBK give same accuracy [34].

Hazra et al. applied a comparative study on different cancer classification approaches such as Naive Bayes, SVM and Ensemble Algorithms on WDBC dataset by measuring each of the classifier's time complexity. Naive Bayes algorithm has the best accuracy with lowest time complexity with the accuracy rate of 97.4% [35].

A research conducted by Walia et. al. evaluated four different kinds of decision tree algorithms: J48, Reduced Error Pruning Tree (REP Tree), Random Forests, and

Random Trees. They built their models and tested on the WDBC data set taken from UCI [13]. All the experiments in this paper are performed on WEKA. According to their experiment, the maximum accuracy is 95.1% for the random forests method and the minimum accuracy is 93.5% for the REP tree. The time taken to build the classification model is also an essential parameter for this research. Random forest algorithm requires around 0.07 seconds, REP tree requires around 0.01 seconds to build the model.

2. METHODS AND MATERIALS

Classification algorithms estimate the target class for each data instance. For instance, in this study we want to diagnose the patient's condition by using data classification approach. In the following sections, we represent the general concepts about classification methods used in this study. Also, we mention the main aim of normalization and how to calculate accuracy of an algorithm. In the last part we define overfitting which is important obstacle in machine learning.

2.1 Classification Algorithms

Classification algorithms provide a way that can be used to build data models on the training set using a given set of variables. Classification helps to estimate target class for given data instances, and is one of the most commonly used methods of data mining (DM) in medical field. Classification algorithms generally has two phases as shown in the Figure 2.1 [14]:

- **Training phase:** It is also called "the learning stage" where a classification model is built using training instances.
- **Testing phase:** It is also called "the classification stage" where the model is used to predict the unseen data. And classifying the outputs of the system is called "labeling."

Classification algorithms are type of supervised learning because a sample data set which is human derived is used to learn the structure of training data set just like a teacher and a student learning relationship and then algorithm try to classify unseen data [36, 37]. After this step, test data is used to calculate the accuracy of the model, and then we continue to apply the model on new data whether the accuracy is expedient or not [3, 14, 36, 38].

As this study's concern is health care system, patients can be classified as "high risk" or "low risk" patient depending on their disease type by using classification model [16, 39]. So, basically each instance group have to be assigned to only one class.

2.2 K Nearest Neighbor Classification (KNN)

K nearest neighbor algorithm is one of the most popular classification algorithms. It was first discovered in 1950s but it did not get popular until 1960s [40]. It is especially used in pattern recognition, facial recognition, optical character recognition, recommended product in shopping websites [14].

KNN method is extremely powerful,simple and effective method. Given an unlabeled test data, KNN finds k nearest records in the training dataset and then assigns them the most suitable label [41]. Fundamentally, KNN assigns label to a data point from its neighbors considering the majority of closest neighbor points. To start processing, algorithm needs to see the test tuple and then perform generalization to classify the data by finding the closest the stored training tuples. Therefore KNN algorithm called as lazy learner and does less work until it starts to label the unseen data [14, 42].

Depending the k value the class of the instance changes. In the Figure 2.4, there are two class which is green square and blue hexagon and when k value equals to 3, it will classified as blue hexagon. When k is 5, the number of green squares constitutes the majority of nearest neighbors [6]. There are two important issues in KNN algorithm. First issue is the choosing right k which means how many neighbours will be chosen for each tuples. It has significant importance on performance of algorithm. Large k values cause to ignore small patterns. Another important case is to calculate the distance between test instance and its neighbours [43].

2.2.1 Distance Measure

KNN algorithm builds a model by considering the majority vote of neighboring data points of a given data point. Therefore distance measurement is the key to find the closest neighbors of a given point. There are many types of distance measures but the most popular ones are the Euclidean, the Minkowski and the Manhattan distances. However, these distances are only valid for continuous variables. The Euclidean distance is more commonly used than the Minkowski and the Manhattan distances.

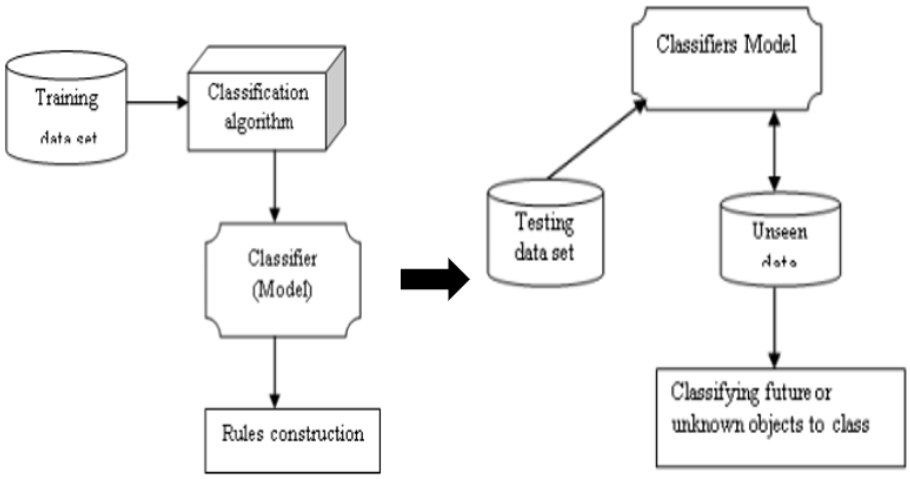


Figure 2.1 : Classification Phases (Modified from [3]).

Let $X_i = (X_1, X_2, ..., X_n)$ and $Y_i = (Y_1, Y_2, ..., Y_n)$ represent feature vectors where n is the dimension of feature space. Then Euclidean distance can be calculated as

$$dist_{Euclidean}(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \tag{2.1}$$

The Manhattan distance formula that calculates the differences between coordinates of pair data points can be written as

$$dist_{Manhattan}(X,Y) = \sum_{i=1}^{n} \mid (X_i - Y_i) \mid \tag{2.2}$$

The Minkowski distance measure is given by following formula

$$dist_{Minkowski}(X,Y) = \left(\sum_{i=1}^{n} \mid X_i - Y_i \mid^{p}\right)^{1/p} \tag{2.3}$$

for a fixed $p \in (0, \infty)$.

2.3 Support Vector Machine (SVM)

Support Vector Machine algorithms was introduced in 1963 by Vapnik in [44]. SVM is a supervised learning model, i.e. given instances are labelled with a label depending on a sample of data with known labels. The aim of SVM is to find a hyperplane, or a decision surface, that divides a data set into two classes as shown in the Figure 2.2 [45]. Data points that determine the hyperplane are called "support vectors" [4, 46].

It is possible that there are many possible hyperplanes that separate data points into two classes. So, the main objective is to find a hyperplane in an n-dimensional space that has "the maximum margin." A margin is the distance between two classes which

depends on the dimension of the hyperplane which in turn is related to number of input features. SVM uses a mechanism called "kernels" to calculate distance between two classes in non-linear models. The choice of kernel affects the performance of built SVM model. To decide the right kernel, the best way is to try kernel functions on data set.

2.3.1 Linear support vector machine

In the linear SVM's the decision plane perfectly separates the classes. In other words, we can draw a straight line between two classes labeled as +1 and -1 as shown in the Figure 2.2 for data points with 2 features.

In short, SVM classifies input vectors into two classes as in

$$S = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n) |\ (x_i, y_i) \in \mathbb{R}^n \times \{+1, -1\}\} \tag{2.4}$$

where n is the number of features, and the labels y_i are either +1 or -1. By using Equation 2.4, we can write the equation of a hyperplane.

Let H_1 and H_2 be the two parallel planes

$$\begin{aligned} H_1 &: w \cdot x_i + b = +1 \quad \text{where } y_i = +1 \\ H_2 &: w \cdot x_i + b = -1 \quad \text{where } y_i = -1 \end{aligned} \tag{2.5}$$

where w is the normal vector to the hyperplane, x_i is the input vector and b is the scalar.

By using Equation 2.5, we can calculate the distance between H_1 and H_2 by subtracting them from each other.

$$\begin{aligned} w \cdot x_1 + b &= +1 \\ w \cdot x_2 + b &= -1 \\ = w \cdot (x_1 - x_2) + b &= 2 \\ = \left(\frac{w}{||w||} \cdot (x_1 - x_2) + b\right) = \frac{2}{||w||} &= \frac{2}{\sqrt{w \cdot w}} \end{aligned} \tag{2.6}$$

From the Equation 2.6, the distance between H_1 and H_2 is $\frac{2}{||w||}$.

Let H_0 be the median between H_1 and H_2 planes. Then H_0 can be written as $w \cdot x_0 + b = 0$. The distance between H_0 and H_1 is d^+ and it is the is the shortest distance from positive input to the hyperplane and can be calculated as $\frac{1}{||w||}$ based on 2.6. d^- is the

distance between H_0 and H_2 and it is the shortest distance from negative input to the hyperplane and can be shown as $\frac{1}{||w||}$.

2.3.2 Non-linear support vector machine

Some data sets do not allow linear separation. In such cases, we use kernel based methods to handle such linearly inseparable data. The main goal is to map the original training data in a different higher dimensional space so that in the new space the data now is linearly separable as shown in Figure 2.3 [5]. In this new space, the image is linearly separable and instead of building curve, all we have to do is to discover a hyperplane that separates two classes smoothly. This is called "the kernel trick" [47, 48].

$$K(x_i, y_j) = \phi(x_i)\phi(y_j) \tag{2.7}$$

According to Equation 2.7, let K be the kernel function and $\phi(x_i)$, $\phi(y_j)$ be the dot product's form of the training tuples. Training tuples appears only in the form of dot products. However, in higher dimensional space, when searching for linear SVM , there is no need to compute dot product of data set. Dot product of data tuples is equal to kernel function as shown in the Equation 2.7. So we can use $K(x_i, y_j)$ instead of $\phi(x_i)\phi(y_j)$ values [14].

