UNIVERSITY OF ÇUKUROVA
INSTITUTE OF NATURAL AND APPLIED SCIENCES

MSc THESIS

Nusaibah Khalid ABDULMAJEED

CLASSIFICATION THE DATA SET IN MEDICAL CARE (BREAST CANCER) BY USING MACHINE LEARNING ALGORITHMS

DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

ADANA-2017

ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

CLASSIFICATION THE DATA SET IN MEDICAL CARE (BREAST CANCER) BY USING MACHINE LEARNING ALGORITHMS

Nusaibah Khalid ABDULMAJEED

MSc THESIS

DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

We certify that the thesis titled above was reviewed and approved for the award of degree of the Master of Science by the board of jury on 27/09/2017.

……………..…….....................	…………………………..………	………………..............................
Assoc. Prof. Dr. Turgay İBRİKÇİ	Assoc. Prof. Dr. Sami ARICA	Assist. Prof. Dr.İrem Ersöz KAYA.
SUPERVISOR	MEMBER	MEMBER

This MSc Thesis is written at the Department of Institute of Natural And Applied Sciences of Çukurova University.
Registration Number:

Prof. Dr. Mustafa GÖK
Director
Institute of Natural and Applied Sciences

This study was supported byResearch ProjectsUnit BAP Coordination Unit of Çukurova University (Project Number: FYL-2016-6805).

ABSTRACT

MSc THESIS

CLASSIFICATION THE DATA SET IN MEDICAL CARE (BREAST CANCER) BY USING MACHINE LEARNING ALGORITHMS

Nusaibah Khalid ABDULMAJEED

UNIVERSITY OF ÇUKUROVA
DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

Supervisor : Assoc. Prof. Dr. Turgay IBRIKCI
Year: 2017, Pages: 87
Jury : Assoc. Prof. Dr. Turgay IBRIKCI
: Assoc. Prof. Dr. Sami ARICA
: Asst. Prof. Dr. İrem Ersöz

Nowadays, datasets are growing daily to obtain knowledge from big dataset. Extraction operation of useful information from the dataset is called data mining that is one of the major techniques to get the diagnostic results especially in medical care fields as breast cancer. Breast cancer is one of the widespread cancers among women if matched with all other tumors all over the world. Classification is widely used in most important and necessary tasks in the real life applications in all fields. This technique has ability of detecting very similarities and differences that a human analyst may be not notice and therefore create and introduction more accurate/useful categories. This thesis presents comparison and analyses breast cancer dataset by using classification decision tree algorithms and kernel function by using Support Vector Machines via Sequential Minimal Optimization (SMO). Decision tree algorithms are applied to these algorithms which are J48, Functional Tree, Random Forest Tree, AD Alternating Decision Tree, Decision stump Tree, Simple Cart and Best First Tree. Moreover, Support Vector With (Normalized Poly Kernel, Poly Kernel, PUK Kernel and RBF Kernel). A computationally efficient classifies of these algorithms by employing Waikato Environment for Knowledge Analysis (WEKA) that is development program which includes a set of machine learning algorithms. These masses included 569 with 357 benign and 212 malignant cases with 32 attributes to test and proof the differences among the classification methods or algorithms.

Key Words: Machine Learning, Support Vector Machines, Decision Trees, Sequential Minimal Optimization, Data Mining.

ÖZ

YÜKSEK LİSANS TEZİ

TIBBİ BAKIMDA (GÖGÜS KANSERI) VERİ KÜMESİNİN MAKINE ÖGRENME ALGORİTMALARI KULLANILARAK SINIFLANDINLMASI

Nusaibah Khalid ABDULMAJEED

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
ELEKTRİK ELKTRONİK MÜHENDİSLİĞİ ANABİM DALI

Danışman : Doç. Dr. Turgay IBRIKCI
Yıl: 2017, Sayfa: 87
Jüri : Doç. Dr. Turgay IBRIKCI
: Doç. Dr. Sami ARICA
: Yrd. Doç. Dr. İrem Ersöz

Günümüzde, büyük veri kümesinden bilgi edinmek için data setleri her geçen gün büyümektedir. Veri tabanından yararlı bilgilerin çıkarılması işlemi veri madenciliği olarak tanımlanır. Bu da özellikle tıbbi bakımda tanısal sonuçları elde etmek için kullanılan en önemli tekniklerden biridir. Meme kanseri, diğer bütün tümör türleri ile karşılaştırıldığında kadınlarda yaygın görülen kanserlerden biridir. Sınıflandırma, gerçek hayat uygulamalarındaki her alanında yaygın bir şekilde kullanılmaktadır. Bu teknik bir insan tarafından fark edilemeyecek küçük benzerlik ve farklılıkları tespit etme yeteneğine sahiptir. Böylece daha faydalıve doğru sınıfların oluşmasını sağlayacaktır. Bu tez, Sıralı Minimal Optimization (SMO) destekli Destek Vektör Makinalarının farklı çekirdek fonksiyonları ile ve karar ağaçları algoritmaları ve meme kanseri verilerini analizve karşılaştırmaları ortaya koymaktadır.Karar ağacı algoritmaları olarak:J48, Functional Tree, Rasgele Orman, AD Değişen karar Ağacı, Karar kütüğü Ağaç, Basit Sepet veEn İyi İlk Ağaç algoritmaları uygulanmaktadır. Ayrıca Destek Vektörü Makinası olarak Normalleştirilmiş Poly çekirdeği,Poly Çekirdeği, PUK çekirdeği ve RBF Çekirdeği kullanılmıştır. Bu algoritmaların sonuçları, bir makina öğrenme ortamı olan WEKA üzerinde ortaya koyulmuştur. Bu sınıflandırma algoritmaları arasındaki farklılıkları test etmek için 32 özellikli 357 iyi huylu ve 212 kötü huylu (malign)sonuçlarına sahip toplamda 569 veriyi içermektedir.

Anahtar Kelimeler: Makine Öğrenme, Destek Vektör Makineleri, Karar Ağaçları, Sıralı Minimal Optimizasyon, Veri Madenciliği.

EXTENDED ABSTRACT

We are living in an informational economy. Millions of data being downloaded on a daily basis into the database. From here, comes the necessity to find a way to filter the only concerned data and represented in a certain pattern in order to ease the accessibility of it. Generally, data mining technique used to obtain updated data that extracted from the database for further strategic analysis.

Data mining in the medical field is a powerful technology with huge potential it offers benefits to caregiver, patients, medical care collective, searchers, and insurers. As well as useful to study the influence income, costs, and operating efficiency, but at the same time keeping high-quality care.

Classification is a data mining *“ machine learning ML, could be expounded as a collection of topics that assure, on making and testing algorithms that could help the process of classification, foretelling, get new modality by using numbers of computer models ”* technique applied to forecast set for data instances.

Automatic classification made based on similarities located in the data. Moreover, automatic classification technique is proving fruitful if the result that drawing via the automatic classifier is agreeable to the clinician.

To perform automatic classification depending on the testing data set by choosing one sub-sample that keeper as the validation data for testing example or model, it obtains out 10 similar sized groups. Every group divided into two levels, 90 called data used for training and 10 called data used for testing. That it gets out a classifier with the algorithm of 90

called data and used on the 10 testing data for group 1. Likewise, give precision in percentage correctly classified, time wastaken, accuracy and measurements of error rates and comparison among them for numbers of cases in the testing dataset. The point of early disclosure and screening of breast tumor is to limit this illness-specific mortality and treatment-dependent morbidity.

Several classification functions have been suggested by searchers in the area of classifying in a breast tumor dataset. The hypothesis is to use a general method for decision trees and the kernel function bong support vector machine via sequential minimal optimization SMO. In order to extract valuable information with high accuracy regarding the attributes in this dataset B.C data set. However, why sequential minimal optimization? Since finding the optimal parameters for a classifier can be a rather tedious process, WEKA toolbox written in the Java programming language that offers some automating ways to solve big quadratic programming QP optimization matter, one of these ways is sequential minimal optimization. That breaches a big QP problem for a series of shorter possible QP problems in a shorter time.

The primary goal of this thesis is to apply several algorithms that used in many research for the purpose of demonstrating the best algorithm with high-resolution and without losing the data classification (misclassify) or (poorly-classify) and taking into account the factors and parameters that contain work branch used. By the way of example in SMO in RBF-kernel selected the C factor SMO - C 8. zero, the precision is 98. 2425%, because C factor impacts the trade-off between ratio of non-separable examples and complex, so that must be chosen by user. C complexly parameter tells a

SVM-SMO optimization how much want to avoid misclassified for all training examples. The relationship between the values of C factor and margin hyper-plane, If the value of C factor was small, will cause the windows optimizer to look for a larger-margin separating hyper-planes, even if that excitable plane mis-classifies more factors. For very small of C, will get misclassified examples, often even if training data is linearly separable. Conversely, In case the value of C factor was large, the optimization will choose a smaller-margin excitable plane if that excitable plane does a much better work for getting all training points classified correctly. Generally an optimal C factor should be larger the size down of data, and the other way round, so if have tiny values for features, be sure to include very large values for the possible C values.

We could say that the study proves with its results that supervised learning algorithms are suitable algorithms into the realm of cancer diagnosis, for potential use in collaboration with the established medical dataset and to help avoid evasive diagnostic tests on patients.

ACKNOWLEDGMENTS

Firstly, I would like to express my sincere gratitude to my advisor Assos. Prof. Dr.Turgay IBRIKCI for the continuous support of my MSc. study and related research, for his patients, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my study.

Besides my advisor, I would like to thank the rest of my thesis committee Assoc. Prof. Dr. Sami ARICA and Assit. Prof. Dr. İrem Ersöz, for their insightful comments and encouragement, but also for the hard question which invented me to widen my research from various perspectives.

Last but not the least, I would like to thank my family, To the fountain of patience and optimism and hope, To each of the following in the presence of God and his messenger, my mother dear, To the big heart my dear father.

To the one who kidnapped beauty by leaving, to the pure soul in my homeland.

To my brother who for supporting me spiritually throughout writing this thesis, To the people who paved our way of science and knowledge.

CONTENTS **PAGE**

LIST OF TABLES **PAGE**

LIST OF FIGURES **PAGE**

LIST OF ABBREVIATIONS

ML : Machine learning
WEKA : Waikato Environment for Knowledge Analysis
WBC : Wisconsin Breast Cancer
BRCA : Breast Cancer susceptibility gene
BMI : Body Mass Index
MRI : Magnetic Resonance Imaging
SVM : Support Vector Machine
SMO : Sequential Minimal Optimization
QP : Quadratic Programming
CSV : Comma-Separated Values
x : Input Attribute Set
y : Out Class Lable
DT : Decision Trees
D_c : No. of Branches
b_x : No. of Bcriteria
$g_t\ (x)$: No. Base Hypothesis
BF : Best First
FP : Functional Tree
AD : Alternating Decision
RF : Random Forest
DS : Decision Stump
SC : Simple Cart
CART : Classification and Regression Tree algorithm
w : Weight
α_i : No.of Incorrect Prediction Made One Example
N : Dimensional Hyperplane

Acc	: Accuracy
TPR	: True Positive Rate
ROC	: Receiver Operating Characteristic
k	: Kappa Statistic
$Pr\ (\alpha)$	**:** Relative Observed Agreement Among Raters
$Pr(^{e})$	: Hypothetical Probability Of Chance Agreement
MAE	: Mean Absolute Error
e_i	**:** Aversge Af Absoulute Errors.
MSE	: Mean Squared Error
MSD	: Mean Squared Deviation
RAE	: Relative Absolute Error
RRSE	: Root Relative Squared Error

1. INTRODUCTION

Nowadays, data sets are growing daily so that is becoming difficult to obtain knowledge from big, massive, noisy or complex data sets without a new technique. In addition, growth in biomedical research so that, from this development which become relates into all facets of modern life. The possibility to use these data sets to extract helpful information for quality health or medical care is critical.

Retrieve of information may help support goodness decision-making and to bypass or avoid a human mistake all of that via computer-aided. Although human decision-making is often near from optimal or ideal, it is still poorly path when there are huge amounts of data sets to classify them. Also, activity and precision of decisions will be lower when humans putting into overwork and massive work. Consequently, this technique able to detect similarities and differences that a human analyst might be not noticedand therefore create and introduction more accurate and useful class.

Extraction operation of useful information from the dataset calling data mining. That is one of the major techniques to get the diagnostic results, especially in medical care fields as breast cancer. Data mining software is one of many analytical tools which handles with the intelligent algorithms in order to analysis the data.

In another side, data mining is a process of finding connections and samples among many fields in huge associated data sets *(Quinlan, J. R., 1986).* The data in medical care *(D. R. Canlas Jr., 2009)* is entity applied to pestilences, treat illness, get a better form of life additionally, avoid or bypass illness which preventable deaths.

Data mining in the medical field is a powerful technology with huge potential it offers benefits to caregiver, patients, medical care collective, searchers, and insurers. As well as useful to study the influence income, costs, and operating efficiency, but at the same time keeping high-quality care.

Classification is a data mining *“ machine learning ML, could be expounded as a collection of topics that assure, on making and testing algorithms that could help the process of classification, foretelling, get new modality by using numbers of computer models ” (Ian H. Witten Eibe, 2011)* technique applied to forecast set for data instances.

Automatic classification made based on similarities located in the data. Moreover, automatic classification technique is proving fruitful if the result that drawing via the automatic classifier is agreeable to the clinician.

Launches a label of breast cancer when cells of breast begin to grow out of control *(Rajesh K, Anand S., 2012).* Breast cancer is a non homogeneous disease and emerge danger dependent on several factors, like tumor size, histological grade, regional lymph a node, involvement, and expression of estrogen (ER) and progesterone hormone receptors (PgR), as well as human epidermal growth factor receptor (HER2) amplification *(Cobain EF, Hayes DF, 2015).*

To perform automatic classification depending on the testing data set by choosing one sub-sample that keeper as the validation data for testing example or model, it obtains out 10 similar sized groups. Every group divided into two levels, 90 called data used for training and 10 called data used for testing. That it gets out a classifier with the algorithm of 90 called data and used on the 10 testing data for group 1. Likewise, give precision in percentage correctly classified, time wastaken, accuracy and measurements of error rates and comparison among them for numbers of cases in the testing dataset. The point of early disclosure and screening of breast tumor is to limit this illness-specific mortality and treatment-dependent morbidity.

This study aimed to display data mining technique could be applied to the medical data sets like breast cancer dataset for classification and compare a good algorithm among others which used in this study.

In current work using WEKA workbench for classifying the breast tumor (Wisconsin breast cancer) dataset into the two types depending on the characteristic of the masses also from other features.

The prime contributions of this thesis are:

(1) To discover the helpful classified precision for a forecast of WBC illness.
(2) Comparison of various data mining algorithms on the illness data set.
(3) Recognize a good performance of an algorithm for the prediction of disease.

1.1. Overview of Breast Cancer

In worldwide, breast cancer is the most illness that threatening the life at womankind and the main reason for cancer dying between female. Useful for early diagnosis saves a large number of women from the risk of death, gives a chance of surviving, less treatment or medication-induced impairment and better quality of life. So that, the aim of this study was to measure the biological characteristics in cells. It may be infected or prone to infected breast cancer and classified via diagnostic computer-based on applications that which using open widely in nowadays. Needful to understand what is this disease, and what types, automatic diagnosis and what are the factors that increase from occurring this disease.

So in this section, It explains in summarizing the most important points relating to this illness.

1.1.1. Cancer and Breast Cancer

Cancer is a broad term for a class of diseases characterized by abnormal cells that grow and invade healthy cells in the body. On commonality kinds of cancer, several of the body's cells start to divide without stopping and catch into near tissues.

But, when cancer evolves, this organized operation collapse. Also, the cells become more and more abnormal, the old cells when they should die, and new cells form when they are not needed. These extra cells can divide without stopping and may form growths survive called tumors. In figure 1.1. shows by steps how the production of cancer cells.

Breast Cancer is one of the widely prevalent cancer. About one in 9 women evolve the breast cancer at several stages in their lives. In the UK 48.000 cases appear every year. Most evolve in women through the age of 50 but younger women are at times affected *(NICE Clinical Guideline, July 2014).*

A benign tumor grows slowly, usually harmless and covered by a membrane which keeps the tumor solitary from the normal surrounding cells. A malignant tumor is almost never covered by a membrane. It usually grows faster than a benign tumor and overrun the surrounding tissues. Malignant cells also, a capacity to spread through the blood and lymphatic systems to distant organs *(Bristol-Myers Squibb, 2014).*

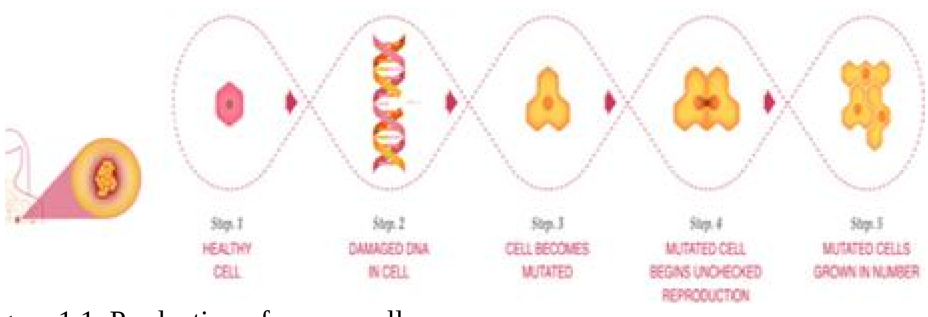


Figure 1.1. Production of cancer cells
Source: http://www.nationalbreastcancer.org/causes-of-breast-cancer.

1.1.2. Types of Breast Cancer

By looking under the microscope the cancer cells divided into several forms:-

- **Benign breast cancer (non-invasive or cancer in situ)** This kind does not propagate or spread to nearby tissue areas and rather than a threat to life. These cells still entirely in-situ, they haven't able to spread outside of these ducts, because these cells not yet developed *(Breast Cancer Organization Types of Breast, 2015).* The Figure 1.2. shows non-invasive cells.

