



**T. C.**

**ALTINBAŐ UNIVERSITY**

**GRADUATE SCHOOL OF SCIENCE AND ENGINEERING**

**Using Data Mining For Classification of Breast Cancer**

**Farah Sardouk**

**M. Sc. Thesis**

**Supervised by Assoc. Prof. Dr. OĐUZ BAYAT**

**Co-Supervisor: Assist. Prof. Dr. Adil Deniz Duru**



# **USING DATA MINING FOR THE CLASSIFICATION OF BREAST CANCER**

**By**

**FARAH SARDOUK**

Submitted to the Graduate Faculty of  
Science and Engineering in partial fulfillment  
Of the requirements for the degree of  
Master of Electrical and Computer Engineering

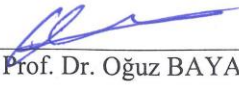
**ALTINBAŞ UNIVERSITY**

**2018**

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

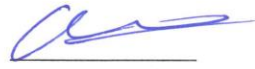



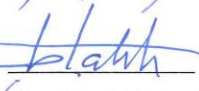
  
Asst. Prof. Dr. Adil Deniz DURU

Co-Supervisor

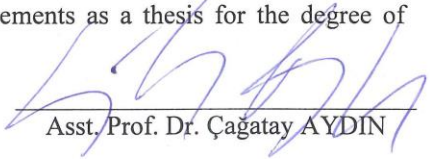
  
Assoc. Prof. Dr. Oğuz BAYAT

Supervisor

Examining Committee Members


Assoc. Prof. Dr. Oğuz BAYAT	School of Natural Science and Engineering, Altınbaş University	
Prof. Dr. Osman Nuri UÇAN	School of Natural Science and Engineering, Altınbaş University	
Asst. Prof. Dr. Adil Deniz DURU	Physical Education and Sports, Marmara University	
Assoc. Prof. Dr. Çağatay AYDIN	School of Natural Science and Engineering, Altınbaş University	
Prof. Dr. Hasan Hüseyin BALIK	Air Force Academy, National Defence University	

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

  
Asst. Prof. Dr. Çağatay AYDIN

Head of Department

Approval Date of Graduate School of  
Science and Engineering: 17 / 01 / 2019

  
Assoc. Prof. Dr. Oğuz BAYAT

Director

## **DECLARATION**

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



**FARAH SARDOUK**

## ACKNOWLEDGEMENT

As they say: It's not going to be easy, but it's going to be worth it. I would like to thank my doctor Adil Deniz since he proves to me that our hard work will be worth it. He was always giving us examples of how to use research in saving people lives and in improving the economics. Furthermore, special thanks to Prof. Oguz Bayat for his constant support during my research. I am grateful to do my research in this beautiful country Turkey so I would like to thank the Turkish community for their hospitality.

Finally, I would like to thank my family especially my grandmother (spirit) who has raised me and taught me to make a difference! Big thanks to my mother, my father, my husband and my brothers for their continuous help.

## **ABSTRACT**

# **Using Data Mining for The Classification of Breast Cancer**

**Farah Sardouk**

[M.S.], [Electrical and Computer Engineering], Altınbaş University,

Supervisor: Dr. Oğuz Bayat

Co-Supervisor: Dr. Adil Deniz Duru

Date: [December 2018]

Pages: 76

Among the most critical issues that threaten human health in current days is breast cancer, [1] the researches reveal that around 12.4 percent of women in the United States are more susceptible to this incident. According to the publications of leading health organizations in the world, WHO reveals that breast cancer is the most propagated disease among women and it may end with mortality. The precautions and regular investigations are the options for preventing this cancer, furthermore, the recognition of the same may begin at the earliest for combating purposes. From data science perspectives, data mining technology is used to uncover the disease according to some parameters like BMI, age and sugar routine database. The deployment of those technologies has resulted in considerable results that may help for breast cancer aid.

In this research Coimbra dataset is collected and studied according to 10 predictors. We used these predictors to estimate if breast cancer is occurring or not. The 6 algorithms used are compared according to their performance in WEKA and in MATLAB. The comparison is useful to prove the possibility of using Data Mining algorithms to help Medicine decision engine with good precision.