There are many kernel functions. Different choices of the kernel function affects the performance of an SVM classifier. The most important and popular kernels are linear, polynomial and Gaussian Radial Basis Function (RBF) [14].

$$K_{linear}(x, y) = x^T y \tag{2.8}$$

$$K_{polynomial}(x, y) = (x^T y + 1)^p \tag{2.9}$$

where p is the degree of a polynomial.

$$K_{RBF}(x, y) = \exp\left(\frac{-\|x - y\|^2}{2\sigma^2}\right) \tag{2.10}$$

where σ is the Gaussian parameter.

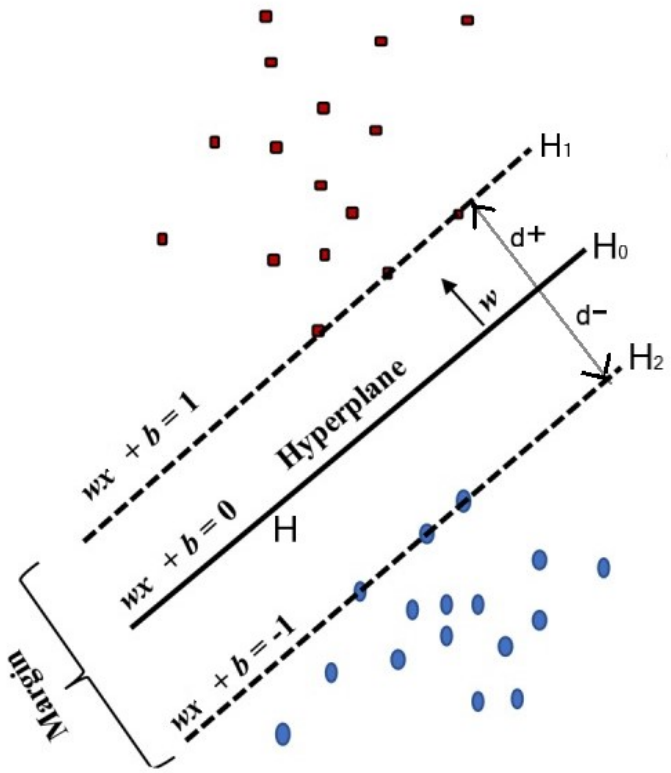


Figure 2.2 : Linear SVM model (Modified from [4]).

2.4 Naive Bayes (NB)

2.4.1 Bayes Theorem

Bayes Theorem is a mathematical formula that finds the conditional probabilities by using given data. It is discovered by mathematician Thomas Bayes in the 18th century.

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \tag{2.11}$$

where X is data tuple on a given class C. $P(C|X)$ indicates a conditional probability and means that C happens given that X happens.

From the equation 2.11, it can be said that

- $P(C|X)$ is the posterior probability of C conditioned on X
- P(C) is the prior probability of class
- P(X|C) is) is the posterior probability of X conditioned on C

- P(X) is the prior probability of predictor

2.4.2 Naive Bayes

Naive Bayes calculates the probability of a given data and finds its class. The most important point is that variables are independent from each other. So the power of this method is that there is no inter-dependency among different feature variables. Generally used for application of Text classification, Spam Filtering, Sentiment Analysis [49].

NB classifier can perform well with large number of data points. The advantage of this method is that NB can work with small amount of training data to predict the probability of a given data. It can be said that naive bayes is a fast, effective and accurate classification algorithm. The steps of NB is that [14, 42, 49, 50]:

✓ **Step 1:** Calculate the prior probability for given class labels

✓ **Step 2:** Find likelihood probability with each attribute for each class

✓ **Step 3:** Calculate posterior probability by using bayes formula

✓ **Step 4:** Find class that has the highest probability

2.4.3 Gaussian Naive Bayes classifier

A Gaussian distribution is also called Normal distribution. It is best if we used in cases when all features are continuous. The advantage of this distribution is that when you work with Gaussian, you do not need to calculate the distribution of data but mean and standard deviation of training data [51].

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}}\exp\Big(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\Big) \tag{2.12}$$

where, μ is the population mean, σ is the standard deviation and σ^2 is the variance.

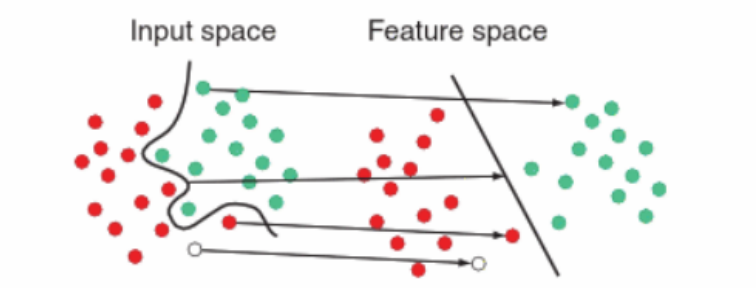


Figure 2.3 : Non-linear SVM model [5].

2.5 An Overview of Classification Methods

It is crucial to measure the performance and accuracy rate of classifier after producing unseen data. So there are some tricks that make the algorithm better and help classifier to avoid inevitable errors. Normalization, accuracy calculation and overfitting is general issue for classification algorithms.

2.5.1 Normalization

It is mentioned that KNN algorithm calculates distance measure between test tuples and training tuples. To measure distance based on every instance brings some consequences because attribute values can be nominal. The objective of doing normalization is that while finding the nearest neighbor to ensure that all attributes have same contribution balance. For instance, if the data set has sum of annual average amount and age attributes, comparison of amount and sum values, it is clear that sum value has large impact over finding the closest neighbour. So by normalizing attributes of data set, their values will be scaled in the range [42, 52].

Normalization simply eliminates the dominated features and reduce them to the same level. There are lots of methods to normalize data. In this study, Z score normalization is used to scale the attributes. The following formula subtracts the mean value of feature x and divides by the standard deviation. The result is called z score. In this method z scores will be scaled in unbounded range of positive and negative numbers [41].

$$X_{new} = \frac{X - \mu}{\sigma} = \frac{X - Mean(X)}{Std(X)} \tag{2.13}$$

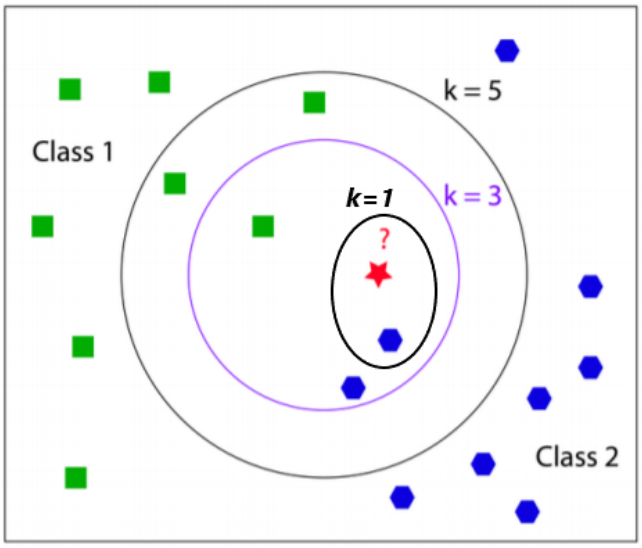


Figure 2.4 : 1-NN rule : it will classified as blue hexagon.
3-NN rule : it will classified as blue hexagon.
5-NN rule : it will classified as green square.
(Modified from [6]).

where Mean(X) is the sum of the all attribute values of X and Std(X) standard deviation of all values of X.

The other main issue is that every attribute can have different importance. Considering above example, annual amount feature can be more important than age feature. If attributes have different contributions, then in KNN algorithms Euclidean distance formula will be changed to

$$\sqrt{w_1(a_1-b_1)^2+w_2(a_2-b_2)^2+\ldots+w_n(a_n-b_n)^2} \tag{2.14}$$

where $w_1, w_2, \ldots, w_n$ are the weights. After scaling the weight values, the sum will be equal to 1 [42, 53].

2.5.2 Accuracy calculation

Accuracy is a need to measure how an algorithm perform and find the percentage of correctly estimated or missed data in built model. Accuracy measures for binary classification can be expressed in four metrics [54, 55]:

✓ TP or true positive, the number of correctly classified positive data

✓ TN or true negative, the number of correctly classified negative data

✓ FP or false positive, the number of missed classified positive data

✓ FN or false positive, the number of missed classified negative data

The sum of all these four values give the total number of instances to classify.