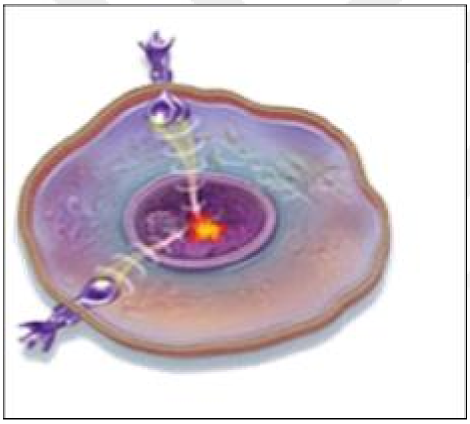


Figure 1.2. Non-Invasive cells

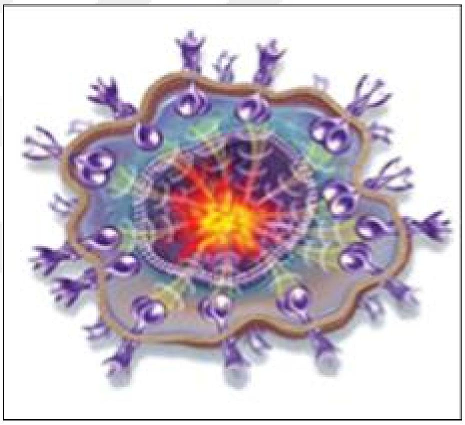


Figure.1.3. Invasive cells

Source: http://www.breastcancer.org/symptoms/diagnosis/invasive

- **Malignant breast cancer (Invasive)** This kind of cancer could be prevalence of breast to another parts of the cells body. So, it is a threat to life. often they could be taken away, but on occasion come back to grow, this occurs if a few cells separated from the main tumors carried in the bloodstream or lymph canals to other parts of the body *(Breast Cancer Organization types Of Breast, 2015).* The Figure 1.3. shows invasive cells.

1.1.3. Risk Factors For Breast Cancer

Although breast tumor that evolution for not clear source, there are many particular risk factors whose increase the probability of B.C cancer and will be from a simple to a more complex form. These risk factors are :-

- **Gender** The major risk factor of B.C is being a women. The illness is almost 100 times more popular amid women than men *(Breast Cancer Organization Risk Factors, 2015).*
- **Age** The peril of evolving breast tumor increase about double for each 10 years of age. The most states happen in women above the age of 50. For example, a percent of starting at the following ages, is as follows *(Breast Cancer Organization Risk Factors, 2015).*

0.44 percent (or 1 in 227) for age 30.
1.47 percent (or 1 in 68) for age 40.
2.38 percent (or 1 in 42) for age 50.
3.56 percent (or 1 in 28) for age 60.
3.82 percent (or 1 in 26) for age 70.

The Figure 1.4. shows increase the risk of breast cancer with age

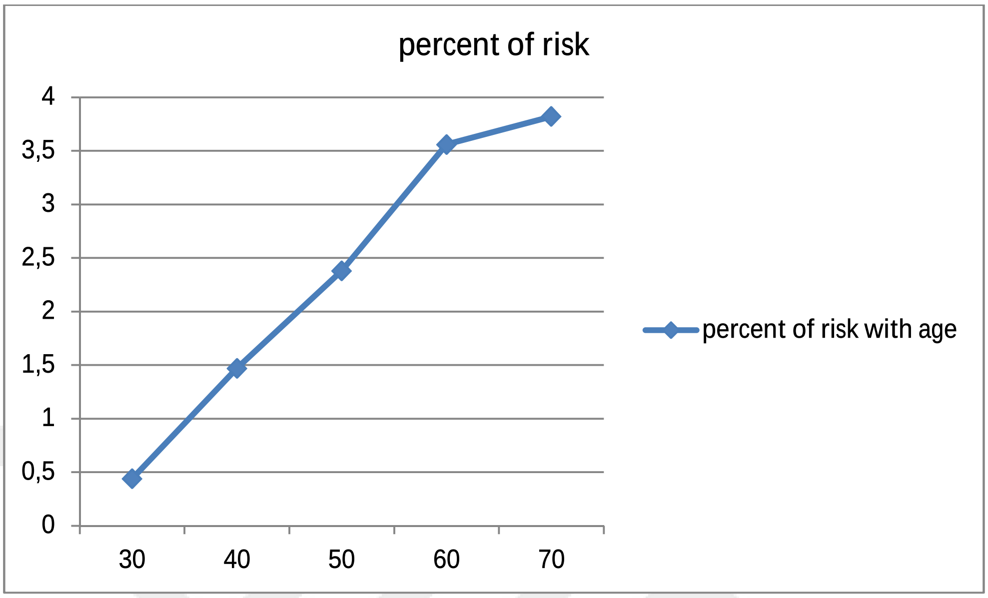


Figure 1.4. Increase the risk of B.C with age

- **Environment** The ratio of B.C difference between countries *(Breast Cancer Organization Risk Factors, 2015).*
- **Family history** A risk B.C growing if the woman's mom, sister or father were diagnosed before with breast cancer and the risk rises if the relative was a diagnosis of the disease before age 50 **(***Breast Cancer Organization Risk Factors, 2015).*
- **Alcoholic, Smokin, Overweight** Alcohol consumption increases the risk of B.C. Smoking-related with the highest risk of B.C in the young and before menopause for women. A fatty women has a higher risk of being diagnosed with B.C from the women with healthy weight *(Breast Cancer Organization Risk Factors, 2015).*
- **Genetic risk factors** % 5 of B.C cases have a strong genetic risk descended from their families *(Breast Cancer Organization Risk Factors, 2015).* There are two dominant genes, BRCA1 & BRCA2, breast cancer susceptibility gene1 and gene2 BRCA1 and BRCA2 are both key elements

in DNA upkeep mechanisms that avoided the DNA damage passing on to further generations of cells. Because, of their protective roles against the passing on of incorrect or mutated DNA, BRCA1 and BRCA2 referred to or naming as tumor suppressor genes *(Sawyer SL, Tian L, 2015)*, that account for most cases of familial B.C. Women with hurtful BRCA leap have 65% to 85% risk of evolving B.C. The Figure 1.5. shows B.C susceptibility genes.

Other risk factors include:

-The highest BMI ***(Body Mass Index)*** after menopause: the age of menopause: Overweight gain after menopause raises the risk of B.C.

-Diet and vitamin intake, lack of exercise.

-Chemicals in the environment.

-Radiance treatment to the breast (before age 30), Late rope at an older age. Hormonal use – postmenopausal.

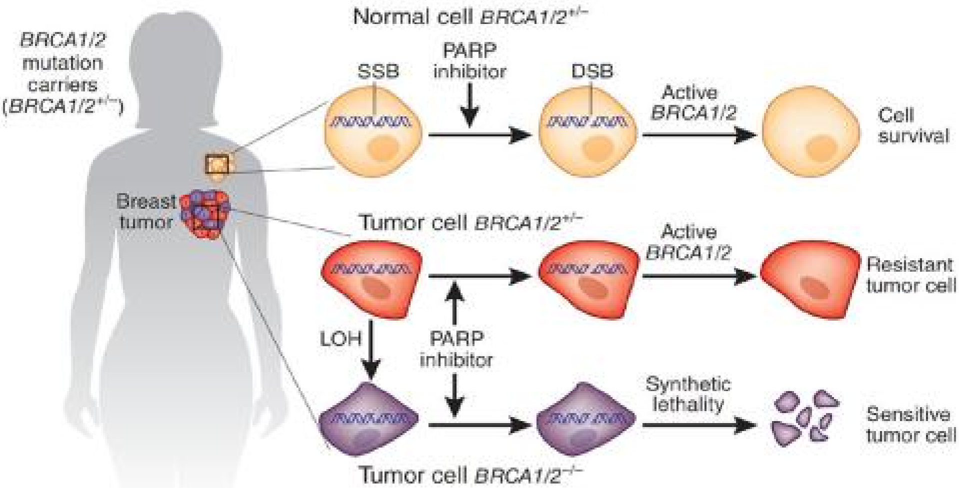


Figure 1.5. Breast cancer susceptibility genes in DNA.
Source: https://cureallcancer.wordpress.com

1.1.4. Diagnosis of Breast Cancer

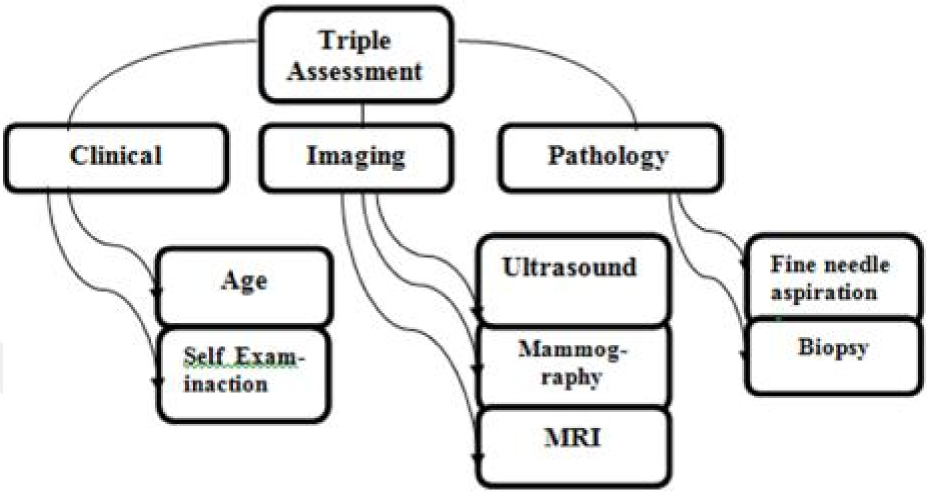


Figure 1.6. Diagnostic tests for breast cancer

The screening and diagnostic tools for detecting and diagnosing breast tumor has become more and more exact since the 1970s, when modern mammography started to become open widely. The Figure 1.6. shows the types of test B.C.

The most common paths for diagnosis B.C are:

- **<u>Clinical</u>**
  1. Ages (The section 1.1.3 explains relationship ages with a diagnosis of disease).
  2. Self-examination the breast tested for mass or suspect parts (variation in textile & sizes) *(Goharian& Grossman, 2004).*

- **Imaging**

1. *Mammography:* The mammography is helpful for discovering early changes in the breast when it might be applied hard to sense the mass. It is an X-Ray of the breast. It applied to diagnose women who have any sign in breast. A mammography shows abnormalities in the breast, which content lesions.
2. *Ultra-sound* option uses audio or sound waves to make images of the breast. This technique is usually used as a follow-up test after an abnormal result on mammography, breast magnetic resonance imaging or clinical breast exam *(Goharian& Grossman, 2004).*
3. *Breast magnetic resonance imaging, MRI* is an expanded medical test used by doctors to diagnose many of these diseases. This technique applies a strong magnetic domain, radio frequency vibration linked to a computer for product details images of a breast. This technique does not use ionizing radiation (X-ray). MRI provides valuable information on many situations of the breast that cannot be acquired by other imaging techniques, like ultrasound and mammography.

- **Pathology**

1. A breast biopsy is the only sure first way to diagnose cancer cells in the breast. Although providing technologies such as MRI and ultrasound can find a doubtful area, they cannot tell whether the cancer is or not. A small part of the suspect tissue removed and grown to determine its type if it is malignant or benign *(Goharian & Grossman, 2004).*
2. Fine needle aspiration in this test, a very thin hollow needle used and injected with a breast to pull a small sample of the tissue with its fluids and test the sample that contains cancer or not *(Goharian& Grossman, 2004).*

1.1.5. Treatment of Breast Cancer

The main treatments used for B.C include:

- **Surgery** It is a way that most people think about the types of treatment. There are many types of surgery for breast cancer, depending on the type, size and location of the tumor. Sometimes only they have the ability to remove the cancerous mass. Sometimes the breasts removed and accompanied by radiotherapy for several weeks *(National breast cancer, 2016).*
- **Radiot** It means high-energy X-rays that force cancer cells to kill them using a special machine. In some cases, it is given with chemical, radiation and called chemo-radiation. Radiation therapy has side effects. Radiation can destroy damaged cells and normal cells. Cancerous cells cannot repair themselves. However, normal cells often recover *(National breast cancer, 2016).*
- **Chemotherapy** Often accompanied by other treatments such as hormonal treatments and targeted therapies. It is given to patients, according to the basis of several factors, including tumor type, size, type of receptors, a number of lymph nodes involved and the degree of the effect of the ureters on the adjoining tissues of the body. Sometimes it's used as a secondary treatment to reduce the size of the tumor before surgery to make sure removal of the tumor easily and more safely *(National breast cancer, 2016).*
- **Hormone therapy** Estrogens & progesterone are chemicals generated by glands in the body. But, sometimes these same hormones can cause cancer to grow. Both types of these drugs help to destroy cancer cells by cutting off their extensions of hormones. The removal of hormonal extensions

(dietary supplements) from the cell leads to cell hunger, death, and self - destruction *(National breast cancer, 2016).*

2. ARTICLE REVIEWS

Clinical diagnosis of the nature of an illness like B.C supports in foretelling the malignant tumor cases. The various widespread paths used for B.C diagnosis, mammography, magnetic resonance imaging MRI, biopsy. The outcome takes out, those paths used to identify modalities or patterns that contribute to helping doctors to classify malignancies and benign tumors. There are various data extraction data mining techniques, M.L algorithms and precise mathematical methods are all applied with this aim. The view of literature presents a summarized description of the researchers on data mining techniques with illness area. The brief description of the algorithms used along with outline results presented in the review.

Sudhir D. Sawarkar et. al. *(Sudhir D. Sawarkar, 2006)* uses ANN and SVM algorithms on the WB.C data set. After treatment data set the outcome of ANN and SVM foretelling models more accurate found from human prognosis, so that, use of this forecasting could be used to avoid biopsy for diagnoses WBC, because of 96.7% high accuracy in these prediction models.

Bin Othman et. al. *(Bin Othman, 2007)* uses different algorithms for simulation outcomes of every algorithm as BNN, RBF, D.T and solitary conjugation ruling learning algorithms. Simulated the errors by using through their work it can be concluded that BNN performs best for B.C data. The time taken to building a model was 0.19 Sec. precision 89.71 % also, least mistake at 0.2140 as corresponding to other algorithms applied.

Zhiyan Yang, p. et. al. *(Zhiyan Yang, P. et, 2010)* applied this method for machinery fault diagnosis for ship intelligent fault diagnosis of chine box. They used one algorithm from decision tree was random forest also, ML as neural networks and genetic algorithms. From comparing, they got the random forest done good results if used to medicine, biology machine learning. Therefore, it is able to apply these algorithms to different data, but in the same field.

Orlando Anunciacao et. al. *(Orlando Anunciacao, 2010)* by using WEKA machine learning tool on dataset found by (Nova de Lisboa University - Medical Faculty - Genetics Department) for 164 keepers & 94 cases discover use D.T algorithms for detection of high-risk B.C sets. To validate the statistical found, applied test if permutation. In this study found a high-risk B.C set contains 13 cases and just 1 control, with Fisher perfect test (validate) the value 9.7* 10-6, a p- of 0.017. The outcomes appeared that it is possible to find statistically significant links with B.C by deriving resolve tree and choose top leaf.

Ahmed Abdelaal et. al. *(Ahmed Abdelaal, 2010)* uses a classifier SVM with tree forest and tree boost in analyzing DDSM dataset to extract features for a mammalian cluster along with age which distinguishes true and false cases. SVM technicality, show favorable outputs to increase diagnosis precision for classifying the cases in a big region under a ROC comparable to values of and tree forests.

Sugandhi C. et. al. *(Sugandhi C, 2011)* shows the classification algorithm for cataract patient illness by using WEKA tool. WEKA has been used to classify the results and for comparison objective. The result for SMO gives 77.34% accuracy of classification which means better execution as compared with other algorithms such as the random tree, J48 and Naive Bayes.

Shrivastava et. al. *(Shrivastava, 2013)* uses classification techniques for classifying the benign or malignant instances of UCI breast cancer dataset neural network and decision tree approaches. She has used WEKA for testing for simplifying the foretelling task. The best result of decision tree algorithms is J48 with the accuracy 98.14%.

Shajahaan et. al. *(Shajahaan, S. Syed, 2013)* explored apply decision trees for breast cancer prediction of classic supervised learning algorithms such as ID3, C4.5, CART and Naïve Bayes. The test performs through WEKA. The authors have concluded that the random tree serves as the best classification algorithm for breast cancer with higher accuracy in the prediction.

Durairaj M. et. al. *(Durairaj M Ranjani V., 2013)* used different data set and applied in data mining for serving healthcare field. The algorithms like NB, KNN, C4.5, and J48 used for classifying and foretell many illnesses such as heart disease, dengue fever, B.C, AIDS, diabetes, dialysis, hepatitis and in vitro fertilization IVF. Compare study analysis uncover top precision 97.7 % for cancer foretelling and about 70 % for in vitro fertilization treatment by using those technicalities.

Dhamodharan S. et. al. *(Dhamodharan S., 2014)* by the usefulness of data mining technicalities diagnoses and resolve liver illnesses applied WEKA workbench via Naïva Bayes and FT tree algorithms. When compared with the outputs of FT tree algorithms better than from Naïve Bayes.

Chaurasia et. al. *(Chaurasia, 2014)* conducted the experiment in WEKA with BF Tree (Best First Tree), IBK (K-nearest neighbor classifier) on breast cancer data. and have taken three estimates standard such as time, correctly classified instances and accuracy for assessing the superiority of each algorithm. The authors had stated that the performance of BF algorithms had been better than the other algorithm in terms of accuracy and low error rate.

Nguyen T. et al. *(Nguyen T., 2014)* layout a system for medical diagnosis through joining fuzzy standard additive model (SAM) with wavelet features. Comparing results from applying fuzzy SAM with wavelets for data reduction in breast cancer and heart disease with probability neural network (PNN), support vector machine (SVM), Fuzzy Adaptive resonance theory (ARTMAP) and Adaptive Neuro-fuzzy inference system (ANFIS). The system proved higher accuracy in medical diagnosis compared to other techniques.

Williams et al. *(Williams, Kehinde, 2015)* focused on two data mining techniques, namely Naïve Bayes and J48 decision trees to foretell breast cancer risks in Nigerian patients, which have collected the dataset from cancer registry of LASUTH, Ikeja in Lagos, Nigeria which contains 69 instances with 17 attributes along with the class label. The analysis made to determine the most efficient and effective model. The dataset carries on 11 non-modifiable factors and five modifiable factors. The test is conducted through WEKA and the authors have claimed J48 decision tree is better for the foretelling B.C risks with a value of accuracy 94.22% and error rates.