It can be visualized by using 2x2 matrix called contingency matrix or confusion matrix as shown in the Figure 2.5. Based on this matrix, it can be computed Recall, Precision, Specificity, Accuracy and most significantly AUC-ROC Curve which are the metrics to measure the classifier's performance. Accuracy metric indicates how well the classifier or algorithm can guess the unseen data. The formula to calculate Accuracy:

$$\begin{aligned} Accuracy &= \frac{NumberOfCorrectPredictions}{TotalNumberOfPredictions} \\ &= \frac{TP+TN}{TP+TN+FP+FN} \end{aligned} \tag{2.15}$$

Recall is only interested in correctly identified positive cases. In other words, it shows that which percentage of positive data predicted accurately formulated as:

$$Recall = \frac{TP}{TP+FN} \tag{2.16}$$

Precision is the answer of which proportion of identified positive data is actually correct and can be calculated as:

$$Precision = \frac{TP}{TP+FP} \tag{2.17}$$

Error rate is the calculation of number of incorrect predictions divided by total number of data set.

$$\begin{aligned} ErrorRate &= \frac{NumberOfIncorrectPredictions}{TotalNumberOfPredictions} \\ &= \frac{FP+FN}{TP+TN+FP+FN} \end{aligned} \tag{2.18}$$

	Predicted YES	Predicted NO
Actual YES	TP	FN
Actual NO	FP	TN

Figure 2.5 : Confusion Matrix

Performance measurement is essential subject for classification problems. These four metrics are also used to compare risks and gains of model. In this study, risk, the number of false negative (FN), refers to incorrectly categorize cancerous patient as a not cancerous patient and gain, the number of true positive, refers to accurate prediction of cancerous patients [14].

Receiver operating characteristic curves (ROC) allow to display a graph that can be useful for comparing classification model's performance. A ROC graph is a two dimensional plot that has false positive rate (FPR) on its x-axis and true positive rate (TPR), also called sensitivity or recall, on its y-axis as shown in the Figure 2.6 [56]. Sensitivity can be calculated as:

$$Sensitivity = \frac{FP}{TN + FP} \tag{2.19}$$

Area under the curve (AUC) expresses how well model can distinguish between classes. If the AUC is higher, it represents that the separability of model is better. It means that the closer the AUC is to 0.5, the less accurate the model.As shown in Figure 2.7, when AUC=1, it means that separation of two classes, in this case 0s and 1s, is perfectly distinguished and it is an ideal situation. However, when AUC decreases in value, the curve is approaching to the diagonal. When AUC fall under the value of 0.5, the model gets poorer and inaccurate estimation rate increases [7].

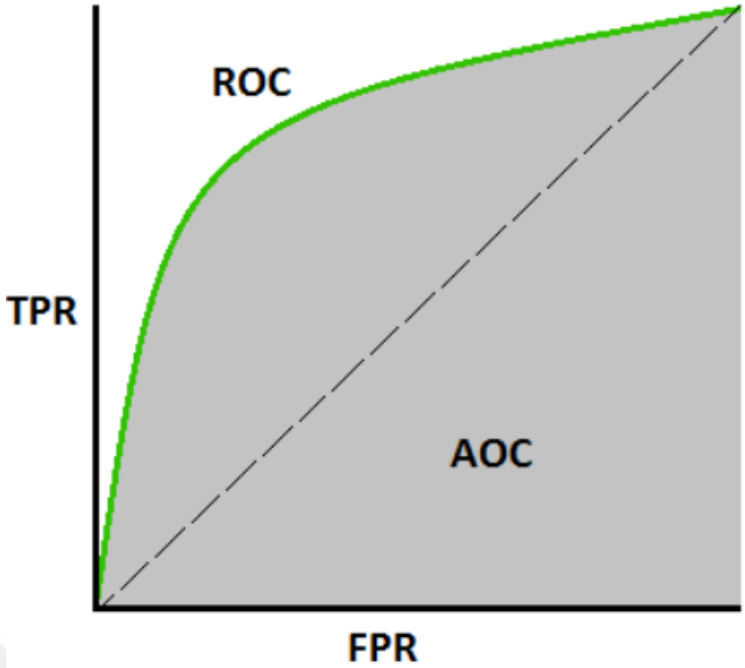
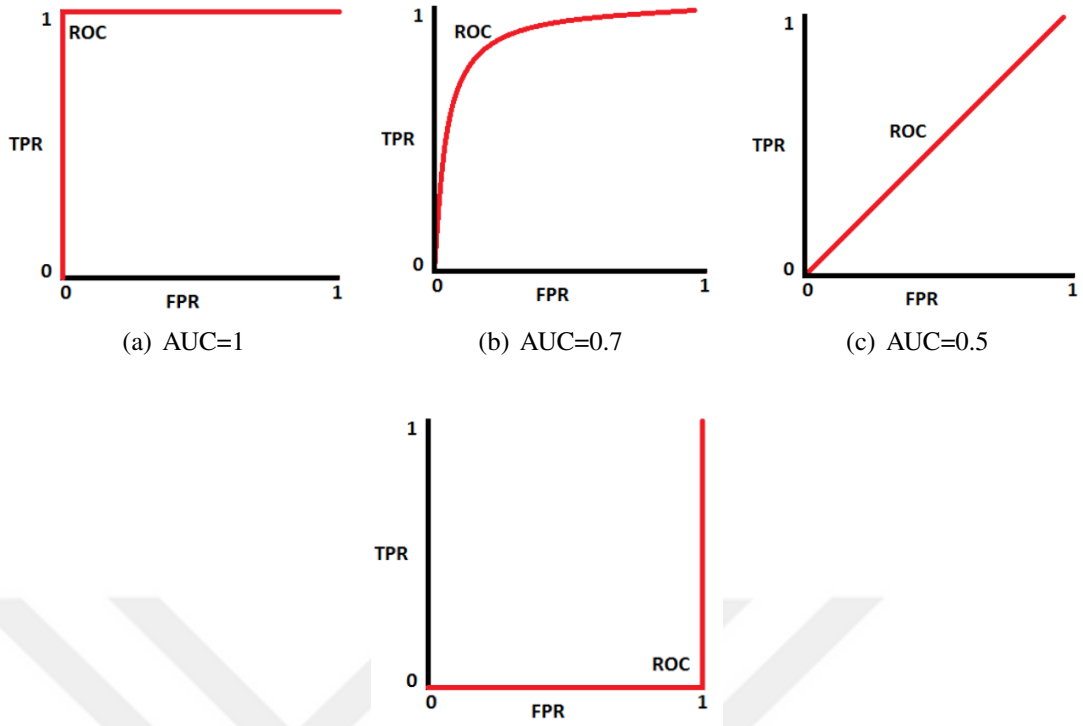


Figure 2.6 : AUC-ROC Curve [7].

2.5.3 Overfitting

Overfitting is the most important obstacle in machine learning model. It literally indicates that the model can only be able to classify training data like memorizing training data that when it faces new data sets, the model does not know how to respond and it leads model to rise misclassification rates [57]. Although, at first, over fitted model seems attractive because of the high accuracy rate and fitting exactly, after large amounts of new data is acquainted to the model, algorithm is demolished and errors will come to light.

The best way to prevent overfitting is to maintain sufficient training data. Also there is a very popular method called cross-validation or k-fold cross. Cross-validation is to partition the training data set that is used in the beginning and split this data into multiple subsets k times, called folds [58, 59]. The first one called validation set or test set and the remaining groups is called training set and use to fit a model. It can be said that high variance which is the changes in model based on training data set and low bias that means to ignore training data are a sign of overfitting [60].



(a) AUC=1 (b) AUC=0.7 (c) AUC=0.5

(d) AUC=0.0

Figure 2.7 : Estimation of Model Performance Based on AUC-ROC Curve [7]

3. EXPERIMENTAL RESULTS

In this section, KNN, SVM and NB classification algorithms are applied on breast cancer data set [13]. The main aim is to build a classification model that allow us to compare the performance of these two methods in terms of their classification accuracy on the test data. The data set used in this work is Wisconsin Diagnostic Breast Cancer (WDBC) data set [1].

3.1 Data Set Description

The data set used in this thesis is publicly available in the University of California at Irvine (UCI) Machine Learning Repository and was created by Dr. William H. Wolberg, physician at the University of Wisconsin Hospital at Madison, Wisconsin, USA. Wisconsin Diagnostic Breast Cancer data set contains two classes and 569 examples. The first class includes the non-cancer patients counted as 357 and the second class consists of 212 cancer patients [13].

The classification methods to which the WDBC data set is applied aim to use 31 columns to estimate the values of the diagnosis column. The WDBC data set attributes and their descriptions is shown in the Table 3.1.

In this study, the WDBC data set is divided into training and test data set, two third $\left(\frac{2}{3}^{rd}\right)$ of data set instances are for training the model which is called training set, while the rest of the WDBC data set records $\left(\frac{1}{3}^{rd}\right)$ are for testing the model which is called test set. In other words, 381 records will be used to train the model and 188 of the data set records to measure the validation of the classification method.

3.1.1 Attributes Description

WDBC data set consists of ID attribute, diagnosis attribute which is bounded variable for this case and ten real-valued features are computed for each cell nucleus which is unbounded variables. All feature values are recorded with four significant digits. There are no missing attribute values in WDBC data set. To compute these features

Table 3.1 : Data set description [1].

No.	Attributes	Attribute Type
1	ID	Numeric
2	Diagnosis (B=benign, M=malignant)	Discrete (Binary outcome)
3	Radius	Numeric
4	Texture	Numeric
5	Perimeter	Numeric
6	Area	Numeric
7	Smoothness	Numeric
8	Compactness	Numeric
9	Concavity	Numeric
10	Concave Points	Numeric
11	Symmetry	Numeric
12	Fractal Dimension	Numeric

used digitized image of a fine needle aspirate (FNA) of a breast mass. These attributes and their types can be summarized as in Table 3.1.

In WDBC data set except diagnosis attribute every other attributes are numerical data. Diagnosis attribute represented as categorical number can take two values $\{\text{B,M}\}$, B stand for benign and M is stand for malignant so in the experiment it is represented as $\{0, 1\}$.