Venkatesan E. *(Venkatesan E., 2015)* analyzed the B.C dataset using four classification algorithms naming J48, CART, AD tree and best first. The authors have conducted the experiment through WEKA tool. The seeker has claimed that the decision trees have a standard construct and easy to understand from the rules can be extracted J48 classifier has the highest accuracy of 99%.

Layla Safwat Jamil *(Layla Safwat, 2016)* used WEKA workbench is with five algorithms are analysis all results. These algorithms are Naïve Bayes classifiers, SVM, D.T, K-Star Instance-based classifier, and ANN which used on B.C data set. The best classifier has the highest accuracy of 75.52% was J48.

3. MATERIAL AND METHODS

3.1. Material

Several classification functions have been suggested by searchers in the area of classifying in a breast tumor dataset. In this work can define the problem is possibilities, explore disease by using machine learning algorithms on a dataset in medical care, this data set containing clinical data for thousands of samples taken from both diseased and healthy tissues. The hypothesis is to use a general method for decision trees and the kernel function bong support vector machine via sequential minimal optimization SMO. In order to extract valuable information with high accuracy regarding the attributes in this dataset B.C data set. However, why sequential minimal optimization? Since finding the optimal parameters for a classifier can be a rather tedious process, WEKA toolbox written in the Java programming language that offers some automating ways to solve big quadratic programming QP optimization matter, one of these ways is sequential minimal optimization. That breaches a big QP problem for a series of shorter possible QP problems in a shorter time. This classification with good activity, efficacy by using supervised learning algorithms. The Figure 3.1. shows the methodology of the study.

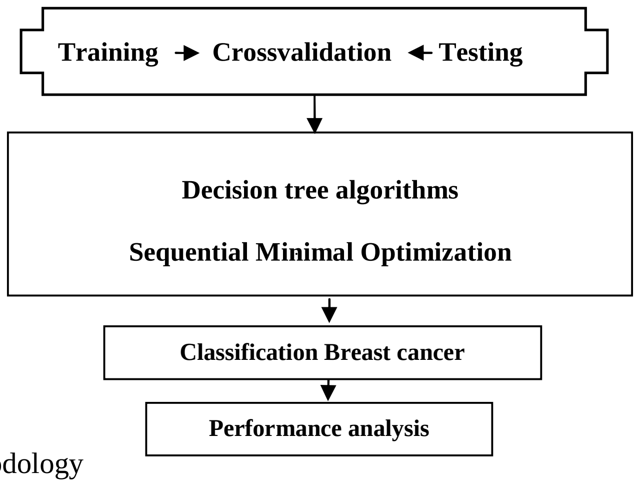


Figure 3.1. Methodology

3.1.1. Dataset and Features

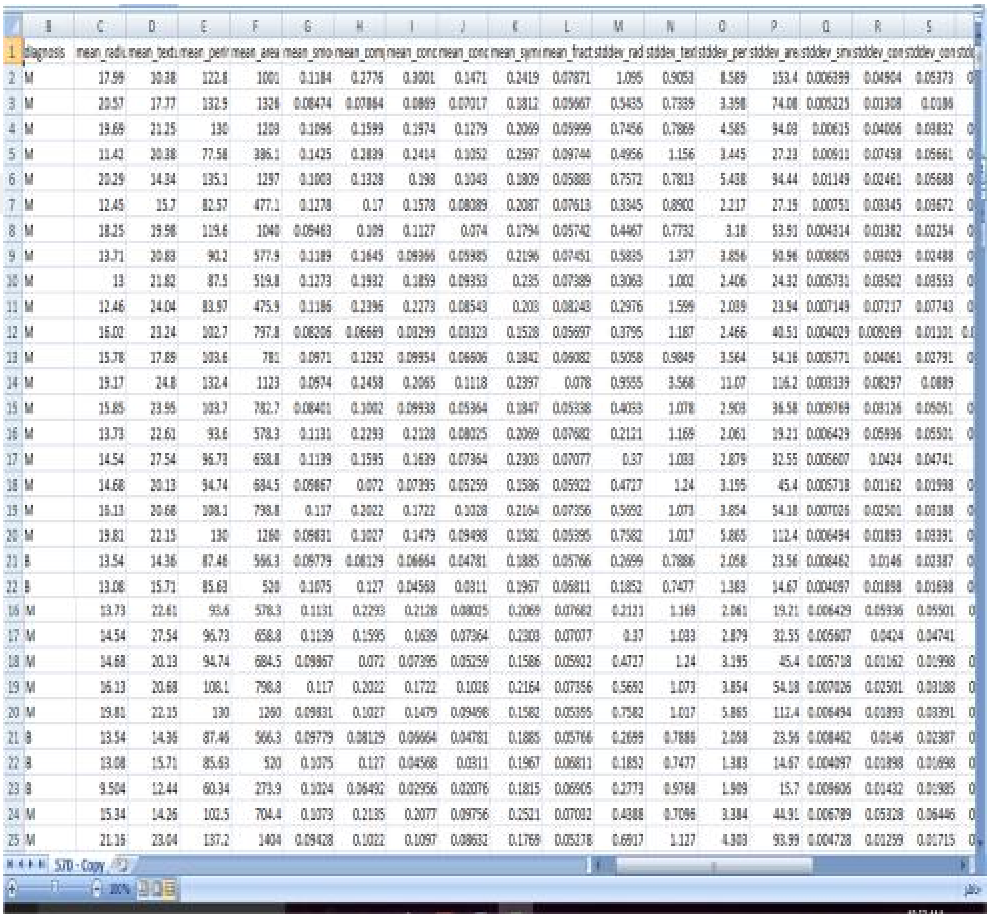
The dataset is collecting of data or statistical data organized so that it can easily be accessed, be administered and be updated where every character represents changing and any instance has its own description. Another view, a data set classified according to types of content, bibliographic, full-text, numerals, and image. Breast cancer data set one of these data, which used in this study. B.C dataset had taken from Wisconsin B.C data via UCI ML storehouse by Dr. William H. Walberg *(UCI Mach. Lone Repository, 1995), (UCI Mach. Learn Repository, 1995).* An attribute is data field, acting as a characteristic or feature of a data object *(Jiawei Han, 2011)*. A feature computed for each cell, namely ID number is the 1st column and the 2nd column is classes naming which are type's diagnosis illnesses B, Benign or M, Malignant in Table 3.2 shows the class label in the dataset. The rest of 30 features are the properties of cells with

- Mean
- Standard errors
- Worst values

of radius, circumference, texture, smoothness, concavity, space, compactness, symmetry, concave points the fractional dimension to all cell nucleus for 569 patients. Figure 3.2 shows a description of B.C dataset. This data set is used for prediction and classification of algorithms in order to compare their accuracy and other measurements using one of WEKA's interfaces (Explorer). In Table 3.1. shows a description of the dataset.

Table 3.1. Description of dataset

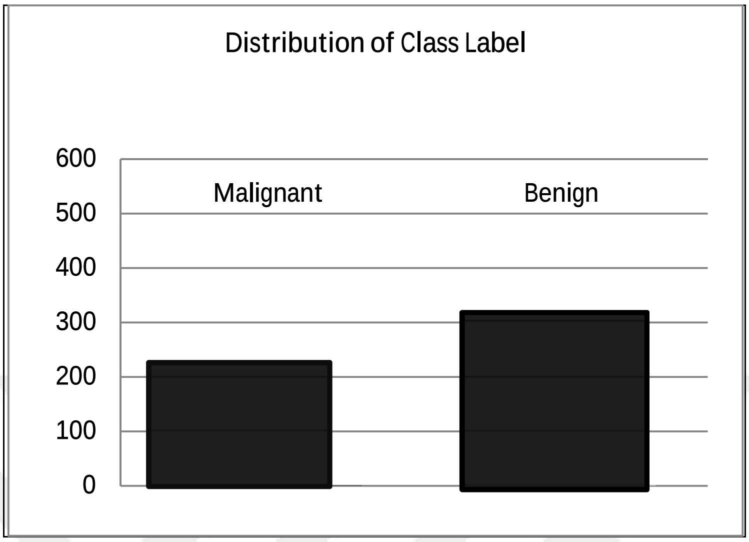
Dataset Characteristics	Multivariate	No. of Attributes:	31
Attribute Characteristics	Real	No. of Instances:	569
Associated Tasks	Classification	No. of Classes:	2



	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	diagnosis	mean_radiu	mean_textu	mean_perir	mean_area	mean_smo	mean_com	mean_conc	mean_conc	mean_sym	mean_fract	stddev_rad	stddev_text	stddev_per	stddev_are	stddev_sm	stddev_con	stddev_con
2	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589	153.4	0.006399	0.04904	0.05373
3	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08	0.005225	0.01308	0.0186
4	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	4.585	94.03	0.00615	0.04006	0.03832
5	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	3.445	27.23	0.00911	0.07458	0.05661
6	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813	5.438	94.44	0.01149	0.02461	0.05688
7	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345	0.8902	2.217	27.19	0.00751	0.03345	0.03672
8	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7732	3.18	53.91	0.004314	0.01382	0.02254
9	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377	3.856	50.96	0.008805	0.03029	0.02488
10	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1.002	2.406	24.32	0.005731	0.03502	0.03553
11	M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.2976	1.599	2.039	23.94	0.007149	0.07217	0.07743
12	M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.3795	1.187	2.466	40.51	0.004029	0.009269	0.01101
13	M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082	0.5058	0.9849	3.564	54.16	0.005771	0.04061	0.02791
14	M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	0.9555	3.568	11.07	116.2	0.003139	0.08297	0.0889
15	M	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	0.4033	1.078	2.903	36.58	0.009769	0.03126	0.05051
16	M	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	0.2121	1.169	2.061	19.21	0.006429	0.05936	0.05501
17	M	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077	0.37	1.033	2.879	32.55	0.005607	0.0424	0.04741
18	M	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922	0.4727	1.24	3.195	45.4	0.005718	0.01162	0.01998
19	M	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.07356	0.5692	1.073	3.854	54.18	0.007026	0.02501	0.03188
20	M	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395	0.7582	1.017	5.865	112.4	0.006494	0.01893	0.03391
21	B	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	0.05766	0.2699	0.7886	2.058	23.56	0.008462	0.0146	0.02387
22	B	13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	0.1967	0.06811	0.1852	0.7477	1.383	14.67	0.004097	0.01898	0.01698
16	M	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	0.2121	1.169	2.061	19.21	0.006429	0.05936	0.05501
17	M	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077	0.37	1.033	2.879	32.55	0.005607	0.0424	0.04741
18	M	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922	0.4727	1.24	3.195	45.4	0.005718	0.01162	0.01998
19	M	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.07356	0.5692	1.073	3.854	54.18	0.007026	0.02501	0.03188
20	M	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395	0.7582	1.017	5.865	112.4	0.006494	0.01893	0.03391
21	B	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	0.05766	0.2699	0.7886	2.058	23.56	0.008462	0.0146	0.02387
22	B	13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	0.1967	0.06811	0.1852	0.7477	1.383	14.67	0.004097	0.01898	0.01698
23	B	9.504	12.44	60.34	273.9	0.1024	0.06492	0.02956	0.02076	0.1815	0.06905	0.2773	0.9768	1.909	15.7	0.009606	0.01432	0.01985
24	M	15.34	14.26	102.5	704.4	0.1073	0.2135	0.2077	0.09756	0.2521	0.07032	0.4388	0.7096	3.384	44.91	0.006789	0.05328	0.06446
25	M	21.16	23.04	137.2	1404	0.09428	0.1022	0.1097	0.08632	0.1769	0.05278	0.6917	1.127	4.303	93.99	0.004728	0.01259	0.01715

Figure 3.2. Screenshot view of B.C dataset

Table 3.2. Distribution of classes label in dataset



3.2. Data Mining Technique (Waikato Environment for Knowlege Analysis)

We are living in an informational economy. Millions of data being downloaded on a daily basis into the database. From here, comes the necessity to find a way to filter the only concerned data and represented in a certain pattern in order to ease the accessibility of it. Generally, data mining technique used to obtain updated data that extracted from the database for further strategic analysis.

That is why data mining becomes very important in this study in order to have constant access to an updated concerned information regarding a certain disease or ascertain patient. In the study, It thought that WEKA tool is a very useful for this study. It is one of these software applications which used for knowledge analysis and it offered free online because it is an open source software. Without the necessity to program that will give us time to handle the SW application as well as experimenting the algorithms and their results *(Jiawei Han, 2011)*.

WEKA tool is an Information system which uses a software capable of self-modification self-learning through algorithms, This is which is also known as the artificial intelligence AI. WEKA tool is one of these software applications which offers me a suitable environment to perform data mining. Because it represses the data and classifies it. Also, reassemble the data to make a certain scheme or a pattern of data according to association rules. We can also apply many algorithms to the medical care dataset in order to find which algorithm deliver us the most correct reading. The algorithm is capable of delivering different readings according to the updated data in the dataset.

Through the WEKA tool also an automated learning schemas or automated learning patterns can be created. These can be easily accessed to be assessed and analyzed. The WEKA tool also provides us with a command line and a user interface which makes it easier to work. We do not have to program anything, all features can be used from the command line. Also, a user guide is available, It trains you to use this free software efficiently.

3.2.1. User Interfaces

The master interface of WEKA is an explorer, however, basically, the same functions could be obtained via a component based on knowing inflow intermediation and from the command line. There is also a tester, that allows for a systematic comparison of the prediction performance of WEKA ML algorithms on a set of data sets.

Explorer face is part of WEKA Graphical User Interface GUI, Explorer face is different panels supplying incoming to prime content of the work-branch:

- A preprocessor includes facilities for importing data from a database, separated by comma value file type CSV etc., and to reprocess a data applying the so-called filtering algorithm. Those filters could be applied convert type of data (such as, convert digital (numeric) attributes to nominal attributes.

The classify panel enables the using CART functions to the resulting dataset, to rating the precision of the outcome foretelling pattern, and to imagine wrong forecasts and characteristic operating curve (ROC) etc. It is imaginable like D.T. Figure 3.3. present mapping of an input attribute *x* in a category *y* classification of the classification.

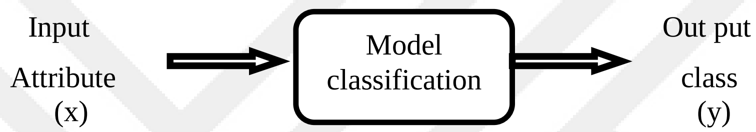


Figure 3.3. Mapping input attribute and output class depending on the classification type

- The access panel supplies getting for assembly ruling know-how or learns that tries to recognize all the significant mutual relations between the features in the dataset.
- The cluster panel allows incoming to clustering technicality in WEKA an easy or simple k-means algorithm. Furthermore a performance of the expectation maximization algorithm to learning an admixture for plain allocations.
- The select features panel provides algorithms to the most foretelling features in the dataset.
- The imagine or visualize panel offers a slanderer plot matrix, where individual scatter plots ability to selected and expanded and further analyzed applying different selection operators.

3.3. Classification Algorithms

A classification is one main technique of data mining. One of a useful feature of these techniques is to find hidden information. Dataset classification is in two-steps practicability, where a classification model is built up called a learning step and classification step where the model used to foretell classes naming for giving dataset *(Jiawei Han, Micheline Kamber, 2011)*.

3.3.1. Decision trees

Decision trees (D.T) are a multi-effect analysis that is one of the classification algorithms. Commonly forms of multiple changeable analysis let us foretell, explain and classify the output. Multiple changeable analysis is especially important in the present-days to solving problems, Because of about all critical results that limited successes are based on or depending on various operators.

Depending on research in cognitive psychology it has proved that the ability to understand and manipulate multiple parts of knowledge is limited by the material and cognitive proselimitations of the short part of that brain *(Tversky A., Miller 1956 & 1974).* This gives many advantages of manipulating techniques and displaying dimensions that are able to maintain and reverse high-dimensional relationships in an easily understandable form. The relationships can be consumed more easily and applied by humans.

D.T is simple and soft, but in the same time very powerful form of diversified variable analysis. D.T is generated by these algorithms are selected in different paths of dividing a data set offshoot as segments. Those segments from an overturned D.T that creates together with the root node at the upper of the tree.

D.T has another naming that is classification trees, that widely used in the classification implementation sand several expansions and differences have been suggested *(R. Rastogi and K. Shim, 1998).* To make a decision by this method and build up classification models, D.T break-down a dataset into smaller and smaller sets whilst keeping at the same time linked decision tree is gradually progressing.

A definitive result is a tree with decision nodes also leaf-nodes. A perfect outcome to make a decision nodes in the tree that matches to top or best foreteller, named root node *(Quinlan, JR, 1986).*

Many advantages for decision trees make us choose it because it is attractive in data mining climate for numerously At first, their outcome is explainable. Second, they don't need any parameter as input. Third, the structuration operations are comparatively speedy. We think that these advantages are also applicable to B.C. D.T is excessively used by many researchers in healthcare field *(Gehrke, G. Ganti, R., 1999)*. Figure 3.4. shows that decision trees are more easy to understand.

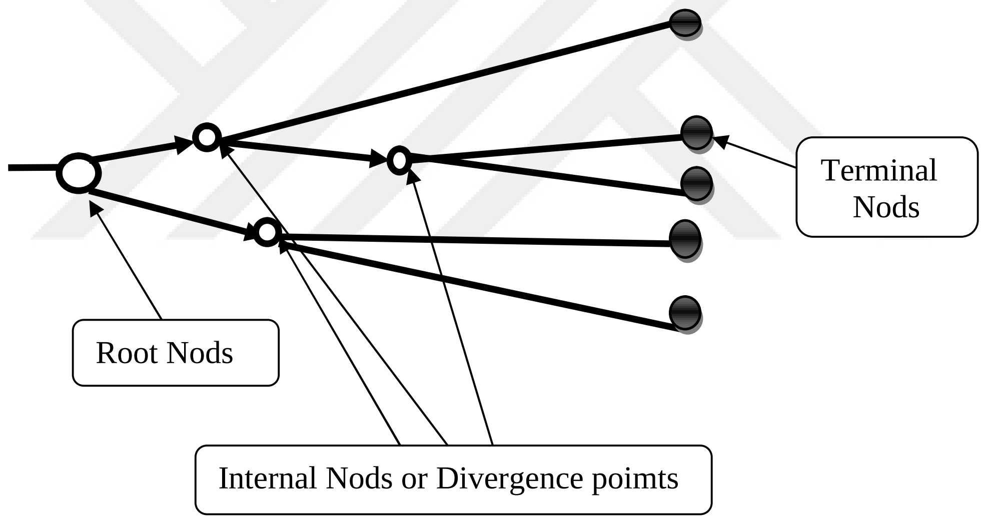


Figure 3.4. Block diagram for decision tree

Where:- A. Nodes of the tree

B. Leaves nodes of the tree / (terminal nodes)

C. Branches of the tree / (decision point)

A Basic Decision Tree Algorithm

$$G(x)\sum_{c=1}^{c}[\![b(x) = c]\!]\, G_c(x) \tag{3.1.}$$

Function decision tree (data $D = \{ (x_n, y_n) \}_{n=1}^{N}$)

If termination criteria met

Return base hypothesis $g_t(x)$

Else

1. Learn branching criteria b_x
2. Split D to C parts D_c = {(x_n, y_n : b(x_n) = c}
3. Build sub – tree $G_c \longleftarrow$ decision tree (D_c)

Where:

D_c No. of branches

b_x No. of branches criteria

$g_t(x)$ No. of base hypothesis

3.3.1.1. J48 Tree Algorithm

J48 is one of widespread decision tree algorithms, because it's actives with specific attributes, lost or missed values of the dataset. By this algorithm can be increased a precision by pruning ***(UCI, 1995) (Kaur G., 2014).*** C4.5 algorithm carries out J48 to an obstetrics strips C4.5 decision tree. To classify a dataset by a perfect as possible, J48 algorithm builds up D.T for data from a collection of training or workout data.

J48 classifier then visits all decision-node frequently after that chooses the most active splits until each leaf is pure and no more any splits are obtainable *(Torgo, L. 1997).*

Algorithm: Breast diagnosis

Input: B.C dataset preprocessed to meet basic requirements of the data mining technique.

Output: J48 D.T predictive models with benign or malignant diagnostic results.

Procedure:

1. Get a dataset from the internet B.C data set.
2. Pre - operation data for using J48 D.T data mining.
3. Upload the dataset in WEKA as a workbench for analysis it.
4. J48 algorithm performed D.T with a leaf node as the type class if there malignant or benign.
5. The new patients diagnosed by reference to the new attribute in the decision tree values, after that follow a channel or path until the leaves node arrives which if either benign or malignant tumor.

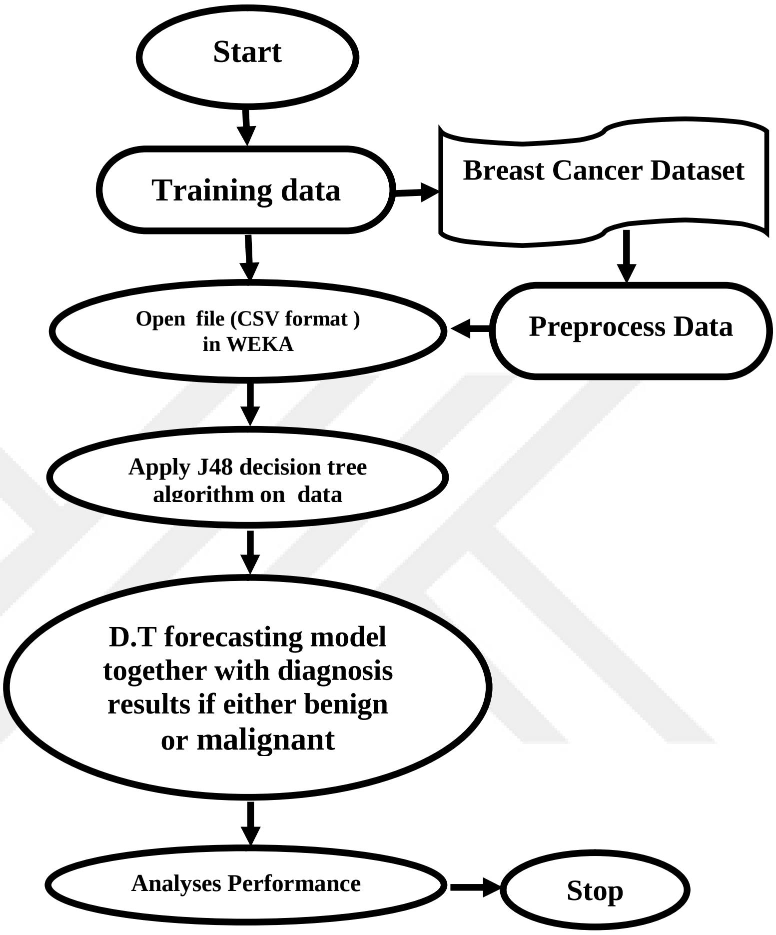


Figure 3.5. Diagram for breast tumor detection

3.3.1.2. Best First Tree Algorithm

Best First (BF) tree algorithms a form of decision tree learning, and may do almost of the characteristics of standard decision learning *(Shi H, 2007).* The naming best-node is saying for nodes that divide leads to the utmost limit of pollution among all nodes that helpful for splitting. It can deal with categorical and numerical variables.

The aim advantage can be taken to various rates *V*, the universe of set*S* is defined for this classification asformula

$$Entropy(s) = \sum_{i=1}^{v} - p_i \, log_2 p_i \qquad (3.2.)$$

Where: p_i is the ratio of *S* relationship to class ***i***, is the rate of *S* relation for type class ***i***. Informations gaining is the predictable *V* in entropy rise by slitting the traineeship dataset according to this feature. Most accurately, gain of the information defined as a:

$$Gain\ (S, A) = Entropy\ (S) - \sum_{v\ \in value\ (A)} \frac{|Sv|}{|S|} Entropy\ (s_v) \qquad (3.3.)$$

Where: The value of *A* is collection of each available values for feature *A* and *S* is subset for which feature *A* the value *V* (Sv = $S \in SAS$ = v).

3.3.1.3 Functional Tree Algorithm

Functional Tree (FT) is one of the form of a multivariate tree ***(J. Gama, 2004).*** A classifier for buildup functional tree ***(Freund Y., 1999)*** that classification trees may have logistic regression tasks on the internal nodes / leaf nodes.
The functional tree can handle with binary and non-binary trees which called multi-ways or multi-classes targeted variables, missing or null values, numeric and nominal attributes.

3.3.1.4. Alternating Decision Tree Algorithm

AD algorithm was created by Yoav Freund and Llew Mason ***(FreundY., 1999)*** for classification that different from the other algorithms because AD includes two types nodes are decision / forecast nodes. Each decision-node contains a division test, whereas all forecast nodes contain a real number for valued *(Frank, E., 2016).*

3.3.1.5. Random Forest Tree Algorithm

RF also sometimes called random decision forest *(Breiman L., 2001)* is active in many various open source programs (toolboxes). RFs are fairly strong with relatively straightforward applications for inexperienced users.

A Supervised RF algorithm can be applied to solve unsupervised learning problem *(Breiman L., 2001).* These advantages make the RF as a suitable algorithm for enabling incremental learning. RF classifier is data mining method developed by Beriman and Cutler that is fast in learning and, runs efficiently in big datasets (*Quinlan J. R.,1986).*

RF increasingly for ML has achieved broad popularity in the last few years, particularly in the field of illnesses because RF offers two aspects that are very useful in data mining: high forecasting accuracy and new information on the variable importance for classification methods *(Quinlan J. R.,1986).*

3.3.1.6. Decision Stump Tree Algorithm

A decision stump *(Sharma P., 2014)* is fundamentally a decision tree with only one single split. A tree in this algorithm divided at the root-level, which depend on a specified feature value pair. Sometimes, this algorithm is also called 1-rule *(Iba.W., 1992).*

3.3.1.7. Simple Cart Tree Algorithm

Simple Cart (SC) it was a method progressed by Leo Breiman in the early 1980s *(Breiman L., 2001).* SC way is CART analysis, abridged for (Classification and Regression Tree algorithm). This way using for prediction and exploration. This algorithm is best for the training data. Classification and regression learning technique, which afford the outputs as either classification or regression tree, depends on categorical or numeric data set. SC could be deal with missing values, empty nominal attributes, binary attributes, unary attributes, nominal attributes and numeric attributes.

It uses cross-validation or a large separate test pattern of data to choose a better tree from the sequence of trees according to the clipping operation.In the enforcement of CART, the data is divided into the two sub-collection these are most different in relation to the outcome.

This step is continued on all subcollection till some minimum subcollection size is arrived *(S. S. Aman, 2011).*

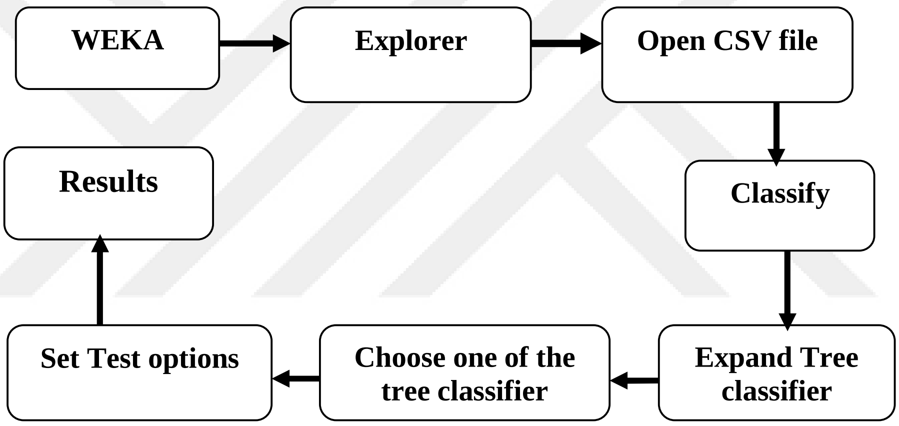


Figure 3.6. Decision tree classification process using WEKA

3.3.2. Support Vector Machine (SVM)

SVMs are teaching methods under supervision. Itis developed by vapnik *(V.N. Vapnik, 1995)* that prominent technicality for resolve classification problem.

The aim of SVM resolved classification and retrogression machinethat reduce atentative risk named training group error and over confirmation which matching to testing group errors *(V.N. Vapnik, 1995 & N.Cristianini, 2000).*

As it's known a SVM most famous machine learning techniqe for two classes (or examples) classification matters, according to labeled train dataset, $(x_{1,}y_1)$, $(x_{1,}y_1)$ where $x_i \in RN$ is the feature vector and $y_i \in \{-1,1\}$ is a class label.

Figure 3.7. shows there are many possible linear factors that can divide the data, But there is only one as much as possible margin, which increases the space between it and the nearest point of data for all categories. This linear classification is called the optimal separation hyper plane.

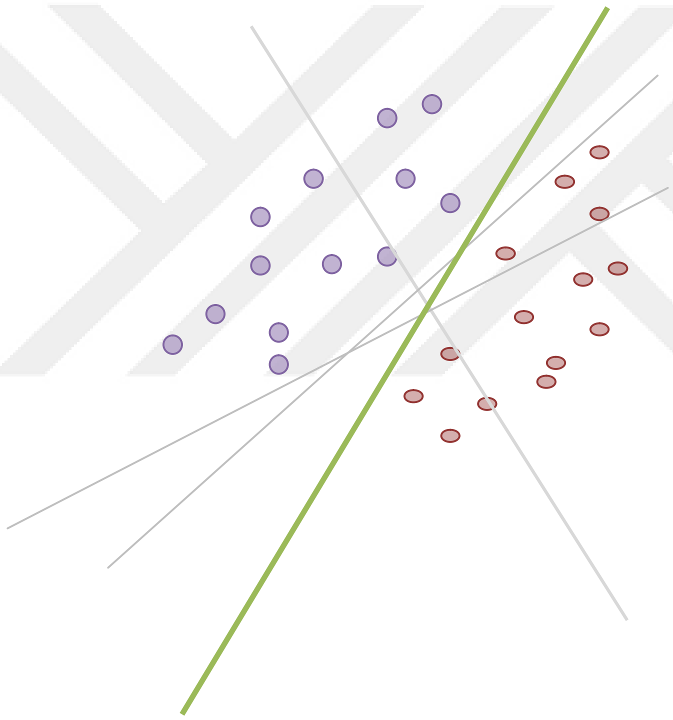


Figure 3.7. Optimal separating hyperplane

The algorithm is looking for determing the decision superficies that affords wider margin or split among data classes while at the same time remission no of errors. However, this decision super face is not produced in input space, but rather in a very H-Dimensional attribute space.

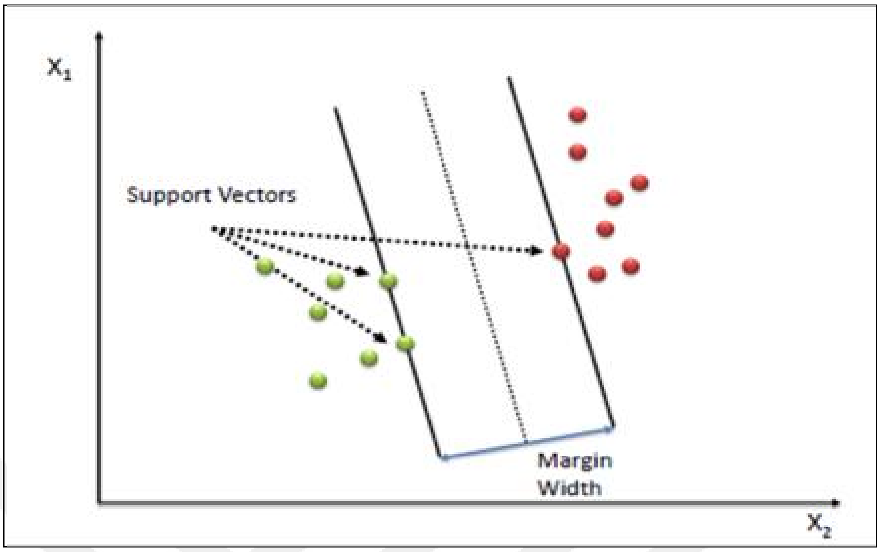


Figure 3.8. Vector (cases) that known the hyperplane are support vector
Sourcr: http://www.saedsayad.com/support_vector_machine.htm

The equation of the margin is given by

$$y_{x=}\, y_{x(Wx+b)} \tag{3.4.}$$

***Where*:** *w* is weight and *b* constant

The equation of the weight is given by

$$w = \sum_{i=1}^{N} \alpha_i y_i\, x_i \tag{3.5.}$$

***Where*:** α_i is the no. of incorrect prediction made one example.

The easiest path to divide the two sets of data with a straight line without a curve, an N-dimensional hyper plane. However, there are cases where a nonlinear area's ability split or divide the sets more efficiency.

SVM deals with a kernel function or kernel trick (nonlinear) to map the data into a various space where a hyperplane (linear) cannot be applied to do the class. In other words, transforming data into a 3-Dimensional space through the mapping makes the problem much easier.

A non-linear task is learning via the linear learning machine in a high-dimensional feature space while the system's ability is controlled by a parameter that does not depend on space dimensions. It is called kernel trick which mean kernel function converts the data into an H-Dimension feature uperficies to make it possible to perform linear separation or push the data to will be close from linear classifier.

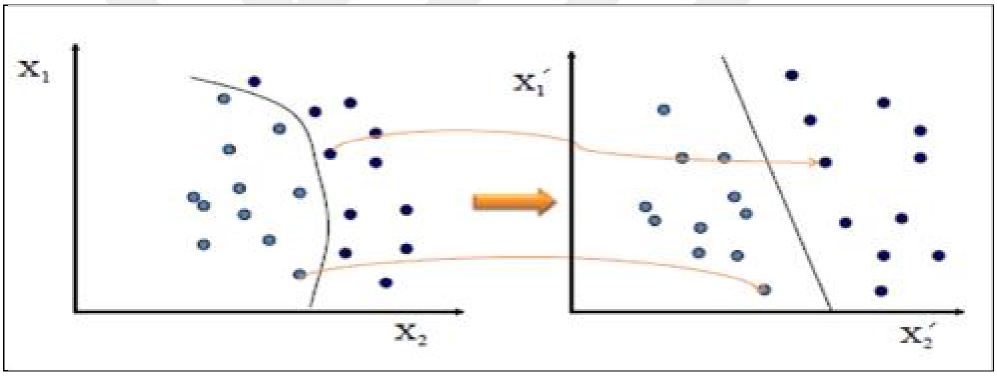


Figure 3.9. Kernel trick
Source: http://www.saedsayad.com/support_vector_machine.htm

The equation of decision rule of kernel function with two classes is given by Figure 3.5.

$$f(x) = sig\left(\sum_i \alpha_i y_i \, \varphi_{(x)} \cdot \varphi_{(xi)+} \, b\right) \tag{3.5.}$$

In this thesis used four types of kernel function with SVM are:

1. Poly nominal kernel with degree ***d***.

*(scale * crossprod (x,y) + offset)* ^{d}

$$K(x, y) = (x^T y) + c)^d \tag{3.6.}$$

2. Normalized-Polynomial kernel with d degree

$$K(x, y) = \frac{\left((x.y) + 1)^d\right)}{\sqrt{((x.x)+1)^d \quad ((y.y) + 1)^d}} \tag{3.7.}$$

3. Radial basis function kernel with dimension or spreadɣ

$$K(x, y) = exp\ [\, - ɣ\ ||\, x - y\,||^{\,2}\ ] \tag{3.8.}$$

4. Pearson puk kernel with Pearson expanse parametersσ, ω.

$$K(x, y) = \frac{\mathbf{1}}{\left[\frac{\mathbf{2}.\sqrt{\|\boldsymbol{x}-\boldsymbol{y}\| \quad ^{\mathbf{2}}\sqrt{\mathbf{2}^{\left(\frac{\mathbf{1}}{\boldsymbol{\omega}}\right)-\mathbf{1}}}} \ ) \quad ^{\mathbf{2}}}{\boldsymbol{\sigma}}\right]} d \tag{3.9.}$$

Where: x is the input attibute and ***y*** is the output class, ***d,*** ɣ are kernel parameters.