Radius is the mean of distances from center to points on the perimeter. Texture is the standard deviation of gray-scale values. Smoothness is the local variation in radius lengths. Compactness calculated as $\frac{Perimeter^2}{Area} - 1.0$. Concavity means severity of the concave portions of the contour. Concave points is represented as number of concave portions of the contour. Fractal dimension is "coastline approximation" -1. These features describe the characteristics of the cell nuclei present in the image of a breast mass [1].

The mean, standard error (SE), and worst or largest (mean of three largest values) values of these features were also computed by using every image and based on these computations generated 30 features. In other words the numerical data is divided into three parts. The first part consists of mean values of 10 basic attributes which is explained in the Table 3.1, the second part indicates the SE calculations and the last part is the worst values of this 10 parameters. There is a column named as "Unnamed: 32". This attribute consists of NaN values so in this thesis this column is omitted [13].

3.1.2 Correlation Heat Map

It calculates the correlations among all the attributes, including the correlation of each attributes with itself. As shown in the Figure 3.1 correlation scores go from -1 which is the perfect negative correlation to 1 which is the perfect correlation. When the color gets darker it approaches to -1. The correlation of the any attributes with itself perfect, so it will be seen as +1 all along the diagonal. It means that the correlation heat map is symmetric.

3.2 Implementation of Methods

In this part, some of the classification methods are applied on WDBC data set. All experiments and code implementation parts are implemented by using Python. In this thesis work, KNN, SVM and NB classification algorithms' results and accuracy rates are compared. The training data set will be the input of this experiment and we expect to generate the most accurate and the best trained model that can predict and classify unseen data. In the following subsections, it will be shown the performance of these classification methods.

3.2.1 Implementation of KNN algorithm on breast cancer data set

While using KNN algorithm on any data set, it is important to decide the right k (nearest number) value. KNN algorithm basically calculates the distance between all data points and then finds the k points that are closest. The main set consists of the great majority of the surrounding points. To find the best accuracy is related to find the right k value.

The Figure 3.2 shows the k values and their accuracy rates. According to the Figure 3.2, k=9 gives the higher accuracy than the other k values. In other words minimum error rate give us the right k value.

The Table 3.2 shows that KNN accuracy and error rate with different k values on breast cancer data set. The lowest performance of KNN algorithm was scored when k=2. As shown in the Table 3.2 the accuracy is 94.2% and the error rate is 5.8% when k=2. The highest performance of KNN algorithm was scored when k=9 and the Table 3.2 shows that the accuracy rate is 96.4% and the error rate is 3.6%.

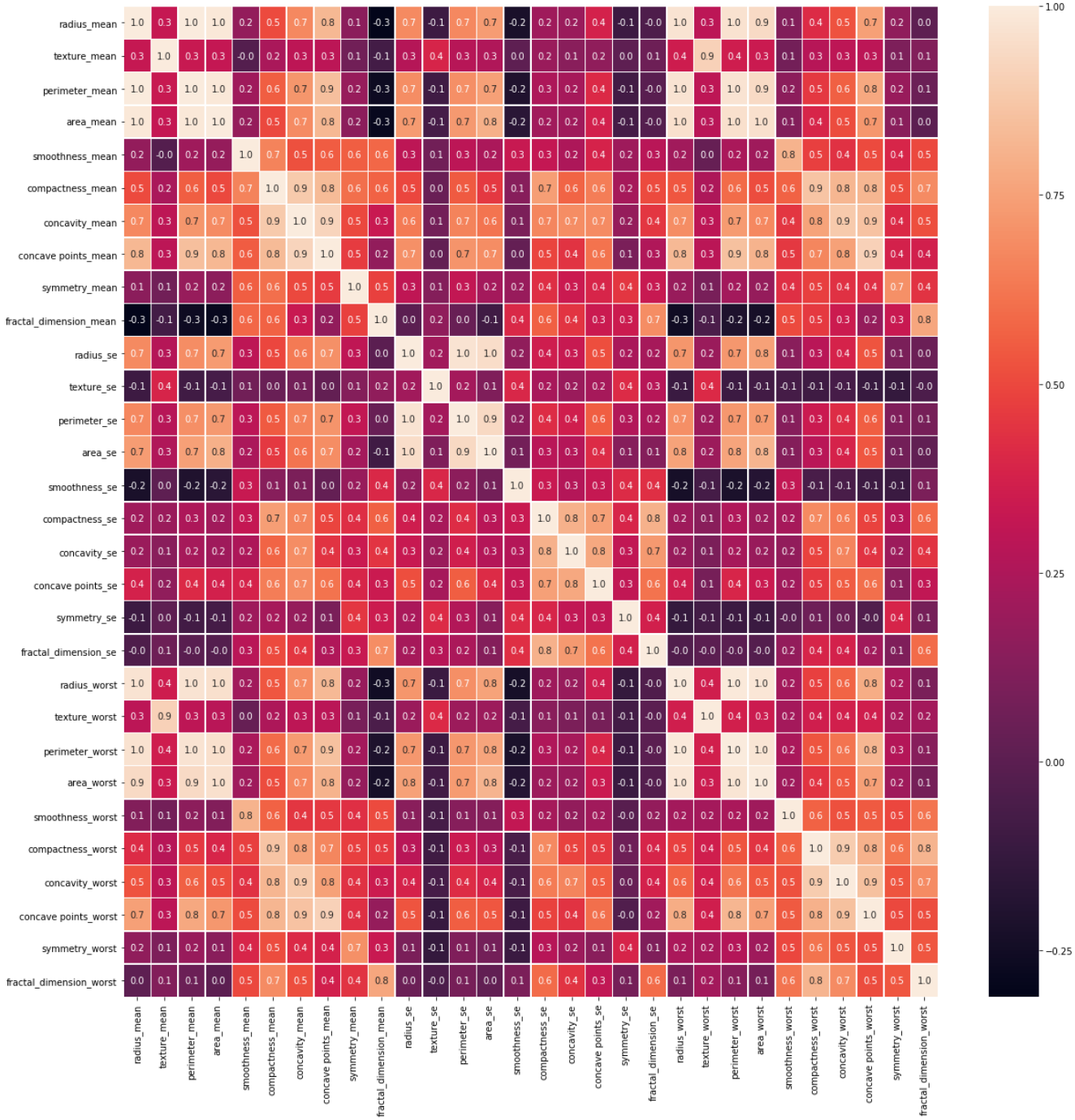


Figure 3.1 : Correlation heat map.

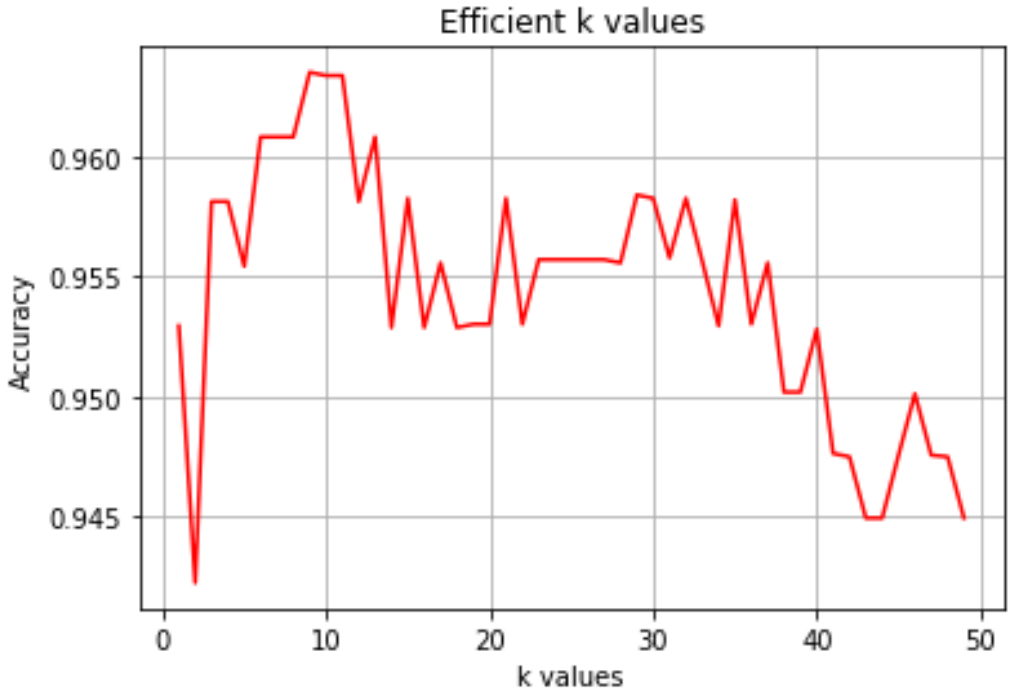


Figure 3.2 : K values accuracy rate.

To see the performance of KNN classification model on WDBC data set, it is constructed a confusion matrix as shown in the Table 3.3.

As it is mentioned before, 381 records used to train the algorithms and 188 records used to test the classification algorithms and find the accuracy rate. So the Table 3.3 indicates the summary of prediction results of testing data. The number of correct and incorrect predictions are summarized with count values.

Table 3.2 : KNN accuracy with different k values

K values	Accuracy	Error Rate
2	94.2%	5.8%
8	96.1%	3.9%
9	96.4%	3.6%
10	96.3%	3.7%

There are two possible predicted classes: “yes” and “no”. In this case “yes” means patients have disease and “no” means they do not have the disease. Also “n” indicates the total number of patients that were being tested for the presence of that disease. According to the Table 3.3 out of those 188 cases, the algorithm predicted “yes” 123

times and “no” 65 times. In reality, 121 patients have the disease and 67 patients do not.