To take out classes of SVM with preferable execution, suitable traineeship is wished for. SVM training requires solving big quadratic programming QP problem.

However, the sequence minimum optimal-SMO with the fixed working group size is among the popular decomposition method of training that can be used even for a very large one *(Shigo Abe, 2005),* or can say SMO resolves the QP matter analytically without additional storage matrix.

Summary:

- Active in H-Dimension. Still effective in states wherever no. of dimensions bigger than no. of patterns. Uses a sub set of training points in the decision function-(called support vectors), so it is also memory efficient.
- Versatile or many-sided: various K- function ability specified for decision function. Common kernels are supplied, but also probable to specify custom kernels. It has a regularisation parameter, that makes the employer think about avoiding over-fitting. Whether number of attributes is bigger than number of patterns, the way is probable to give badexecutions. SVMs do not directly supplied probability assessment, So that must be calculated applying indirect techniques.

3.4. Performance Assessments

In machine learning applications, recall and other specificity measures are widely used bastides the common accuracy measure. Formal definitions for these performance criterias that described as shown below.

-True Positive (TP) of positive samples of correctly predicted to be positive.

-True Negative (TN) of negative samples of correctly predicted to be negative.

-False Positive (FP) of negative samples erroneously predicted to be positive.

-False Negative (FN) of positive samples erroneously predicted to be negative.

- (N) number of sample instances.

Confusion Matrix: It is so common in M.L field, contains information on actual and projected classifications performed by the rating system.
The performance of these systems is usually assessed using the data in the matrix.

Table 3.3 Confusion Matrix.

		Predicted	
		P	***N***
Actual	***P***	TP	FN
	N	FP	TN

Table 3.3 Representconfusion matrix with TP (true positive), FN (false negative), FP (false positive) and TN (true negative).

Precision: It is the ration of the true positive-number of total number of forecasted positive samples. Precision is TP / predicted positive.

$$\text{Precision} = \frac{TP}{(TP+FP)} \tag{3.10.}$$

Recall or sensitivity: It is a ratio of positive samples were correctly specified in the total number of positive samples. It is known as sensitivity analysis of the true positive ratio (TPR).

$$\text{Recall or sensitivity} = \frac{TP}{(TP+FN)} \tag{3.11.}$$

Accuracy: It is a measure of the ratio for correctly predicted.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FN + FP)} \quad (3.12.)$$

F-Measure: is the average of accuracy to recall. It is an important measure that combines precision and recall.

$$\text{F} - \text{Measure} = \frac{(2*Recall*Precision\ )}{(precision+Recall)} \quad (3.13.)$$

Receiver Operating Characteristic-ROC: is the active part of estimating the goodness or performance of diagnosis tests. ROC is the relationship between sensitivity and specificity for a specific diagnosis. For ROC curves the *y-axis* represents a sensitivity and an *x-axis* is representational specificity

TPR is plotted over the *y-axis* and FP is plotted along *x-axis.* Perform of all classifiers acted like a point on ROC curve.

Kappa Statistic: is a measure of correcting the chance of agreement between classifications and real classes. It is calculated by taking the expected agreement by accident away from the note agreement and dividing it as far as possible agreement. A value greater than 0 means that the workbooks do a better job than chance.

$$\text{K} = \frac{\Pr(\alpha) - Pr}{1 - \Pr(e\ )} \quad (3.14.)$$

Where: *Pr (α)* it is a remarkable agreement comparison between the residents, and *Pr (^{e})* is the hypothetical possibility of a chance. An agreement is hypothetical probability of chance agreement.

Mean Absolute Error-MAE it applied for measurement how much closed predictions are for final results. MAE usually similar in magnitude but slightly smaller than the root mean square error.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i| \tag{3.15.}$$

Where e_i is average of absoulute errors.

The Mean Squared Error-MSE/ Mean Squared Deviation-MSD of a measure (from an estimate for an unobserved quantity) measures the average squares of errors or deviations.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} \tag{3.16.}$$

Absolute Error and Relative Error are two kinds of experimental error. RAE absolute in certainty is the unbelief in measure, who expressed using the pertinent unity. As well, the absolute error can be applied to express the inaccuracy in measuring the relative error is absolute error divided by the exact value size. Values are ratios, not have units. Therefore, when you choose the form, it can have a larger link and a smaller error estimate.

$$RAE = \frac{\sum_{i=1}^{n}\left|p_{i-a_i}\right|}{\sum_{i=1}^{n}\left|a_{-a_i}\right|} \tag{3.17.}$$

Where: p_i is the predicted target value, and a_i the actual target value.

$$RRSE = \frac{\sum_{i=1}^{n}\left(p_{i-a_i}\right)^2}{\sum_{i=1}^{n}\left(p_{-a_i}\right)^2} \tag{3.18.}$$

4. EXPERIMENTAL RESULTS

The problem in this thesis is the define possibilities explore disease by using machine learning algorithms on data-sets in medical care, this data set containing clinical data for thousands of samples taken from both diseased and healthy tissues. The first step for applying the machine learning techniques is to prepare the breast cancer data-set, which consists 31 features (attributes) are properties of cells with mean, standard errors, worst values of the radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and the fractional dimension to each cell nucleus for 569 patients. The second step tells us which algorithm is best.

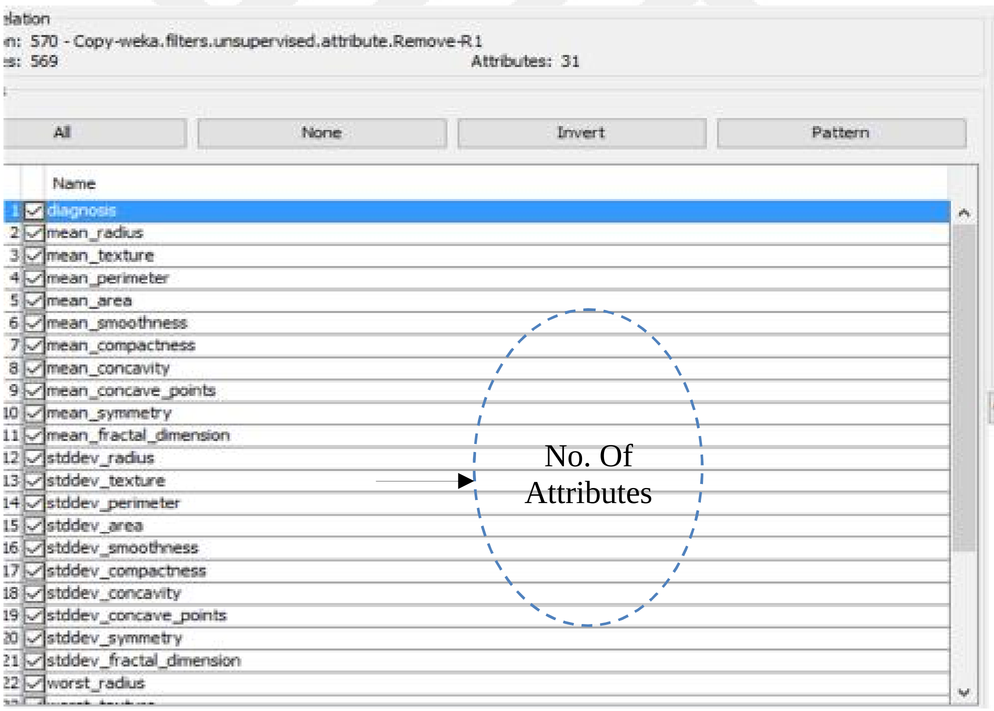


Figure 4.1. Input of features (attributes) for dataset

The work uses the WEKA workbench for classifying the breast tumor (Wisconsin breast cancer) data-set into the two types depend on the characteristic of the masses also from other features. The main purpose of the work is to test the success of the classifier of the results by choosing optimum parameters and to compare among them by a few known methods.

4.1. Decision Trees Result

4.1.1. J48 Algorithm Tree Results

In WEKA all rows of data are considered an instance and features of the data are naming attributes. The results show different parameters as numbers of correctly classified instances/ incorrectly classified instance, Kappa statistic, mean absolute, root mean squared error, relative absolute error, root relative squared error and confusion matrix.

From the Table 4.1. shown below, we acknowledge that by using the J48 algorithm the correct classification accuracy will read 93.145%, and the incorrectly classified accuracy reads 6.854%. From the same Table shows details of each type of errors.

Table 4.1. Summary of classification J48 algorithm includes types of error and accuracy

Correctly Classified Instances	530
Correctly Percentage	93.145%
Incorrectly Classified Instances	39
Incorrectly Percentage	6.854%
Kappa statistic	0.854
Mean absolute error	0.074
Root mean squared error	0.257
Relative absolute error %	15.836
Root relative squared error %	53.331

Lists difference evaluation standard results for J48 method on test data to tell us the types of rates, how much covers the classifiers under ROC space and the F- measure which it is harmonic mean of J48 algorithm precision and recall in Table 4.2.

Table 4.2. Performance analysis results of J48 algorithm on B.C data set

	TP rate	**FP rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Class**
	0.925	0.064	0.895	0.925	0.910	0.927	M
	0.936	0.075	0.954	0.936	0.945	0.927	B
Avg. W	0.931	0.071	0.932	0.931	0.932	0.927	

The confusion matrix for the classification results obtained from J48 classifier in Table 4.3.

A confusion matrix, also famed as an error matrix is a specific Table that gives a visualization of perform an algorithm, typically a supervised learning one also, in unsupervised learning it is usually called a matching matrix *(Stehman, Stephen V, 1979).*

Each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class (in some cases vice versa). From Table 4.3 can be seen that all the 569 instances belonging to contact class have been correctly classified.

Table 4.3. Prediction error matrix (confusion matrix) for J48 algorithm.

M	**B**	**<Classified as>**
196	16	M
23	334	B

4.1.2. Best First Tree Algorithm Results

The Table 4.4 shows the classification accuracy and various errors for best first algorithm on B.C data with two classes benign, malignant cancer and the measure of correcting the chance of agreement between classifications and real classes. From same table, we readable the numbers of correctly and incorrectly classified instances or the accuracy respectively is 92.970% and 7.029 %.

Table 4.4. Summary of classification for best first algorithm includes accuracy and types of error

Correctly Classified Instances	529
Correctly Percentage	92.970%
Incorrectly Classified Instances	40
Incorrectly Percentage	7.029%
Mean absolute error	0.081
Root mean squared error	0.252
Relative absolute error %	17.498
Root relative squared error %	52.210
Kappa statistic	0.849

In Table 4.5. consists of values analysis results of the best first algorithm on B.C data set.

Table 4.5. Performance evaluation of the best first algorithm with respect to 10-fold- cross- validation

	TP rate	FP rate	Precision	Recall	F-Measure	ROC Area	Class
	0.901	0.053	0.910	0.901	0.905	0.938	M
	0.947	0.099	0.942	0.947	0.944	0.938	B
Avg. W	0.930	0.082	0.930	0.930	0.930	0.938	

By this error matrix assigns letters M & B to the class values and provides expected class values in rows and predicted class values ("classified as") for each column in Table 4.6. We can clearly see that M = tested_negative and B = tested _positive for for best first algorithm.

Table 4.6. Prediction error matrix (confusion matrix) for best first algorithm

M	B	<Classified as>
191	21	M
19	338	B

4.1.3. Functional Tree Algorithm Results

From the Table 4.7. clear that the simulation result of the functional tree algorithm as correctly classified instance value reads 97.715 % and incorrectly classified instance the classification value reads 2.284 % more than, simulation errors on B.C data set.

Table 4.7. Training and simulation errors and accuracy of functional tree algorithm

Correctly Classified Instances	556
Correctly Percentage	97.715 %
Incorrectly Classified Instances	13
Incorrectly Percentage	2.284 %
Kappa statistic	0.950
Mean absolute error	0.042
Root mean squared error	0.145
Relative absolute error %	9.141
Root relative squared error %	29.983

Table 4.8. Different performance metrics running by using functional tree algorithm in WEKA

	TP rate	**FP rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Class**
	0.958	0.011	0.981	0.958	0.969	0.990	M
Avg.	0.989	0.042	0.975	0.989	0.982	0.990	B
W	0.977	0.031	0.977	0.977	0.977	0.990	

Table 4.8. reads different evaluation standard classification results obtained from the functional tree method on B. C test data set by WEKA.

Table 4.9. displays the error matrix for the classification results obtained from the functional tree classifier.

Table 4.9. Prediction error matrix (confusion matrix) for functional tree algorithm

M	**B**	**<Classified as>**
203	9	M
4	353	B

4.1.4. Alternating Decision Tree Algorithm Results

The different types of error measurement like the mean absolute error, root mean squared error, root relative squared error, relative absolute error, numbers of correctly classified instances and incorrectly classified instance, kappa statistic, accuracy that reads 94.024 % and the average for them all that reported in Table 4.10.

Table 4.10. Precise measurements of classification for the alternating decision tree algorithm obtaining by WEKA

Correctly Classified Instances	535
Correctly Percentage	94.024 %
Incorrectly Classified Instances	34
Incorrectly Percentage	5.975 %
Kappa statistic	0.871
Mean absolute error	0.083
Root mean squared error	0.201
Relative absolute error %	17.603
Root relative squared error %	41.759

Table 4.10. consists of values analysis results for alternating decision tree method on test data obtaining by WEKA.

Table 4.11. Performance evaluation of the alternating decision tree algorithm for B.C data set with respect to 10- fold- cross- validation

	TP rate	**FP rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Class**
	0.910	0.042	0.928	0.910	0.919	0.985	M
	0.958	0.090	0.947	0.958	0.953	0.985	B
Avg. W	0.940	0.072	0.940	0.940	0.940	0.985	

The confusion matrix for the classification results obtained from for alternating decision tree classifier in Table 4.12

Table 4.12. Prediction error matrix (confusion matrix) for AD algorithm

M	**B**	**<Classified as>**
193	19	M
15	342	B

4.1.5. Random Forest Tree Algorithm Results

Table 4.13 reads errors measurement, numbers of correctly and incorrectly classified instance and kappa statistic.

The accuracy of classification using random forest tree algorithm from Table 4.34 is 96.660 %.

Table 4.13. Summary classifier for random forest tree algorithm include accuracy & types of error

Correctly Classified Instances	550
Correctly Percentage	96.660 %
Incorrectly Classified Instances	19
Incorrectly Percentage	3.339 %
Kappa statistic	0.928
Mean absolute error	0.070
Root mean squared error	0.176
Relative absolute error %	15.144
Root relative squared error %	36.55

The performance matrix with different simulation results like TP rate, FP rate, and precision, recall, F-measure and ROC area during training and testing by using random forest tree algorithm method on B.C dataset reported through Table 4.14.

Table 4.14. Different performance metrics running by using RF algorithm in WEKA

	TP rate	FP rate	Precision	Recall	F-Measure	ROC Area	Class
	0.948	0.022	0.962	0.948	0.955	0.989	M
	0.978	0.052	0.969	0.978	0.974	0.989	B
Avg. W	0.967	0.041	0.967	0.946	0.967	0.989	

The confusion matrix for the classification results obtained from random forest tree algorithm classifier in the 4.15.

Table 4.15. Prediction error matrix (confusion matrix) for alternating decision tree

M	B	<Classified as>
193	19	M
15	342	B

4.1.6. Decision Stump Algorithm Results

Table 4.16 tells summary classification accuracy and types of error for B.C data set that include 2 classes for 569 cases by using decision stump algorithm. Furthermore, shows the numbers of correctly classified instances and incorrectly classified instances. From Table 4.34. The accuracy of classification using a decision stump algorithm is 88.927 %.

Table 4.16. Performance measures in training and test data set by using decision stump algorithm

Correctly Classified Instances	506
Correctly Percentage	88.927 %
Incorrectly Classified Instances	11.072 %
Incorrectly Percentage	63
Kappa statistic	0.755
Mean absolute error	0.169
Root mean squared error	0.313
Relative absolute error %	36.173

In schedule 4.16. reads different evaluation standard classification results obtained from the decision stump algorithm on B. C test data set by WEKA.

Table 4.17. Performance analysis results of decision stump algorithm on B.C data set

	TP rate	**FP rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Class**
	0.778	0.045	0.912	0.778	0.840	0.874	M
	0.955	0.222	0.879	0.955	0.915	0.874	B
Avg. W	0.889	0.156	0.891	0.889	0.887	0.874	

The confusion matrix for the classification results obtained from decision stump algorithm classifier in Table 4.18. From Table 4.18 can be seen that all the 569 instances which classified on the classes M and B.

Table 4.18. Prediction error matrix (confusion matrix) for decision stump

M	**B**	**<Classified as>**
5	47	M
6	341	B

4.1.7. Simple Cart tree Algorithm Results

Table 4.19 shows the classification accuracy and all types of error for simple cart tree algorithm on B.C data set that include 2 classes for 569 cases. Furthermore, Table 4.34 tells the accuracy of classification using simplecart tree algorithm from Table is 93.145 %.

Table 4.19. Summary classifiers for simple cart tree algorithms include accuracy & types of error

Correctly Classified Instances	530
Correctly Percentage	93.145 %.
Incorrectly Classified Instances	39
Incorrectly Percentage	6.854 %
Kappa statistic	0.852
Mean absolute error	0.084
Root mean squared error	0.251
Relative absolute error %	18.010
Root relative squared error %	52.097

Lists different evaluation standard results for stump method of test data in Table 4.20.

Table 4.20. Performance analysis results of simplecart algorithmon B.C data set

	TP rate	**FP rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Class**
	0.896	0.048	0.918	0.896	0.907	0.940	M
	0.952	0.104	0.939	0.952	0.946	0.940	B
Avg. W	0.931	0.083	0.931	0.931	0.931	0.940	

The confusion matrix for the classification results obtained simple cart tree classifier in Table 4.21.

From Table 4.21 could be seen that all the 569 instances belonging to contact class have been correctly classified.

Table 4.21. Prediction error matrix (confusion matrix) for simplecart tree

M	B	<Classified as>
190	22	M
17	340	B

4.2. Support Vector Algorithm Results

4.2.1. Support Vector with Normalized Poly Kernel

Table 4.22 shows the classification accuracies and all types of error for SMO with normalized poly kernel algorithm on B.C data set that include 2 classes for 569 cases.

Additionally from Table 4.35, the accuracy of classification using SMO normalized poly kernel algorithm is 96.485 %.

Table 4.22. Summary classifiers for SMO normalized poly kernel algorithm. include accuracy and types of error

Correctly Classified Instances	549
Correctly Percentage	96.485%.
Incorrectly Classified Instances	20
Incorrectly Percentage	3.514%
Kappa statistic	0.924
Mean absolute error	0.051
Root mean squared error	0.173
Relative absolute error %	10.919
Root relative squared error %	35.805

Lists difference evaluation standard results for SMO normalized poly kernel method on test data to tell us the types of rates, how much covers the classifiers under ROC space and the F- measure which it is harmonic mean of SMO normalized poly kernel algorithm for precision and recall in Table 4.22.

Table 4.23. Performance analysis results for SMO normalized poly kernel algorithm. B.C data set

	TP rate	**FP rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Class**
	0.939	0.020	0.966	0.939	0.952	0.991	M
	0.980	0.061	0.964	0.980	0.972	0.991	B
Avg. W	0.965	0.046	0.965	0.965	0.965	0.991	

Table 4.24 shows the confusion matrix for the classification results obtained from SMO normalized poly kernel algorithm classifier in Table 4.18.From Table 4.24 can be seen that all the 569 instances belinging to contact class have been correctly classified.

Table 4.24. Prediction error matrix (confusion matrix) for SMO normalized poly kernel algorithm

M	**B**	**<Classified as>**
199	13	M
7	350	B

4.2.2. Support Vector with Poly Kernel Algorithm

The different types of error measurement of SMO poly kernel algorithm like mean absolute error, root mean squared error, root relative squared error, relative absolute error, numbers of correctly classified instances and incorrectly classified instance, kappa statistic, and the average for them reported through the Table 4.25. Additionally, the Table 4.35 include the accuracy of classification for all types kernel functions. From that Table the SMO poly kernel algorithm is 97.188 %.

Table 4.25. Report classifiers for SMO poly kernel algorithm

Correctly Classified Instances	553
Correctly Percentage	97.188 %.
Incorrectly Classified Instances	16
Incorrectly Percentage	2.812 %
Kappa statistic	0.939
Mean absolute error	0.038
Root mean squared error	0.145
Relative absolute error %	8.213

Lists different evaluation standard results for SMO poly kernel method on reside test data on test data in Table 4.26

Table 4.26. Performance analysis results for SMO poly kernel algorithm. B.C data set

	TP rate	**FP rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Class**
	0.958	0.020	0.967	0.958	0.962	0.994	M
	0.980	0.042	0.975	0.980	0.978	0.994	B
Avg. W	0.972	0.034	0.972	0.972	0.972	0.994	

The confusion matrix shows the classification results obtained from SMO poly kernel classifier in Table 4.27.

Table 4.27. Prediction error matrix (confusion matrix) for SMO poly kernel algorithm

M	B	<Classified as>
203	9	M
7	350	B

4.2.3. Support Vector With PUK Kernel Algorithm Results

The Table 4.28 exhibts the classification accuracy and all types of errors for SMO puk kernel algorithm on B.C data set. From report clasiffiers, shows numbers of correctly classified instance / incorrectly classified instance.

The accuracy of classification using SMO puk kernel algorithm from Table 4.35 is 95.079 %.

Table 4.28. Summary classify for SMO PUK kernel algorithm. includes accuracy with errors

Correctly Classified Instances	541
Correctly Percentage	95.079 %
Incorrectly Classified Instances	28
Incorrectly Percentage	4.920 %
Kappa statistic	0.896
Mean absolute error	0.049
Root mean squared error	0.221
Relative absolute error %	10.589
Root relative squared error %	45.713

Lists different evaluation standard results for SMO puk kernel method on reside test data on test data in Table 4.29.

Table 4.29. Performance measures for SMO puk kernel models

	TP rate	**FP rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Class**
	0.972	0.062	0.904	0.972	0.936	0.976	M
	0.938	0.028	0.982	0.938	0.960	0.984	B
Avg. W	0.951	0.041	0.953	0.951	0.951	0.981	

The confusion matrix is a very important factor for calculating many performance measures such as, sensitivity and specificity.

In Table 4.30. shows the classification results obtained from the SMO puk kernel model.

Table 4.30. Prediction error matrix (confusion matrix) for SMO puk kernel algorithm

M	**B**	**<Classified as>**
206	6	M
22	335	B

4.2.4. Support Vector With RBF kernel Algorithm Results

Table 4.31 includes the experimental results for 569 instances for RBF algorithm on B.C data set in order to comparison with other existing classsification approaches.

From Table 4.34 the accuracy of classification using SMO RBF kernel algorithm is 98.242 %.

Table 4.31. Summary classifies for SMO RBF kernel algorithm include accuracy and types of error

Correctly Classified Instances	559
Correctly Percentage	98.242 %
Incorrectly Classified Instances	10
Incorrectly Percentage	1.757 %
Kappa statistic	0.962
Mean absolute error	0.017
Root mean squared error	0.132
Relative absolute error %	3.757
Root relative squared error %	27.418

Lists different evaluation standard results like ROC, F-measurs, etc. for SMO RBF kernel method on reside test data on test data in Table 4.32.

Table 4.32. Performance analysis results for SMO RBF kernel algorithm

	TP rate	**FP rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Class**
	0.962	0.006	0.990	0.962	0.976	0.978	M
	0.994	0.038	0.978	0.994	0.986	0.978	B
Avg. W	0.982	0.026	0.983	0.982	0.982	0.978	

The confusion matrix for the classification results obtained from SMO RBF kernel classifier in Table 4.33.

Table 4.33. Prediction error matrix (confusion matrix) for SMO RBF kernel algorithm

A	**b**	**<Classified as>**
204	8	M
2	355	B

4.3. Comparison of Algorithms

The accuracies of all classifiers have been represented in tables 4.34 and 4.35. Also, the time taken to test models of training and testing data for each D.T and SMO algorithms which reported through tables 4.36 and 4.37. The results can be shown the best accuracy in FT algorithm with 97.715% by the time taken by a second is 0.08 Sec. The SMO RBF algorithm is taken 0.04 Sec to reach 98.242% accuracy.

Table 4.34. Accuracy (%) for D.T algorithms

Algorithms	**Accuracy %**
FT	97.715
RF	96.660
AD	94.024
J48	93.145
SC	93.145
BF	92.970
DS	88.927

Table 4.34 Shows accuracy with best and worst percentage D.T algorithms, the best percentage D.T to classification B. C data set is FT algorithms with 97.715 % and the worst percentage can see in DS algorithms with 88.927 %.

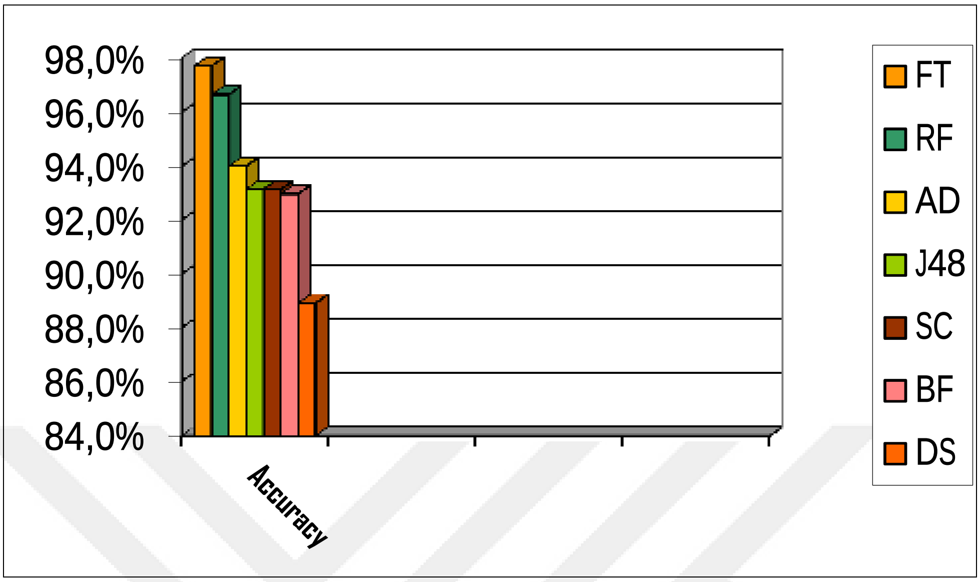


Figure 4.2. Graph showing the accuracy in percentage values for different D.T algor ithms

From the graphic 4.2. shows planned for accuracy with best and worst percentage D.T algorithms. FT achieved the highest results from the rest of the selected algorithms.

Table 4.35. Accuracy (%) of SVM - SMO algorithms

Algorithms	**Accuracy %**
SMO RBF	98.242
SOM PO	97.188
SMO N.PO	96.485
SMO PUK	95.079

Table 4.35. shows the comparison accuracy rate of SMOs algorithms with best percentage. It clearly from this Table the best classification on B.C data set is RBF kernel to have 98.242 % accurcy.

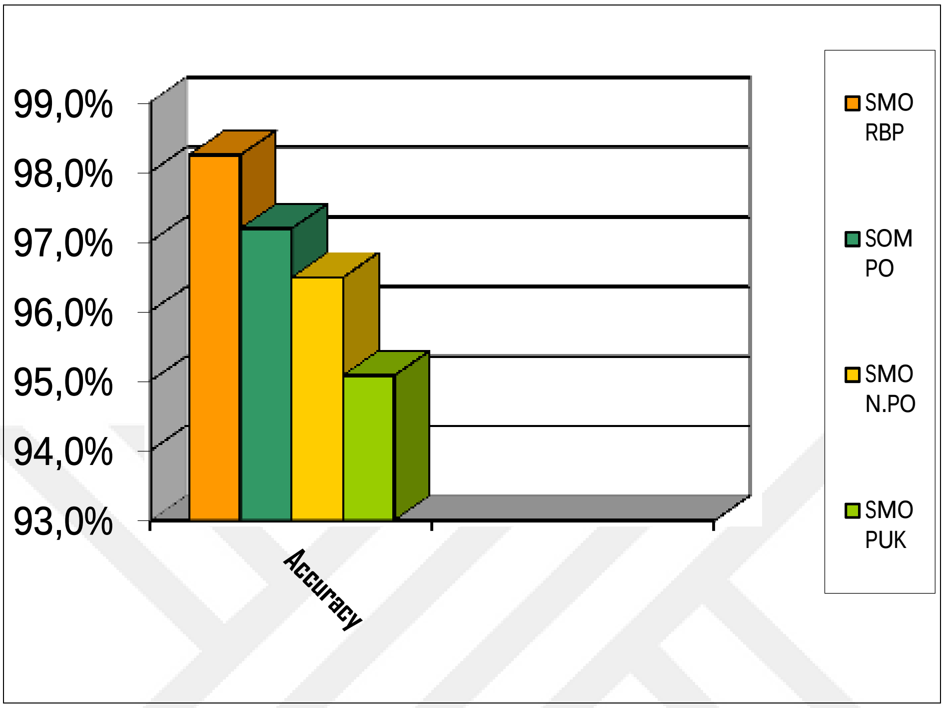


Figure 4.3. Performance correct instances analysis (accuracy) in percentage values For SMO algorithms

The Figure 4.3 shows the accuracy percentage of SMO algorithms.

Table 4.36. Time taken to test the model of D.T algorithms

Algorithms	**Time (Sec.)**
FT	0.08
RF	0.03
AD	0.10
J48	0.02
SC	0.09
BF	0.08
DS	0.01

The time taken to test model on training and testing data for each D.T algorithms in Table 4.36.

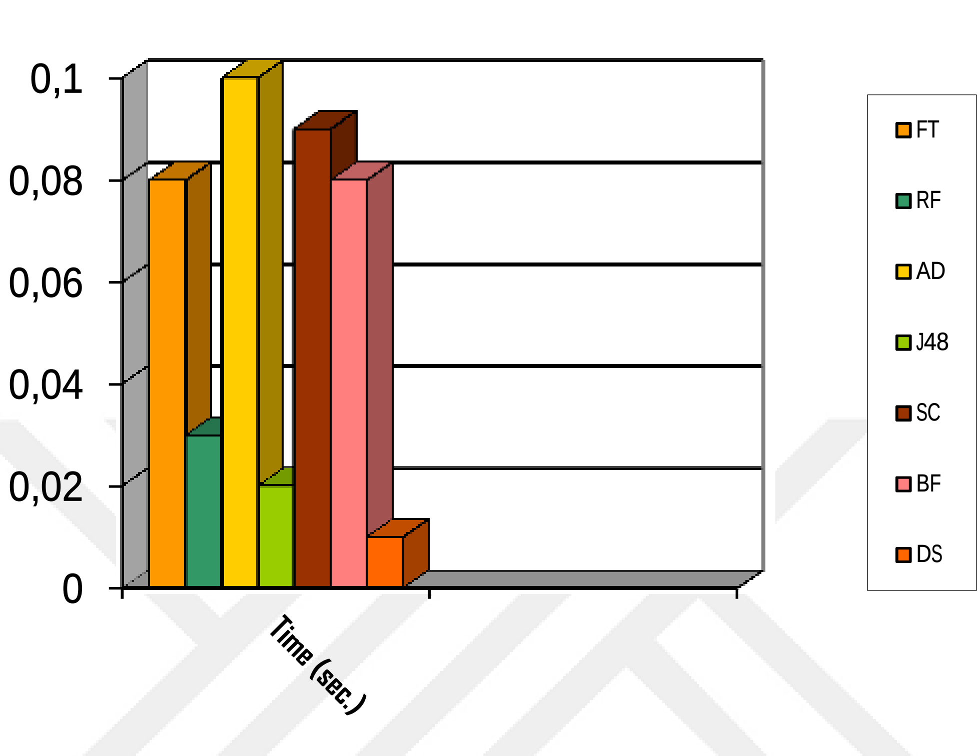


Fıgure 4.4. line chart for time taken to test model on training and testing data for each D.T algorithms

Table 4.37. Time taken to test the model of SVM-SMO algorithm

Algorithms	Time (sec.)
SMO RBF	0.04
SOM PO	0.01
SMO N.PO	0.14
SMO PUK	0.05

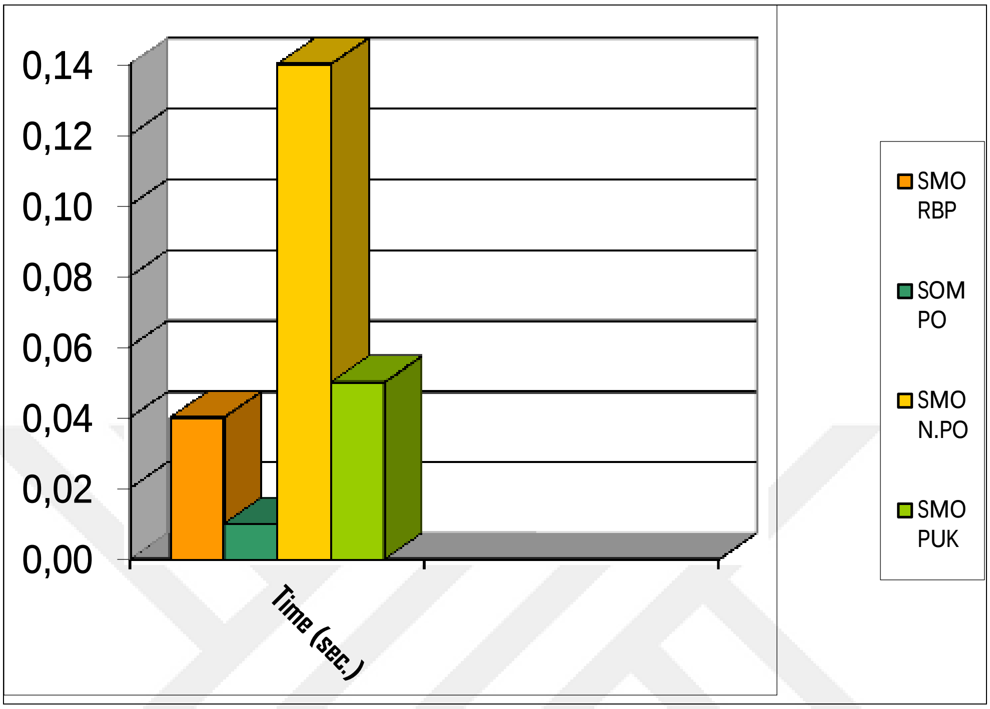


Figure 4.5. The time taken to test model on training and testing data for each SMO algorithms.

Table 4.38. Weighted average receiver operating characteristics for algorithms

Algorithms	**W.Avg.ROC**
FT	0.990
RF	0.989
AD	0.985
SC	0.940
BF	0.930
J48	0.927
DS	0.874
SMO RBF	0.978
SOM PO	0.994
SMO N.PO	0.991
SMO PUK	0.981

In the Figure 4.6 exhibits comparison weighted average receiver operating characteristics for algorithms for shows how much covers the classifiers under

ROC space. ROC curves provide a comprehensive and visually attractive way to summarize the accuracy of predictions.

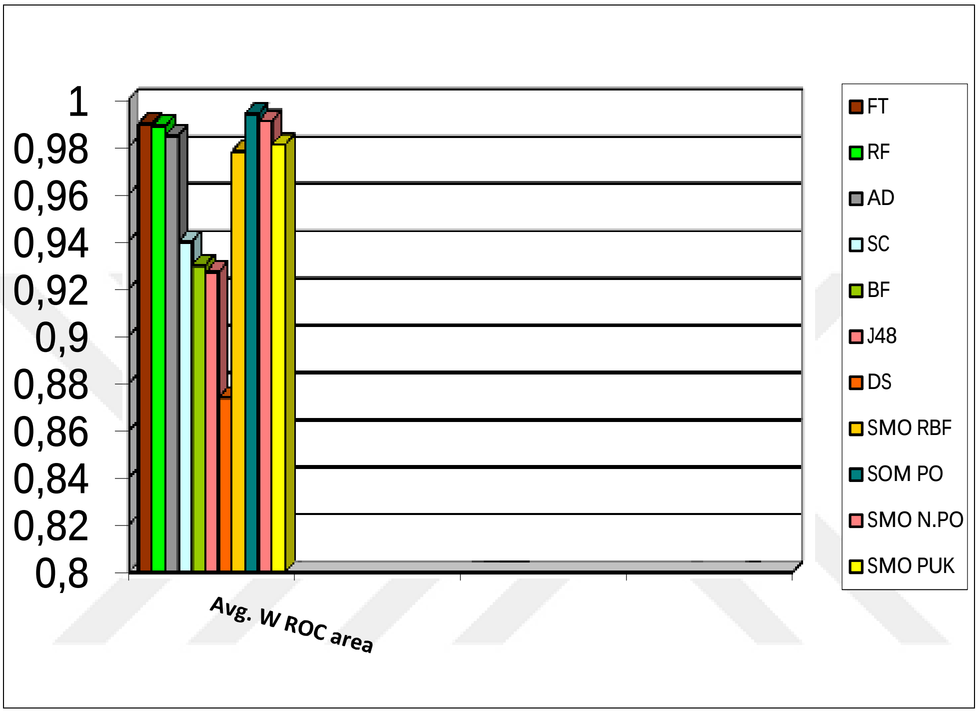


Figure 4.6. Graphical representation comparison weighted average ROC area for all algorithms.