Table 3.3 : KNN confusion matrix

n =188	**Predicted YES**	**Predicted NO**	
Actual YES	117	4	121
Actual NO	6	61	67
	123	65	

As a result of KNN confusion matrix, 4 patients diagnosed as non-cancer patient but actually they have cancer. Also 6 patients diagnosed as cancer patient but in reality, they do not have cancer. Out of 188 patients, the algorithm predicted 178 of them correctly and 10 of them are misestimated.

Briefly, KNN algorithm has the best performance when k=9 with 96.4% accuracy rate and 3.6% error rate, as shown in the Table 3.4.

Table 3.4 : KNN performance on breast cancer data set

Algorithm	Accuracy	Error Rate
KNN (k=9)	96.4%	3.6%

3.2.2 Implementation of SVM algorithm on breast cancer data set

SVM creates the hyper plane $w \cdot x + b = 0$ that have the largest margin in a high-dimensional space to separate WDBC data set into two different classes i.e. malignant and benign.

The margin between two classes is the longest distance between closest data points as shown in the Figure 3.3. And w is the normal to the hyper plane which is called weight vector and b is the bias. According to Figure 3.3 $(w \cdot x + b = 1)$ for positive

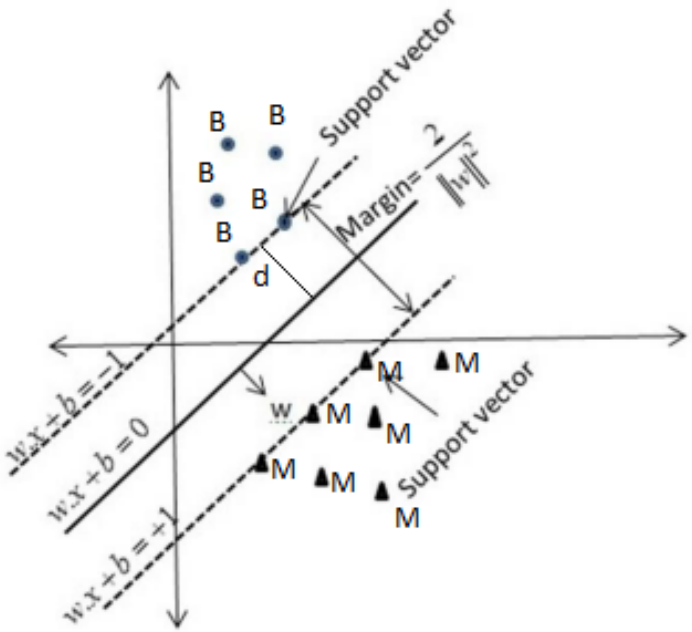


Figure 3.3 : SVM margin separating data (Modified from [4])

class which is malignant in this case and $(w \cdot x + b = -1)$ for negative class which is benign in this case. The main aim of this algorithm is to separate the two classes using kernel function and find right classifier that will work well on unseen examples. The right choice of kernel is the key determines the performance of SVM algorithm.

In this thesis, three basic kernels which is linear, radial basis and polynomial kernels used to train WDBC data set. The best performance is calculated by using radial basis function as shown in the Table 3.5.

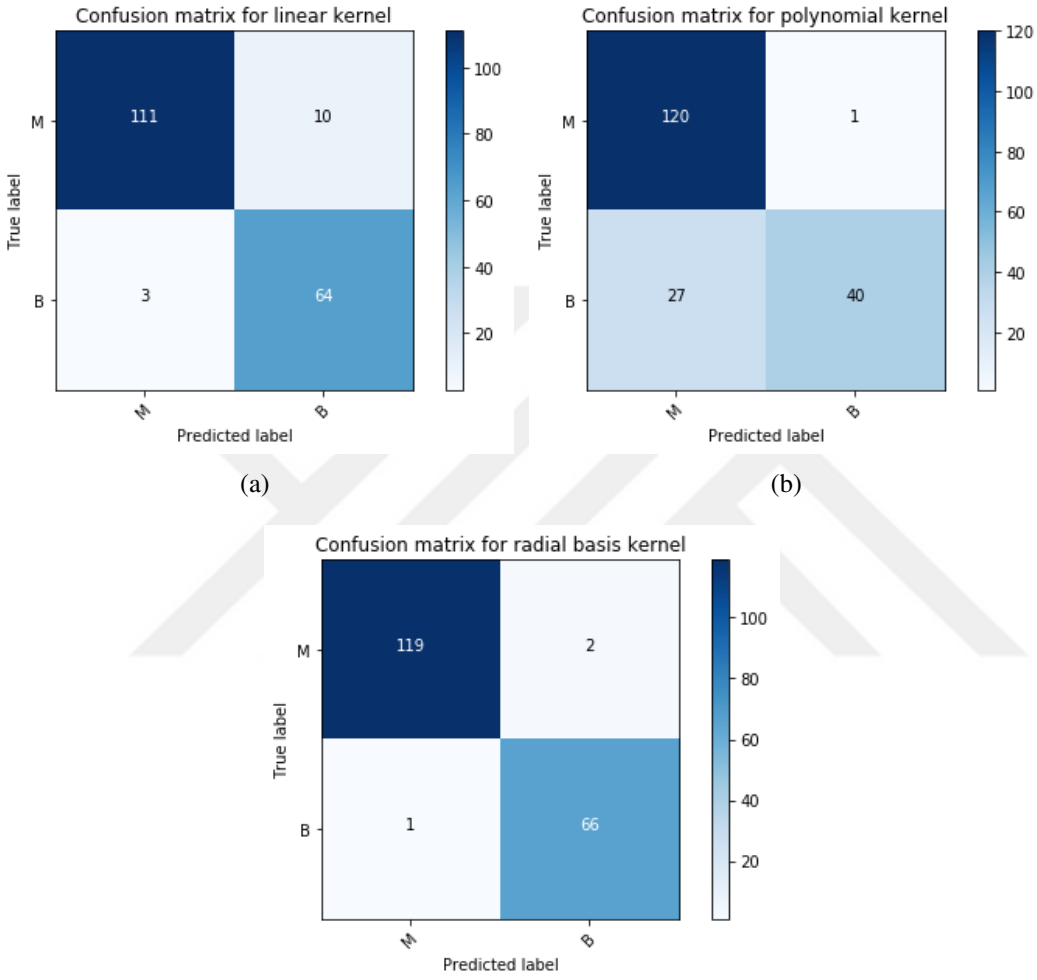
Table 3.5 : SVM kernel performance comparison on breast cancer data set

Kernel	Accuracy	Error Rate
Linear Kernel	93.1%	6.9%
Polynomial Kernel	85.1%	14.9%
RBF Kernel	98.4%	1.6%

Linear kernel confusion matrix shows that 10 of the patients who has malignant cell estimated as non-cancer patient and 3 of the non-cancer patients estimated as cancer patient. Linear kernel has the 93.1% accuracy rate and 6.9% error rate as shown in the Table 3.5. Confusion matrix for polynomial kernel indicates that only 1 of the

malignant patient misestimated and 27 of the benign patients estimated as malignant patient so polynomial kernel has 85.1% accuracy rate and 14.9% error rate.

According to the figure 3.4, RBF kernel has the most accurate estimation that 2 of the cancer patients marked as non-cancer and only 1 of the non-cancer patient estimated wrongly. RBF has the 98.4% accuracy rate and 1.6% error rate.



(a) (b)

(c)

Figure 3.4 : SVM confusion matrix for types of kernel

While using SVM the other important issue is to choose the right C and gamma which are the 2 important parameters. A low C which is the bearable error term creates a small-margin hyperplane and the larger value of C creates a larger-margin hyperplane. SVM's main aim is to find maximum margin out of all possible hyperplanes. A higher value of gamma will fit the training data set exactly which can cause the over-fitting. The optimal value of C and gamma for WDBC data set is shown in the Figure 3.5.

The Figure 3.5 implicates that the most yellow area has the highest estimation rate, best fitted C value is the 5 and gamma value is 0.01. While calculating the accuracy and error rate shown as above in the Table 3.5, it was taken these values into account.

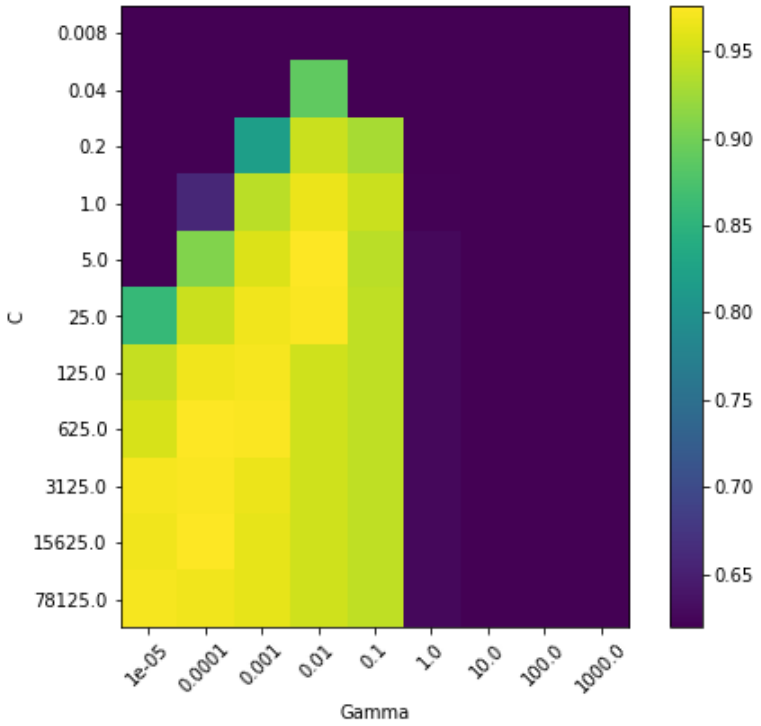


Figure 3.5 : Selecting hyper-parameter C and gamma in SVM

3.2.3 Implementation of NB algorithm on breast cancer data set

Naive Bayes algorithm calculates the probability of being malignant or benign of a cell using given data. NB is known as a fast and efficient classification method. In this thesis, NB algorithm is calculated based on Gaussian distribution in python.