4.4. Summary

This chapter demonstrates the results of classification, medical dataset by applying many methods via WEKA as a workbench. Analysis shows that the activity of these algorithms when used the B.C dataset for D.T and four types of kernel functions with SVM. İn other hand, many comparisons used according to precise measurements like Kappa Statistic which is the percentage of possible predictions correctly classified by chance, time taken to building the models and how much covers the classifiers under ROC space etc. More that we could say that the study proves with its results that supervised learning algorithms are suitable algorithms into the realm of cancer diagnosis with good output for most algorithms which used in this work.

5. DISCUSSIONS

This study proves that the forecast of the UCI breast cancer dataset is under the frames of classification algorithms in data mining. The eight classification algorithms, including J48, Function Tree, Random Forest Tree, AD Alternating Decision Tree, Decision Stump, Simple Cart, SVM (SMO) and Best First, were put to practical use as opposed to being theoretical to the breast cancer data set. The results accomplished in this research show near from perfect result with a high degree of accuracy when compared with previous results. Different types of experiments carried out to cover all available cases.

Many reasons to apply to decision tree algorithms don't need any factor framework, and thus are suitable for exploratory knowledge discovery in references *(Williams and et.all, 2015)*. DT could be deal with multidimensional dataset. Furthermore, SMO solves the QP problem analytically without any extra or additional matrix storage.

The experiential results showed that highest precision 98.242% for SMO RBF classifier with highest correctly numbers of instance 557, and 97.715% found in FT classifier with highest correctly numbers of instance 550. But, precision 88.927% is found in the decision stump with lowest correctly numbers of instance 509. The best of the classification results of all the algorithms is the realization of SMO RBF.

We could say that the study proves with its results that supervised learning algorithms are suitable algorithms into the realm of cancer diagnosis, for potential use in collaboration with the established medical dataset and to help avoid evasive diagnostic tests on patients.

Finally, it can be said the proposed technology has performed a fully compliant or competitive taxonomy on the given dataset. This Table shows that all research on the dataset with their results.

Table 5.1. Accuracy of reviewed articles

No	Researchers	Type of algorithms	Accuracy %	Workbench	Source of data
1	Bin Othman et al.*(Bin Othman, 2007)*	Bayes Net. Radial Basis Function Decision Tree and Pruning	89.714 87.428 85.714	WEKA	B.C data. Total of 6291 patients with dimension of 699 rows and 9 attributes.
2	Zhiyan Yang.p. et al. *(Zhiyan Yang,P. et, 2010)*	Decision tree	87.230	WEKA	164 controls Genetics of faculty of Medical Sciences or keepers and 94 cases
3	Sugandhi C. et al. *(Sugandhi C, 2011)*	SMO J48	77.343 73.828	WEKA	UCI machine Learning Repository
4	Shrivastava et al. *(Shrivastava, 2013)*	J48	98.141	WEKA	UCI machine Learning Repository
5	Shajahaan et al. *(Shajahaan, S. Syed, 2013)*	ID3 CART C4.5	92.990 92.420 95.570	WEKA	University of Wisconsin Hospitals
6	Williams et al. *(Williams, Kehinde, 2015*	J48	94.220	WEKA	The data contains 69 instances with 17 attributes along with the class labe
7	Venkatesan E. *(Venkatesan E., 2015)*	J48	99.000	WEKA	Swami Vivekananda Diagnostic Centre Hospital

Table 5.1. Contınuous

8	Layla Safwat Jamil *(Layla Safwat, 2016)*	J48 SMO Naïve Bayes	75.520 69.580 71.670	WEKA	Data set includes 201 instances of one class and 85 instances of another class
9	Nusaibah AL-SALIHY *(Nusaibah Kh. AL-SALIHY, 2017)*	FT RF AD J48 SC BF SMO RBF SMO PO SMO N.PO SMO PUK	97.715 96.660 94.024 93.145 92.970 88.927 98.242 97.188 96.485 95.079	WEKA	Breast Cancer Wisconsin dataset includes **569 instance** with **two classes**

From The Table 5.1. the researcher Safwat Jamil uses WEKA as a tool of classifying data and knowing (testing) its accuracy by classifying algorithms. But we note the difference in the accuracy of the classification in the case of the use of example algorithm J48 because data set includes 201 instances of one class and 85 instances of another class, but, the dataset which uses in this work include 569 instance with two classes. From that diffrences. We understand the classification depending on the volume and quality of the data, the number of cases contained in that data and how to process it prior to the process of classification.

6. CONCLUSIONS

This study includes several stages could be mentioned in a short phase, outlined what is breast cancer and its relation to age, sex and factors that increase the likelihood of contracting the disease, moreover the technical diagnosis and treatment of cancer to saved many lives of women. We asked our self what these techniques and why we choose them?

According to the general concept of prospecting evidence across many techniques used by researchers, begins to wonder about the benefits that organizations can achieve from using data mining? What organizations that began to invest this technique in their activities?

One of the professional sectors started to take advantage of this concept is health care. With the growth in electronic health records (electronic health records), more and more facilities and collect massive amounts of digital data to the patient, so that health care providers and researchers use data mining of massive data stores to detect previously unknown patterns and then use this information to build predictive models for improving diagnosis and health care outcomes.

Data mining is to see data from a new perspective and here are the added value and which unpredictable data given by traditional methods. Data mining gives these added values in numerical fields if used.

The primary goal of this thesis is to apply several algorithms that used in many research for the purpose of demonstrating the best algorithm with high-resolution and without losing the data classification (misclassify) or (poorly-classify) and taking into account the factors and parameters that contain work branch used.

By the way of example in SMO in RBF-kernel selected the C factor SMO -C 8. zero, the precision is 98. 2425%, because C factor impacts the trade-off between ratio of non-separable examples and complex, so that must be chosen by user.

C factor shows the trade-off between the margin maximization and errors on training data set.The mean of C complexly parameter tells a SVM-SMO optimization how much want to avoid misclassified for all training examples. The relationship between the values of C factor and margin hyper-plane, If the value of C factor was small, will cause the windows optimizer to look for a larger-margin separating hyper-planes, even if that excitable plane mis-classifies more factors. For very small of C, will get misclassified examples, often even if training data is linearly separable. Conversely, In case the value of C factor was large, the optimization will choose a smaller-margin excitable plane if that excitable plane does a much better work for getting all training points classified correctly. Generally an optimal C factor should be larger the size down of data, and the other way round, so if have tiny values for features, be sure to include very large values for the possible C values.

REFERENCES

Brıstol-Myers, Squibb Foundation and Corporate Social Responsibility, 2014.

Breast Cancer Organızatıon, 2015. http://www.breastcancer.org/ types/factors.

Breast Cancer Organızatıon, 2015. http://www.breastcancer.org/ risk/factors.

Breıman, L., 2001. Random forests, Machine Learning 45 (1), 5-32.

Bın Othman M. F , Yau, T. M. S., January 2007. Comparison of Different Classification Techniques Using WEKA For Breast Cancer, In 3rd Kuala Lumpur International Conference on Biomedical Engineering Springer Berlin Heidelberg, Pp.520-523.

Cobaın Ef, Hayes Df., 2015. Indications for Prognostic Gene Expression Profiling in Early Breast Cancer. Curr. Treat Options Oncol. Pp. 23.

Chaurasıa, Vıkas, And Saurabh Pal, 2014. “A novel Approach For Breast Cancer Detection Using Data Mining Techniques.” International Journal of Innovative Research in Computer and Communication Engineering vol.2, pp.2456-2465.

Dhamodharan, S., 2014. Liver Disease Prediction Using Bayesian Classification , Special Issues, 4th National Conference on Advance Computing , Application Technologies.

Duraıraj M. and Ranjanı V., 2013. Data Mining Applications In Healthcare Sector a Study. International Journal Sciences & Research Technology IJSTR, vol 10.

Freund Y. and Mason L, 1999. The Alternating Decision Tree Algorithm. Proceedings 16th International Conference on Machine Leaning, pp124-133.

Frank, E., Hal, M. A., And Wıtten I, 2016. The Weka Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufman.

Gehrke, J. Gantı,V., R. Ramakrıshnan, And Loh.W-Y, 1999. BOAT—optimistic decision tree construction. In Proceedings of the 1999 ACM SIGMOD international conference on Management of Data pp. 169–180). Philadelphia, PA, USA.

Goharıan & Grossman, 2016. Data Mining Classification, Illinois Institute Technology Internet. Available from: http://ir.iit.edu/~nazli/cs422/CS422-Slides/ DM-Classification.pdf.

Ian H. Wıtten Eıbe, Frank Mark A. Hall " 2011,Third Edition" "Data Mining Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Burlington, MA 01803, USA.

Iba, W., Langley, P. 1992, Induction of One-Level Decision Trees, In: Proceeding of ML-92, 9th International Conference Workshop on Machine Leaning, pp. 233–240.

Joao. Gama, 2004. Functional Tree Machine Learning. vol. 55 .pp. 219-250

Jıaweı Han, Mıchelıne Kamber And Jıan Peı, 2011. Data Mining Concepts and Techniques, 3rd ed. The Morgan Kaufmann Series in Data Management Systems.

Kaur, G., Chhara. A. 2014. Improved J48 classification algorithm for the Prediction Of Diabetes. Int. Jour. of Computer Applications., vol.98pp.13–7.

Kım, H. and Loh, 2001. Classification Trees With Unbiased Multi-Way splits, Journal of the American Statistical Association, vol. 96, pp.589-604.

Layla Safwat Jamil, 2016. International Journal of Engineering Sciences & Research Technology "Data Analysis on Data Mining Algorithms Using WEKA" : https://zenodo.org/record/59630/files/26.pdf.

Miller, G.A., 1956. The magical Number Seven Plus Or Minus Two: Some Limits On Our Capacity For Processing Information. Psychological Review, vol.63, pp.81–97.

N.Cristianini , J.Shawe-Taylor, 2001. “An Introduction To Support Vector Machines: And Other Kernel-Based Learning Methods. New York Cambridge University Press.

Nationalbreastcancer, 2016.http://www.nationalbreastcancer.org/breast-cancer-hormone-therapy.

Nıce Clinical Guideline, July 2014. http://patient.info/health/breast-cancer-leaflet.

Quinlan, J. R. 1986, Induction of Decision Trees. Machine Leaning vol. 1 pp. 81-106.

Quinlan, J. R., 1986 Induction of Decision Trees. Machine Learning. R. RASTOGI AND K. SHIM. Public: A decision tree classify that integrates building and pruning. In Proc. of 1996 International Journal on Very Large Databases (VLDB96) pp.81-106. .

R. D. Canlas Jr., 2009 “Data Mining in Medical Care: Current Applications and Issues”.

Rajesh K, Anand S., 2012 April. Analysis of SEER Dataset for Breast Cancer Diagnosis using C4.5 Classification Algorithm. International Journal of Advanced Research in Computer and Communication Engineering, vol.1 pp. 72–7.

Sawyer Sl, Tian L, Kahkonen M, et al., 2015. Balletic mutations in BRCA1 cause a new Franconia anemia subtype. Cancer Discovery. vol.2, pp.135-142.

S. S. Aman, Kumar Sharma, 2011. "A Comparative Study of Classification Algorithms for Spam Email Data Analysis," International Journal on Computer Science and Engineering, vol. 3, pp- 1890-1895.

Shajahaan S. Syed, S. Shanthi, and V. M. Chitra, 2013. “Application of Data Mining Techniques to Model Breast Cancer Data International Journal of Emerging Technology and Advanced Engineering vol.11, pp. 362-369.

Sharma,P., Ratnoo, S. 2014, A review on Discovery of Classification Rules Using Pittsburgh Approach, International Journal Artificial Intelligent. Know l. Discover, vol. 4 pp.16.

Shi. H, 2007. Best-first Decision Tree Learning. Master Thesis, http://hdl.handle.net/10289/2317.

Shiego Abe, Springer 2005. Book, “Support Vector Machines For Pattern Classification” Advances in Pattern Recognition. Springer-Verlag New York, Inc. Secaucus, NJ, USA.

Stehman, Stephen V. 1997. "Selecting And Interpreting Measures of Thematic Classification Accuracy". Remote Sensing of Environment.

Sugandhi C, Ysodha P, Kannan M, 2011. Analysis of a Population of Cataract Patient Database in WEKA Tool, International Journal of Scientific and Engineering Research.

T. Nguyen, A. Khosravi, D. Creıghton, S. Nahavandi, 2014. “Medical Diagnosis By Fuzzy Standard Additive Model With Wavelets,” in IEEE International conference on Fuzzy Systems,(IJCNN) pp. 1-8.

Torgo, L. 1997. Functional Models For Regression Tree Leaves. In D. Fisher, Machine. Learning, Proceedings of the 14th International conference pp. 385–393

Tversky A., 1974. Judgment Under Uncertainty: Heuristics And Biases. Science, vol. 185, pp.1124–1131.

Uci Mach. Learn, 1995. Repository Center for Machine Learning and Intelligent Systems. Breast cancer Wisconsin (diagnostic) dataset.

Venkeatesan, E., and T. Velmurugan, 2015. “Performance analysis of decision tree algorithms for breast cancer classification.” Indian Journal of Science and Technology vol.29, pp.1-8

V.N. Vapnik, Springer, New York, 1995. Book, The Nature Of Statistical Learning Theory.

Williams, Kehinde, Peter Adebayo Idown, Jeremiah Admoolabalogun, Andadeniranisholaoluwaranti, 2015. “Breast Cancer Risk Prediction Using Data Mining Classification Techniques.” Transactions on Networks and Communications vol. 2, pp. 1-11.

CURRICULUM VITAE

Nusaibah Khalid ABDULMAJEED, she was born 11.4.1988 Iraqi - Baghdad. She got her B.Sc. in Electrical and Electronic Engineering from Al Anbar University in 2008-2009. Then she worked as an intern with a training certificate in electrical maintenance supported by the International Committee of the (Red Cross (ICRC 2012 and an ECG device at the College of Technology in Baghdad 2013. She worked as the Manager of the Maintenance section in the women's hospital and children in al- Ramadi. She is completing her Master's degree in Electrical and Electronics Engineering at Çukurova University, Adana, Turkey. Her research interests include: Software Engineering, Biomedical dataset and Image Processing, Computer Vision, Artificial neural networks and Machine Learning algorithms

APPENDIX

76

Classifying Breast Cancer by Using Decision Tree Algorithms

Nusaibah Khalid ABDULMAJEED
Department of Electrical and Electronics Engineering,
Cukurova University, Adana,
01330, TURKEY
(+90) 5340643270
nusaibah.shather@yahoo.com

Turgay IBRIKCI
Department of Electrical and Electronics Engineering, Cukurova University, Adana,
01330, TURKEY
(+90) 322 338 68 68
ibrikci@cu.edu.tr

ABSTRACT

Nowadays, datasets are growing daily to obtain knowledge from big dataset. Extraction operation of useful information from the dataset is called data mining that is one of the major techniques to get the diagnostic results especially in medical care fields as breast cancer. Breast cancer is one of the widespread cancers among women if matched with all other tumors all over the world. Classification is widely used in most important and necessary tasks in the real life applications in all fields. This technique has ability of detecting very similarities/differences that a human analyst may be not notice and therefore create and introduction more accurate/useful categories. This study presents comparison and analyses breast cancer dataset by using classification decision tree algorithms. Decision tree algorithms are applied to these algorithms which are J48, Function Tree, Random Forest Tree, AD Alternating Decision Tree, Decision stump and BF Best First. A computationally efficient classifies of these DT algorithms by employing Waikato Environment for Knowledge Analysis (WEKA) that is development program which includes a set of machine learning algorithms. These masses included 569 with 357 benign and 212 malignant cases with 32 attributes to test and proof the differences among the classification

Breast cancer starts when cells in the breast begin to grow out of control [14]. Breast cancer is non homogeneous disease and the emergence of danger dependent on several factors, like tumors size, histological grade, regional lymph a node, involvement, and expression of estrogen (ER) and progesterone hormone receptors (PgR), as well as human epidermal growth factor receptor (HER2) amplification [18]. Although breast cancer-related mortality has decreased since the 1980s, due to advances in early detection and adjuvant therapies, breast cancer is still increasingly common diagnosed cancer in Canadian women, with an estimated 25,000 new cases this year [2]. Several types of analytical software are available by data mining; one main method of data mining is a classification. A classification in general is a method to find a form which shows and recognize data classes to aim be able to use a form to forecasting the class of objects whose class labels are unknown [11][7]. A classification is path using for help and support medical staff in their decision making operation to discovering illnesses and useful to study the influence income, costs, and operating efficiency, but at the same time keeping high-quality care [3]. Predictions by

methods or algorithms. These results found by manner which involves reserving a particular sample of a medical dataset on which do not train the model. The decision tree classification forms forecast breast tumor with lower error average and higher precision of correctly classified cases 97.7%. The predicted accuracy correctly classified instances for decision stump algorithm 88.0% model is the lowest of all.

Keywords
Machine Learning; Decision Trees; Data Mining; Classification; Biomedical Information; WEKA.

INTRODUCTION

During recent years, the prevalent using of computers in today's society means that massive amount of data is kept electronically. Also in past the decade has seen an explosive growth in biomedical research so from this development which become relates into all facets of modern life. We get large data sets and is a useful resource of the pretty tools are obtainable putting it to use. The process of extracting useful information from set is called data mining. Data mining software is one of many analytical tools in order to analysis the data. This technique so useful because the software has the ability of detecting very minute similarities / differences that a human analyst may be not notice and therefore create and introduction more accurate / useful categories. In other side, data mining is the operation of finding connections / samples among many fields in huge associated data sets [12]. Data mining is described in the medical field is a powerful technology with huge potential it offers benefits to caregiver, patients, medical care collective, searchers, and insurers. It is very sorrowful to know that more than one million females is influenced by breast tumor, the most widespread aggressive cancer in females and among them 600,000 proved to be deadly every year.

machine learning applications are vast; one such particular application to be investigated is in regards to classifying whether a breast tumor is malignant or benign. The primary objective of this work addresses the various classification decision trees algorithms and comparison among them such as time is taken, accuracy and measurements of error rates. In current work using WEKA workbench for classifying the breast tumor dataset into the two types depend on the characteristic of the masses also from other features

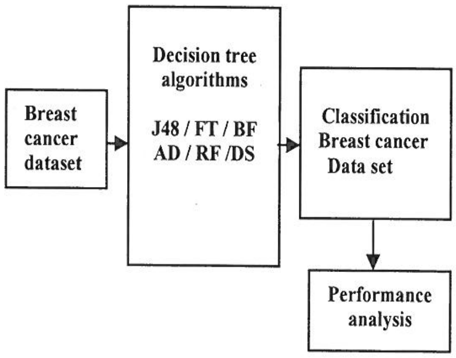


Figure 1. Methodology

9][12]. C4.5 algorithm carries out J48 to obstetrics strips C4.5 decision tree. To classify a dataset by a perfectly as possible, J48 algorithm builds decision trees for data from a set of training data. J48 classifier then visits all decision-node frequently after that chooses the most active splits until each leaf is pure and no more any splits are obtainable [15].

MATERIALS AND METHODS

Several classification functions that have been suggested by searchers in the area of classification applications in a breast tumor dataset using decision tree algorithms via using WEKA toolbox that is written in Java programming language. DT is one of the easier and important types of algorithm for forecasting, modeling machine learning. DT divided into two sets as binary tree and other according a result, that give away non-binary multi-way trees [10]. For classification of non-homogeneous disease dataset is non-binary to good activity, efficacy by supervised learning algorithms. The Figure1 also shows the methodology of the study.

2.1. Data and Features

The dataset had taken from Wisconsin Breast Cancer Data from the UCI Machine Learning Repository [17]. Attribute is data field, acting as a characteristic or feature of a data object [8]. A feature computed for each cell namely ID number is a 1st column and the 2nd column are classes naming which are type's diagnosis illnesses (B, Benign or M, Malignant). The rest of 30 features are properties of cells with

- Mean.
- Standard errors.
- Worst values.

of the radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and the fractional dimension to each

2.2.2 *Best First Algorithm*

Best First(BF) tree algorithms a form of decision tree learning, and may do almost of the characteristics of standard decision learning [21]. The naming best-node is saying for node that divides leads to the utmost limit of pollution among all nodes that helpful for splitting. It can deal with categorical and numerical variables.

2.2.3.Functional Tree Algorithm

Functional Tree (FT) is one of form of multivariate tree [6]. Classifier for buildup functional tree [16] that classification trees that may have logistic regression tasks on the internal nodes / leaf nodes. The functional tree can handle with binary and non-binary trees which called multi-ways or multi-classes targeted variables, missing or null values, numeric and nominal attributes.

2.2.4 Alternating Decision Tree Algorithm

AD algorithm was created by Yoav Freund and Llew Mason [23] for classification that different from the other algorithms because AD includes two types nodes are decision / forecast nodes. Each decision-node contains a division test, whereas all forecast nodes contain a real number for valued [22].

2.2.5. Random Forest Tree Algorithm

The RF algorithm, also sometimes called random decision forest [19] is active in many various open source programs. RFs are fairly strong with relatively straightforward applications for inexperienced users. RF increasingly for ML has achieved broad cell nucleus for 569 patients.

Dataset Characteristics:	Multivariate	No. of Attributes:		32
Attribute Characteristics:	Real	No. of Instances:		569
Associated Tasks	Classification	No. of Classes:		2

Table 1. Description of dataset

2.2. Classification Algorithms

A classification is one main technique of data mining. These techniques useful find hidden information. Datasets classification is in two-step practicability, where a classification model is built up called a learning step and classification step where the model used to foretell classes naming for giving dataset [8]. DT has another naming that is classification trees, that widespread used in classification implementation sand several expansions/differences have been suggested [13]. To make a decision by this method and build up classification models, decision tree break-down a dataset into smaller and smaller sets whilst keeping in the same time linked decision tree is gradually progressing. Definitive outcome is a tree with decision-nodes also leaf-nodes. A perfect outcome to make a decision-node in a tree that matches to best predictor, called root-node [12].

popularity in a last few years, particularly in the field of illnesses because RF offers two aspects that are very useful in data mining: high forecasting accuracy and new information on variable importance for classification methods[1]. A Supervised RF algorithm can be applied to solve unsupervised learning problem [19]. These advantages make the RF as a suitable algorithm for enabling incremental learning. RF classifier is data mining method developed by Beriman and Cutler that is fast in learning and, runs efficiently in big datasets [1].

2.2.6. Decision stump Algorithm

A decision stump [5] is fundamentally a decision tree with only one single split. A tree in this algorithm divided at the root-level, which depend on a specified feature value pair. Sometimes this algorithm also called 1-rule [20].

3. EXPERIMENTAL RESULTS

The results of this study are given in Table 2, Table 3 and Table 4 with different parameters. These results are with six algorithms for the medical dataset as input UCI breast cancer which outline in the section 2.1. The WEKA workbench [4] is data mining environment which contains many different machine learning

2.2.1 J48 Algorithm

J48 it's one of widespread decision tree algorithms, because it's actives with specific attributes, lost or missed values of the dataset, by this algorithm can be increased a precision by pruning

algorithms. Selected decision trees are compared between these decision tree algorithms. The validation is 10 that given significant expression that applied in these measures that are described as shown.

-**True Positive** *(TP)* of positive samples of correctly forecasted to be positive.

-**True Negative** *(TN)* of negative samples of correctly forecasted to be negative.

-**False Positive** *(FP)* of negative samples erroneously forecasted to be positive.

-**False Negative** *(FN)* of positive samples erroneously forecasted to be negative.

-*N* number of samples instances.

Precision: is a measure of accuracy. It is ration of the true positive-number to total number for forecasted positive samples

$$\frac{TP}{TP+FP} \qquad (1)$$

Recall: is measure of accuracy. It is the proportion of positive samples that were correctly identified to the total number of positive samples. It is also known as sensitivity or true positive rate (TPR).

$$\frac{TP}{TP+FN} \qquad (2)$$

Accuracy: is a measure of ratio for correctly forecasting.

$$\frac{TP+TN}{TP+TN+FN+FP} \qquad (3)$$

F-Measure: is the mean of precision t and

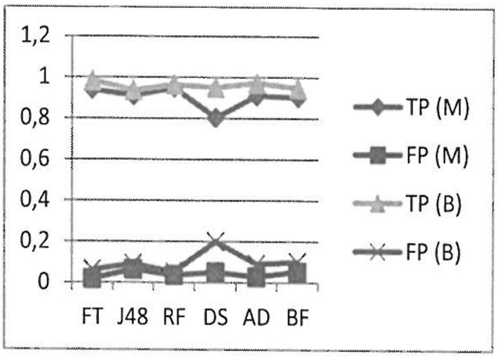


Figure 2. The results of each class's values.

Algorith ms	TP/Av g.	FP/Av g.
FT	0.967	0.045
J48	0.926	0.080
RF	0.958	0.046
DS	0.895	0.143
AD	0.949	0.067
BF	0.930	0.082

Table 3 . Results of averages TP, FP for classification classes

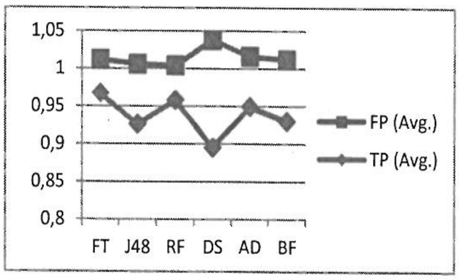


Figure 3. Results of average of FP and TP

Algorithm s	Time (sec.)
FT	0.03
J48	0.22
RF	0.15
DS	0.13
AD	0.11
BF	0.04

Table 4. Results time of the decision tree

recall. It is an important measure as it gives equal importance to accuracy and recall.

$$\frac{2*Recall*precision}{precision+Recall} \quad (4)$$

Receiver Operating Characteristic (ROC) Curve: is effective method of evaluating the quality or performance of diagnostic tests. TPR is plotted along the y axis and FP is plotted along the x axis. Performance of each classifier represented as a point on the ROC curve.

Algorithms	TP	FP	Classes
FT	0.939	0.017	M
	0.983	0.061	B
J48	0.910	0.064	M
	0.936	0.090	B
RF	0.948	0.036	M
	0.964	0.052	B
DS	0.802	0.050	M
	0.950	0.198	B
AD	0.910	0.028	M
	0.972	0.090	B
BF	0.901	0.053	M
	0.947	0.099	B

Table 2. Results of TP, FP for each classes with algorithms

Decision Tree	Precision	Recall	F-Measure	ROC Curve
FT	0.977	0.977	0.977	0.990
J48	0.934	0.933	0.933	0.931
RF	0.967	0.967	0.967	0.989
DS	0.891	0.889	0.887	0.874
AD	0.940	0.940	0.940	0.985
BF	0.930	0.930	0.930	0.938

algorithms for classification

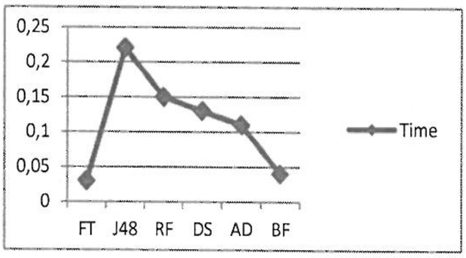


Figure 4. Comparison of the decision tree algorithms according to their completed time

Algorithms	Numbers of correctly classified instance	Numbers of Incorrectly classified instance
FT	550	19
J48	527	42
RF	545	24
DS	509	60
AD	538	31
BF	530	39

Table 7.Results no. of correctly/incorrectly instances the decision tree algorithms for classification analysis

Table 5. Comparison of decision tree classification algorithms

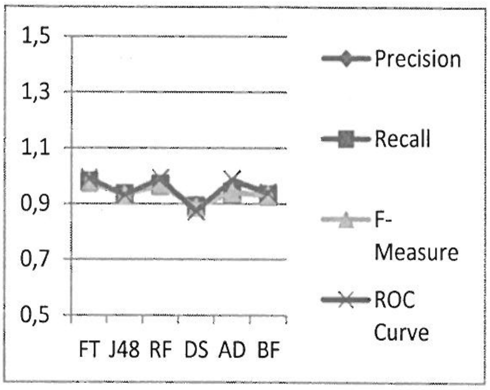


Figure 5. Comparison of decision tree according to Precision, Recall, F-Measure, ROC Curve

Algorithms	Accuracy
FT	97.7%
J48	93.1%
RF	96.6%
DS	88.0%
AD	94.0%
BF	92.0%

Table 6. Accuracy of decision tree algorithms

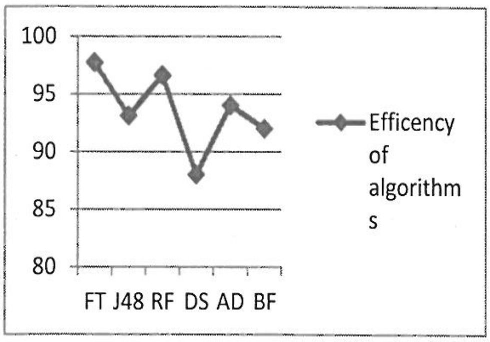


Figure 6..Performance correct instances analysis (Accuracy) in percentage values

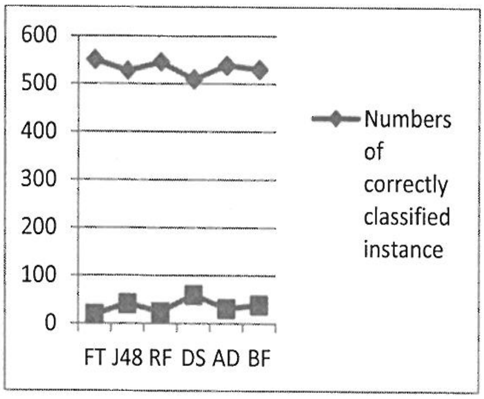


Figure7. Performance no. of correctly/incorrectly instances the decision tree algorithms for classification analysis

CONCLUSION

This study proves that the forecast of the UCI breast cancer dataset is under the frames of classification algorithms in data mining. Many reasons to apply decision tree algorithms in this study like don't need any factor framework, and thus are suitable for exploratory knowledge discovery. DT could be deal with multidimensional dataset. Depending on precision for J48, FT, AD, Random Forest, BF and Decision stump algorithms using different precision measures in section3. The experiential results showed that highest precision 97.7% found in FT classifier with highest correctly numbers instance (550) but, precision 88% is found in Decision stump with lowest correctly numbers instance (509). The best on the classification results of all the six algorithms, realization of FT better than another five algorithms. We can say that it is clear that there are promising results in the area of applying supervised learning algorithms into the realm of cancer diagnosis, for potential use in collaboration with the established medical tests and to help avoid evasive diagnostic tests on patients.

FUTURE WORK

WEKA IT IS AN OPEN SOURCE PROGRAM CONTAINS A LOT OF CLASSIFICATION ALGORITHMS AS SUCH AS SUPPORT VECTOR MACHINE NOT LIMITED TO D.T. SO WE'RE IN OUR FUTURE WORK USING OTHER MODELS AND THE EFFECTIVENESS OF ITS WORK WITH HUGE MEDICAL DATA WITHIN THIS AREA. WE CAN USE OTHER SURFACES (APPLICATIONS) LIKE KONWLEDAGEFLOW, NOT ONLY EXPLORER.

ACKNOWLEDGMENTS

This research is supported by Cukurova University Research Funds under ***FYL-2016-6805***. We are thankful to the fund that greatly assisted the research to provide our knowledge in this paper.

REFERENCES

[1] Breiman, L., Random forests, Mach. Learn 45(1), 5-32, 2001.

[2] Canadian Cancer Society's Steering Committee on Cancer Statistics. Canadian Cancer Statistics 2015 Toronto Canadian Cancer Society 2015

[3] Carol J. Buck, Step-by-Step Medical Coding, Elsevier Health Sciences (Book), 2016 Edition.

[4] Frank E, Hall M, Trigg L, Holmes G, Witten IH Bioinformatics. 2004, 12; 20(15):2479-81

[5] Iba, Wayne; and Langley, Pat 1992; Induction of One-Level Decision Trees, in ML92: Proceedings of the Ninth Int. Conf. on Mach. Learn, Aberdeen, Scotland, 1–3 July 1992, San Francisco, 233–240.

[6] J. Gama, Functional trees, Mach. Leaning, 2004.

[7] J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2001.

[8] Jiawei Han, Micheline Kamber and Jian Pei Data Mining: Concepts and Techniques, 3rded.The Morgan Kaufmann Series in Data Management Systems; 2011, 40.

[9] Kaur G, Chhabra A. Improved J48 classification algorithm for the prediction of diabetes. Int. Jour. of Computer Applications. 2014 , 98(22)13–7.

[10] Kim, H. and Loh, Classification trees with unbiased multi-way splits, Jour. of the American Statistical Association, W.-Y.2001, vol. 96, 589-604.

[11] P. Sharma and S. Ratnoo, "A review on Discovery of Classification Rules using Pittsburgh Approach," Int. J. Artif. Intell. Know l. Discov., vol. 4(3) 16, 2014.

[12] Quinlan, J. R. Induction of Decision Trees.

[20] Shi H, Best-first Decision Tree Learning. Master Thesis, 2007

[21] Vanassche, A., Krzywania, D., Vaneyghen, J., Struyf, J., Blockeel, H.: First order alternating decision trees. In: 13th Int. Conf. on Inductive Logic Programming, 116–125 2003.

[22] Yoav Freund and Llew Mason.The Alternating Decision Tree Algorithm. Proceedings 16th Int. Conf. on Mach. Learn, 124-133 1999.

Mach. Learn. 1986, 81-106.

[13] R. Rastogi and K. Shim. Public: A decision tree classifier that integrates building and pruning. In Proc. of 1996 Int. Conf. on Very Large Databases(VLDB96), 1998.

[14] Rajesh K, Anand S. Analysis of SEER Dataset for Breast Cancer Diagnosis using C4.5 Classification Algorithm. Int. Jour, of Advanced Research in Computer and Communication Engineering. 2012 Apr; 1(2):72–7.

[15] Tarek Gaber, Aboul Ella Hassanien, Khaled Shaalan, Ahmad Taher Azar, FahmyTolba Proceedings of the Int. Conf. on Advanced Intelligent Systems, 2016.

[16] Torgo, L. 1997. Functional models for regression tree leaves. In D. Fisher, Mach. Learn, proceedings of the 14th Int. conf.,385–393

[17] UCI Mach. Learn Repository: Center for Mach. Learn. and Intelligent Systems. Breast cancer Wisconsin (diagnostic) dataset: 1995.

[18] Cobain EF, Hayes DF. Indications for prognostic gene expression profiling in early breast cancer.Curr. Treat Options Oncol. 2015;16(5):23.

[19] *Ho, Tin Kam 1995. Random Decision Forests(PDF). Proceedings of the 3rd Int. Conf. on Document Analysis and Recognition, Montreal, 278–282.*

Authors' background

Your Name	Title*	Research Field	Personal website
Nusaibah Khalid ABDULMAJEED	Master Student	Data mining, neural networks	
Turgay İbrikçi	Associate Professor	Neural networks, Bioinformatics	http://eemb.cu.edu.tr/eng/detay.aspx?pageid=1488

***This form helps us to understand your paper better, the form itself will not be published.**

***Title can be chosen from: master student, PhD candidate, assistant professor, lecture, senior lecture, associate professor, full professor**