According to confusion matrix of NB from Figure 3.6, 13 patients diagnosed as non-cancer but in reality they have cancer. Besides, 3 patients had benign cancer cells but were actually measured as having cancer. Therefore, it can be said that 16 patients were misestimated.

Table 3.6 : NB performance on breast cancer data set

Algorithm	Accuracy	Error Rate
NB	91.5%	8.5%

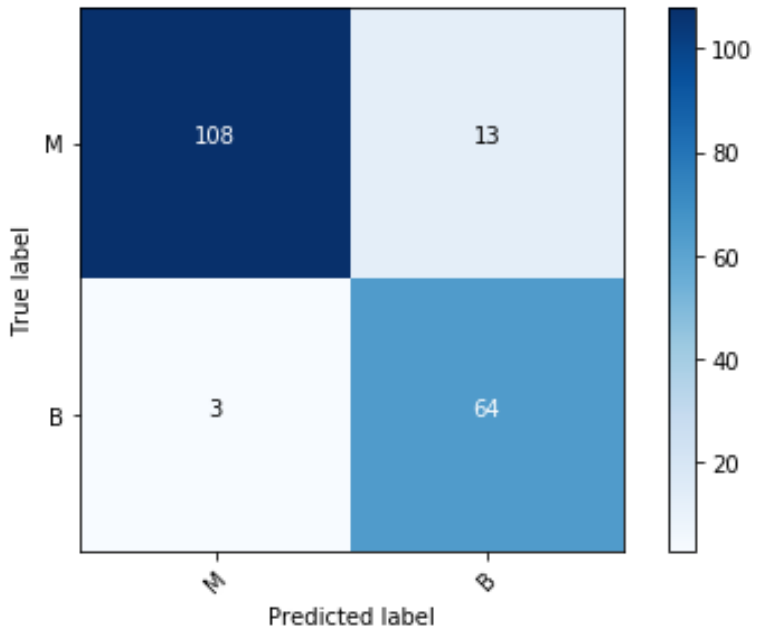


Figure 3.6 : Naive bayes confusion matrix

From Table 3.6, the accuracy of naive bayes algorithm can be computed as 91.5%. So far it is the lowest estimation that we get and error rate of this algrithm is 8.5%.

3.3 Results

The main objective of this thesis is to compare classification performance of classification algorithms that were applied in this study. Taking into consideration of the kernel types and deciding to use RBF kernel since it has the best score, the SVM algorithm has the best accuracy rate comparing to KNN and NB algorithms and it reached 98.4% accuracy rate. KNN algorithm has the 96.4% accuracy rate which is not the worst score but comparing with SVM algorithm KNN is the second position. Also NB algorithm has the 91.5% accuracy rate and has the lowest accuracy rate comparing with applied algorithms. The Table 3.7 summarize the results obtained from applying different classification methods on Wisconsin Diagnostic Breast Cancer (WDBC) data set.

As shown in the Figure 3.7, SVM algorithm has better result to distinguish between positive and negative classes. It is the closest curve to the ideal situation. NB has resulted the least performance among the three models.

Table 3.7 : Comparison of classification methods

Algorithm	Accuracy	Error rate
KNN	96.4%	3.6%
SVM (RBF)	98.4%	1.6%
NB	91.5%	8.5%

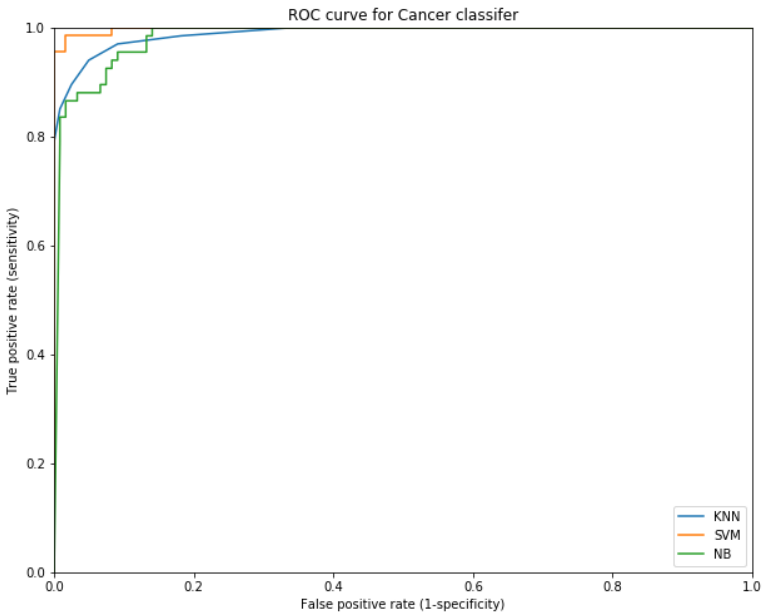


Figure 3.7 : ROC Curve

4. CONCLUSIONS AND RECOMMENDATIONS

Current research to diagnose diseases in medical field mainly focuses on using data mining algorithms in an efficient and accurate way. By applying classification algorithms on medical data set, it is possible to capture the uncertainty and variability among patients on certain diseases.

4.1 Conclusion

In this thesis, firstly it was mentioned that machine learning and data mining concepts, techniques, its applications and their relationships with artificial intelligence. Also it was given a common knowledge about the breast cancer abiding by the thesis topic and shared survey results about worldwide risk of this disease.

In second part, it was presented and explained the background information of classification methods which are conducted in this thesis. Support Vector Machine, K Nearest Neighbor and Naive Bayes classification algorithms are applied and evaluated taking into consideration of their performances and usefulness in every aspect.

In last part which the experimental works were done, the WDBC data set was explained. Then KNN, SVM and NB methods were implemented on using breast cancer data set by using Python. Each of these two methods was carried out individually on WDBC data set and tried the find the best and optimal parameter values by taking into consideration of algorithm's specialty and accuracy rate to implement them. In SVM implementation, three most common kernel function was chosen and compared with each other. Distance metrics with nine nearest neighbor was found as optimal KNN algorithm parameters.

As a result of the coding part, the number of accurate and inaccurate predictions were come to the light with the help of using confusion matrix. Also the accuracy rate and

error rate of each method were calculated and compared. The SVM classifier with Radial Basis Kernel gave the highest success rate compared to NB and KNN classifier.

4.2 Future Works

This thesis achieve and present the performance of the most used classification algorithms and comparison of them.

Future researches and ideas to improve the performance of this work include the following

- Extending data set with increasing size of data would be useful to see performance of algorithms and how algorithms would react in such case.
- Since constituting data set with lots of attributes like WDBC is not an easy work, it would be emergent to implement this data set on other algorithms that used for image recognition
- To perform this algorithm by using same type of algorithm on different breast cancer data sets apart from WDBC data set.

REFERENCES

[1] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic), (Date accessed: March 2019).

[2] **M. Fayyad, U.**, **Piatetsky-Shapiro, G. and Smyth, P.** (1996). From Data Mining to Knowledge Discovery in Databases, *AI Magazine*, *17*, 37–54.

[3] **Annasaheb B., A. and Verma K., V.** (2016). Data Mining Classification Techniques: A Recent Survey, *International Journal of Emerging Technologies in Engineering Research (IJETER),*, *4*.

[4] **Huang, S.**, **Cai, N.**, **Pacheco, P.P.**, **Narrandes, S.**, **Wang, Y. and Xu, W.** (2017). Applications of Support Vector Machine (SVM) Learning in Cancer Genomics, *Cancer Genomics Proteomics*, *15*, 41 – 51.

[5] **Nisbet, R. and et al.**, (2009), Handbook of Statistical Analysis and Data Mining Applications, Elsevier Science Technology.

[6] **Cover T., M. and Hart P., E.**, (2018), Nearest Neighbor Pattern Classification.

[7] https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5, (Date accessed: April 2019).

[8] http://www.imaginis.com/general-information-on-breast-cancer/what-is-breast-cancer-2, (Date accessed: March 2019).

[9] **American Cancer Society** (2017). Breast Cancer Facts Figures 2017-2018, *A Cancer Journal for Clinicians*.

[10] **Soliman, O.S. and AbuElhamed, E.** (2014). Classification of Breast Cancer using Differential Evolution and Least Squares Support Vector Machine, *International Journal of Emerging Trends Technology in Computer Science (IJETTCS)*, *3*, 155–161.

[11] **Ferlay, J.**, **Colombet, M.**, **Soerjomataram, I. and et. al.** (2018). Global and Regional Estimates of the Incidence and Mortality for 38 Cancers: GLOBOCAN 2018, *International Agency for Research on Cancer/World Health Organization*.

[12] **Bray, F.**, **Ferlay, J.**, **Soerjomataram, I.**, **Siegel, R.L.**, **Torre, L.A. and Jemal, A.** (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *A Cancer Journal for Clinicians*, *68*, 394–424.

[13] **Dua, D. and Graff, C.**, (2017), UCI Machine Learning Repository, http://archive.ics.uci.edu/ml.

[14] **Han, J. and Kamber, M.**, (2006), Data Mining: Concepts And Techniques, University of Illinois, San Francisco.

[15] **Hurwitz, J. and Kirsch, D.**, (2018). Machine Learning For Dummies, John Wiley Sons, Inc., ibm limited edition edition.

[16] **Mitcell M., T.**, (1997), Machine Learning, McGraw-Hill, New York.

[17] **Turing, A.M.** (1950). Computing Machinery and Intelligence, *Mind*, *LIX*(236), 433–460.

[18] **Ertel, W.**, (2017), Introduction, Springer, Cham.

[19] **Mourya, S. and Gupta, S.**, (2012), Alpha Science International.

[20] https://codeburst.io/supervised-machine-learning-for-dummies-part-1-overview-15c18f2269ba, (Date accessed: April 2019).

[21] **Witten, I. and Frank, E.**, (2005), Data Mining: Practical Machine Learning Tools And Techniques, University of Waikato, San Francisco.

[22] **Sahu, H.B.**, **Shrma, S. and Gondhalakar, S.** (2011). A Brief Overview on Data Mining Survey.

[23] **World Health Organization**, (2018), WHO position paper on mammography screening, https://www.who.int/cancer/PRGlobocanFinal.pdf?ua=1, (Date accessed: April 2019).

[24] http://gco.iarc.fr/tomorrow/home, (Date accessed: April 2019).

[25] Basic Information About Breast Cancer, https://www.cdc.gov/cancer/breast/basic_info, (Date accessed: March 2019).

[26] **Memorial Sloan Kettering Cancer Center**, (2019), Stages of Breast Cancer, https://www.mskcc.org/cancer-care/types/breast/diagnosis/stages-breast, (Date accessed: March 2019).

[27] **World Health Organization**, (2014), WHO Position Paper on Mammography Screening. Geneva: World Health Organization, https://www.ncbi.nlm.nih.gov/books/NBK269535/, (Date accessed: April 2019).

[28] https://www.cancer.gov/types/breast/mammograms-fact-sheet#q5, (Date accessed: April 2019).

[29] **Alam, R.**, **Rafiq, M.**, **Suleman, T.**, **Arslan Tariq, M. and Sajid Farooq, M.** (2018). Data Mining Algorithms for Classification of Diagnostic Cancer Using Genetic Optimization Algorithms.

[30] **Saygili, A.** (2018). Classification and Diagnostic Prediction of Breast Cancers via Different Classifiers, *2*, 48–56.

[31] **Polat, K. and Güneş, S.** (2007). Breast Cancer Diagnosis Using Least Square Support Vector Machine, *Digit. Signal Process.*, *17*(4), 694–701.

[32] **Übeyli, E.D.** (2007). Implementing Automated Diagnostic Systems for Breast Cancer Detection, *Expert Syst. Appl.*, *33*(4), 1054–1062.

[33] **Othman, F. and Yau, T.**, (2007). Comparison of Different Classification Techniques Using WEKA for Breast Cancer, volume 15, pp.520–523.

[34] **Salama, G.**, **Abdelhalim, M.B. and Zeid, M.** (2012). Breast Cancer Diagnosis on Three Different Datasets using Multi-classifiers, *International Journal of Computer and Information Technology*, *1*, 36–43.

[35] **Hazra, A.**, **Kumar, S. and Gupta, A.** (2016). Study and Analysis of Breast Cancer Cell Detection using Naive Bayes, SVM and Ensemble Algorithms, *International Journal of Computer Applications*, *145*, 39–45.

[36] **Paluszek, M. and Thomas, S.**, (2017), MATLAB Machine Learning, Apress, Berkeley, CA.

[37] **Aggarwal C., C.**, (2015), Data Mining, Springer, United States.

[38] **Aggarwal C., C. and et al**, (2014), Data Classification: Algorithms and Applications, CRC Press, United States.

[39] **Tomar, D. and Agarwal, S.** (2014). A survey on Data Mining Approaches for Health care, *International Journal of Bio-Science and Bio-Technology*, *5*.

[40] **Sharma, A. and Suryawanshi, A.** (2016). A Novel Method for Detecting Spam Email using KNN Classification with Spearman Correlation as Distance Measure, *International Journal of Computer Applications*, *136*, 28–35.

[41] **Lantz, B.**, (2015), Machine learning with R, Packt Publishing, Birmingham.

[42] **Bramer, M.**, (2016), Principles of Data Mining, Springer, London.

[43] **Zhang, Z.** (2016). Introduction to machine learning: K-nearest neighbors, *Annals of Translational Medicine*, *4*, 218–218.

[44] **Vapnik, V.**, (1963). Pattern Recognition Using Generalized Portrait Method, pp.774–780.

[45] **Akinsola, J.E.T.** (2017). Supervised Machine Learning Algorithms: Classification and Comparison, *International Journal of Computer Trends and Technology (IJCTT)*, *48*, 128 – 138.

[46] **K., D. and Srivastava, Lekha, B.** (2010). Data Classification using Support Vector Machine, *Journal of Theoretical and Applied Information Technology*.

[47] **Murty, M.N. and Raghava, R.**, (2016), Kernel-Based SVM, Springer International Publishing.

[48] **Huang, M.W.**, **Chen, C.W.**, **Lin, W.C.**, **Ke, S.W. and Tsai, C.F.** (2017). SVM and SVM Ensembles in Breast Cancer Prediction, *PLoS One*, *12(1):e0161501*.

[49] **Kaur, G. and Oberai, E.N.** (2014). A Review Article On Naive Bayes Classifier With Various Smoothing Techniques.

[50] **Langarizadeh, M. and Moghbeli, F.** (2016). Applying Naive Bayesian Networks to Disease Prediction: a Systematic Review, *Acta Informatica Medica*, *24*, 364.

[51] **Sammut, C. and Webb, G.I.** (2017). *Encyclopedia of Machine Learning and Data Mining*, Springer Publishing Company, Incorporated, 2nd edition.

[52] **Chih-Min, M.**, **Wei-Shui, Y. and Bor-Wen, C.** (2014). How the Parameters of K-nearest Neighbor Algorithm Impact on the Best Classification Accuracy: In Case of Parkinson Dataset, *Journal of Applied Sciences*, 171–176.

[53] **Saranya, C. and Manikandan, G.** (2013). A study on normalization techniques for privacy preserving data mining, *5*, 2701–2704.

[54] **Baldi, P.**, **Brunak, S.**, **Chauvin, Y.**, **Andersen, C. and Nielsen, H.** (2000). Assessing the accuracy of prediction algorithms for classification: An overview, *Bioinformatics (Oxford, England)*, *16*, 412–24.

[55] **Galdi, P. and Tagliaferri, R.**, (2018). Data Mining: Accuracy and Error Measures for Classification and Prediction.

[56] **Majnik, M. and Bosnic, Z.** (2013). ROC analysis of classifiers in machine learning: A survey, *Intelligent Data Analysis*, *17*, 531–558.

[57] **Pham, H. and Triantaphyllou, E.**, (2008). The Impact of Overfitting and Overgeneralization on the Classification Accuracy in Data Mining, pp.391–431.

[58] **Arlot, S. and Celisse, A.** (2010). A survey of cross-validation procedures for model selection, *Statist. Surv.*, *4*, 40–79.

[59] Machine Learning Challenges: Choosing the Best Model and Avoiding Overfitting, https://www.mathworks.com/, (Date accessed: March 2019).

[60] **James, G.**, **Witten, D.**, **Hastie, T. and Tibshirani, R.**, (2013), An Introduction to Statistical Learning, Springer, New York, NY.

APPENDICES

APPENDIX A.1 : Python Code

APPENDIX A.1

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

veriler=pd.read_csv('data.csv')

veriler.drop(columns=['id','Unnamed:32'],inplace=True)
veriler.head()
print(veriler)

x=veriler.iloc[:,1:31].values
y=veriler.iloc[:,0:1].values

from sklearn.preprocessing import LabelEncoder
le=LabelEncoder()
y[:,0]=le.fit_transform(y[:,0])

from sklearn.preprocessing import OneHotEncoder
one=OneHotEncoder(categorical_features='all')
y=one.fit_transform(y).toarray()

from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y[:,1:],
test_size=0.33,random_state=0)


from sklearn.preprocessing import StandardScaler
sc= StandardScaler()

X_train =sc.fit_transform(x_train)
X_test=sc.fit_transform(x_test)

Y_train =sc.fit_transform(y_train)
Y_test=sc.fit_transform(y_test)

from sklearn.neighbors import KNeighborsClassifier
knn=KNeighborsClassifier(n_neighbors=9, metric='minkowski')
knn.fit(X_train,y_train.ravel())
y_pred=knn.predict(X_test)

from sklearn.metrics import confusion_matrix
```

```
cm=confusion_matrix(y_test,y_pred)
print(cm)

print(list(range(569)))

sonuc=pd.DataFrame(data=y[:,1:],index=range(569),columns=['diagnosis'])
print(sonuc)


sonuc2=pd.DataFrame(data=x ,index=range(569),
columns=['radius_mean','texture_mean',
'perimeter_mean','area_mean','smoothness_mean',
'compactness_mean','concavity_mean',
'concave␣points_mean','symmetry_mean',
'fractal_dimension_mean','radius_se',
'texture_se','perimeter_se','area_se',
'smoothness_se','compactness_se',
'concavity_se','concave␣points_se',
'symmetry_se','fractal_dimension_se',
'radius_worst','texture_worst',
'perimeter_worst',
'area_worst','smoothness_worst',
'compactness_worst',
'concavity_worst','concave␣points_worst',
'symmetry_worst',
'fractal_dimension_worst'])
print(sonuc2)
s=pd.concat([sonuc,sonuc2],axis=1)
print(s)
sonuc_new = (sonuc2 -np.min(sonuc2))/
(np.max(sonuc2)-np.min(sonuc2)).values


f,ax = plt.subplots(figsize=(20, 20))
sns.heatmap(sonuc_new.corr(), annot=True,
linewidths=.5, fmt= '.1f',ax=ax)

from sklearn.model_selection import cross_val_score
knn=KNeighborsClassifier()
k_range=list(range(1,50))
k_scores=[]
for k in k_range:
    knn=KNeighborsClassifier(n_neighbors=k)
    scores=cross_val_score(knn, X_train,y_train[:,0],
    cv=10,scoring='recall')
    k_scores.append(scores.mean())
```

```
plt.plot(k_range,k_scores,color="blue")
plt.xlabel('k␣values')
plt.ylabel('Recall')
plt.title('Efficient␣k␣values')
plt.grid(True)

from sklearn.model_selection import cross_val_score
knn=KNeighborsClassifier()
k_range=list(range(1,50))
k_scores=[]
for k in k_range:
    knn=KNeighborsClassifier(n_neighbors=k)
    scores=cross_val_score(knn, X_train,y_train[:,0],
    cv=10,scoring='accuracy')
    k_scores.append(scores.mean())
print(np.round(k_scores,3))

plt.plot(k_range,k_scores,color="red")
plt.xlabel('k-values')
plt.ylabel('Accuracy')
plt.title('Efficient-k-values')
plt.grid(True)
plt.show()

from sklearn.model_selection import GridSearchCV
k_range=list(range(1,50))
param_grid=dict(n_neighbors=k_range)
scores = ['accuracy', 'recall']
for sc in scores:
    grid_knn=GridSearchCV(knn,param_grid,cv=10,scoring=sc )
    print("#Tuning-hyper-parameters-for%s" % sc)
    grid_knn.fit(X_train,y_train[:,0])
    print(grid_knn.best_params_)
    print(np.round(grid_knn.best_score_,3))

from sklearn.metrics import confusion_matrix
cm=confusion_matrix(y_test,y_pred)
print('KNN')
print(cm)

from sklearn import metrics
from sklearn.svm import SVC
svc=SVC(kernel='linear' ,gamma=.01, C=5,probability=True )
svc.fit(X_train,y_train.ravel())
y_pred=svc.predict(X_test)

cm=confusion_matrix(y_test,y_pred)
```

```
print('SVC–lin')
print(cm)

from sklearn.metrics import accuracy_score
print(accuracy_score(y_test, y_pred))

from sklearn.svm import SVC
svc=SVC(kernel='poly' ,gamma=.01, C=5 ,probability=True)
svc.fit(X_train,y_train.ravel())
y_pred=svc.predict(X_test)

cm=confusion_matrix(y_test,y_pred)
print('SVC–poly')
print(cm)
print(metrics.accuracy_score(y_test, y_pred))


from sklearn.svm import SVC
svc=SVC(kernel='rbf' ,gamma=.01, C=5,probability=True )
svc.fit(X_train,y_train.ravel())
y_pred=svc.predict(X_test)

cm=confusion_matrix(y_test,y_pred)
print('SVC–RBF')
print(cm)

from sklearn.metrics import accuracy_score
print(accuracy_score(y_test, y_pred))

import numpy as np
import matplotlib.pyplot as plt
from sklearn.svm import SVC
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix
from sklearn import preprocessing
import pandas as pd
import random
import itertools
import seaborn as sns

bc = pd.read_csv('data.csv')
bc.head(1)

bcs = pd.DataFrame(preprocessing.scale(bc.ix[:,2:32]))
bcs.columns = list(bc.ix[:,2:32].columns)
bcs['diagnosis'] = bc['diagnosis']

X = bcs.ix[:,0:30]
```

```
y = bcs['diagnosis']
class_names = list(y.unique())

def plot_confusion_matrix(cm, classes,
                          normalize=False,
                          title='Confusion matrix',
                          cmap=plt.cm.Blues):
    plt.imshow(cm, interpolation='nearest', cmap=cmap)
    plt.title(title)
    plt.colorbar()
    tick_marks = np.arange(len(classes))
    plt.xticks(tick_marks, classes, rotation=45)
    plt.yticks(tick_marks, classes)

    if normalize:
        cm = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]
        print("Normalized confusion matrix")
    else:
        print('Confusion matrix for radial basis kernel')

    print(cm)

    thresh = cm.max() / 2.
    for i, j in itertools.product(range(cm.shape[0]),
                                  range(cm.shape[1])):
        plt.text(j, i, cm[i, j],
                 horizontalalignment="center",
                 color="white" if cm[i, j] > thresh else "black")

    plt.tight_layout()
    plt.ylabel('True label')
    plt.xlabel('Predicted label')
mtrx = confusion_matrix(y_test,y_pred)
np.set_printoptions(precision = 2)

plt.figure()
plot_confusion_matrix(mtrx,classes=class_names,
title='Confusion matrix for rbf kernel')

from sklearn.naive_bayes import GaussianNB
gnb=GaussianNB()
gnb.fit(X_train,y_train.ravel())
y_pred=gnb.predict(X_test)
cm=confusion_matrix(y_test,y_pred)
print('GNB')
print(cm)
```

```
from sklearn import metrics
print(metrics.accuracy_score(y_test, y_pred))
print(1-metrics.accuracy_score(y_test, y_pred))
import numpy as np
import matplotlib.pyplot as plt
from sklearn.svm import SVC
from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix
from sklearn import preprocessing
import pandas as pd
import random
import itertools
import seaborn as sns

bc = pd.read_csv('data.csv')
bc.head(1)

bcs = pd.DataFrame(preprocessing.scale(bc.ix[:,2:32]))
bcs.columns = list(bc.ix[:,2:32].columns)
bcs['diagnosis'] = bc['diagnosis']

X = bcs.ix[:,0:30]

y = bcs['diagnosis']
class_names = list(y.unique())

def plot_confusion_matrix(cm, classes,
                          normalize=False,
                          title='Confusion␣matrix',
                          cmap=plt.cm.Blues):
    plt.imshow(cm, interpolation='nearest', cmap=cmap)
    plt.title(title)
    plt.colorbar()
    tick_marks = np.arange(len(classes))
    plt.xticks(tick_marks, classes, rotation=45)
    plt.yticks(tick_marks, classes)

    if normalize:
        cm = cm.astype('float') /
            cm.sum(axis=1)[:, np.newaxis]
        print("Normalized␣confusion␣matrix")
    else:
        print('Confusion␣matrix␣for␣Naive␣Bayes')

    print(cm)

    thresh = cm.max() / 2.
```

```
    for i, j in itertools.product(range(cm.shape[0]),
                                  range(cm.shape[1])):
        plt.text(j, i, cm[i, j],
                 horizontalalignment="center",
                 color="white" if cm[i, j] > thresh else "black")

    plt.tight_layout()
    plt.ylabel('True label')
    plt.xlabel('Predicted label')
mtrx = confusion_matrix(y_test,y_pred)
np.set_printoptions(precision = 2)

plt.figure()
plot_confusion_matrix(mtrx,classes=class_names,
           title='Confusion matrix for NB ')

from sklearn.svm import SVC
svc=SVC(kernel='rbf' ,gamma=.01, C=5,probability=True )
svc.fit(X_train,y_train.ravel())
y_pred=svc.predict(X_test)

cm=confusion_matrix(y_test,y_pred)
print('SVC-RBF')
print(cm)

from sklearn.metrics import accuracy_score
print(accuracy_score(y_test, y_pred))

y_proba=knn.predict_proba(X_test)
y_score = knn.predict_proba(X_test)[:,1]

print(y_proba[:,0].ravel())
y_new=y_proba[:,0].ravel()

from sklearn import metrics
fpr,tpr,thold=metrics.roc_curve(y_test,y_new)


print(fpr)
print(tpr)

y_pred_knn_p =knn.predict_proba(X_test)[:,1]
y_pred_svc_p =svc.predict_proba(X_test)[:,1]
y_pred_lgr_p =gnb.predict_proba(X_test)[:,1]

models=[y_pred_knn_p,y_pred_svc_p,y_pred_lgr_p]
label=['KNN','SVM','NB']
```

```
plt.figure(figsize=(10, 8))
m=np.arange(3)
for m in m:
    fpr, tpr,thresholds= metrics.roc_curve(y_test,models[m])
    print('model:',label[m])
    print('thresholds:',np.round(thresholds,3))
    print('tpr:       ',np.round(tpr,3))
    print('fpr:       ',np.round(fpr,3))
    plt.plot(fpr,tpr,label=label[m])
plt.xlim([0.0,1.0])
plt.ylim([0.0,1.0])
plt.title('ROC curve for Cancer classifer')
plt.xlabel('False positive rate (1-specificity)')
plt.ylabel('True positive rate (sensitivity)')
plt.legend(loc=4,)
```

CURRICULUM VITAE

Name Surname: Burcu Meral

Place and Date of Birth: Istanbul, 25/10/1990

E-Mail: burcum1905@gmail.com

EDUCATION:

- **B.Sc.:** 2013, Bahcesehir University, Faculty of Arts and Sciences, Mathematics
- **B.Sc.:** 2014, Bahcesehir University, Faculty of Engineering and Natural Sciences, Software Engineering (Double Major)