## ÖZET

### Meme Kanserinin Sınıflandırılması için Veri Madenciliğini Kullanmak

**Farah Sardouk**

**Elektrik ve Bilgisayar Mühendisliği, Altınbaş Üniversitesi**

**Danışman: Dr. Oğuz Bayat**

**Eş Denetçi: Dr. Adil Deniz Duru**

**Tarih: Aralık, 2018**

**Sayfa: 76**

Günümüzde insan sağlığını tehdit eden en kritik konular arasında meme kanseri bulunuyor. Araştırmacılar, Amerika Birleşik Devletleri'ndeki kadınların yaklaşık yüzde 12,4'ünün bu olaya daha duyarlı olduğunu gösteriyor. Dünyanın önde gelen sağlık örgütü "Dünya Sağlık Örgütü (WHO) yayınlarına göre, meme kanserinin kadınlarda en çok yayılan hastalık olduğunu ve ölümlerle sonuçlanabileceğini ortaya koymaktadır. Önlemler ve düzenli araştırmalar, bu kanseri önleme seçenekleridir, ayrıca bu hastalığın tanınması, hastalıklarla mücadele için erken aşamalarda başlayabilir. Veri bilimi perspektifinden veri madenciliği teknolojisi, BMI (vücut kitle indeksi), yaş ve şeker rutin veritabanı gibi bazı parametrelere göre hastalığı ortaya çıkarmak için kullanılır. Bu teknolojilerin yayılması, meme kanseri çaresine yardımcı olabilecek önemli sonuçlara neden olmuştur.

Bu çalışmada Coimbra veri kümesi toplanmış ve 10 prediktöre göre çalışılmıştır. Bu prediktörleri meme kanserinin olup olmadığını tahmin etmek için kullandık. Kullanılan 6 algoritma WEKA ve MATLAB'da performanslarına göre karşılaştırıldı. Karşılaştırma, tıbbi karar motoruna iyi hassasiyetle yardımcı olmak için veri madenciliği algoritmalarını kullanma olasılığını kanıtlamak için kullanışlıdır.

**Anahtar Kelimeler:** Meme kanseri; Veri madenciliği; Sınıflandırma modelleri.

# TABLE OF CONTENTS

## Pages

List of Figures.....	V
List of Tables .....	VI
List of Abbreviations.....	VII
<b>1 INTRODUCTION .....</b>	<b>1</b>
1.1 Overview .....	1
1.2 Problem formulation .....	2
1.3 Aim of the work.....	2
1.4 Previous Work.....	3
1.5 Proposed Work.....	3
1.6 Thesis organization.....	4
<b>2 BACKGROUND STUDY .....</b>	<b>5</b>
2.1 Roles in data mining classification.....	5
2.2 Model construction .....	6
2.3 Classification methods .....	7
2.4 Learning in classification rules.....	8
2.5 Regression tree and classification .....	9
2.6 Naïve k-mean algorithm.....	9
<b>3 LITERATURE SURVEY .....</b>	<b>10</b>
<b>4 METHODOLOGY .....</b>	<b>18</b>
4.1 Principal component analysis (PCA) .....	18
4.2 Performance metrics.....	19



4.2.1	performance measures .....	19
4.2.2	Validation measure .....	21
4.3	Simulation tools.....	22
5	<b>RESULTS .....</b>	<b>23</b>
5.1	Dataset characteristics .....	23
5.2	Attribute Information.....	24
5.3	The Main Interface .....	25
5.4	Data 3D virtualization.....	26
5.5	Performance of classification .....	27
5.6	Analysis results .....	29
5.6.1	Class-wise Performance of Benign Cases .....	29
5.6.2	Class-wise Performance of Malignant Cases.....	30
5.7	Classification of Cancer Data using Weka Tool .....	32
5.8	Classification performacne using Weka .....	36
5.8.1	Class-wise Performance of Benign Cases in Weka .....	36
5.8.2	Class-wise Performance of Malignant Cases in Weka .....	39
5.8.3	Average Performance of Classification with Both Cases of Weka.....	41
6	<b>CONCLUSION.....</b>	<b>.....</b>
7	<b>APPENDIX(THE MATLAB CODE) .....</b>	<b>74</b>
8	<b>References .....</b>	<b>74</b>

## **LIST OF FIGURES**

Figure 4.1: The outcomes of the principal component analysis algorithm.....	19
Figure 4.2: The matrix of confusion in performance assessment of classifiers.....	20
Figure 4.3: The main frontend of paradigm as in Matlab platform. ....	22
Figure 5.1: The graphical user interface of the paradigm.....	25
Figure 5.2: The 3D virtualization of project's data. ....	26
Figure 5.3: The class performance of Benign cases. ....	30
Figure 5.4: Average Performance of Classification with Both Cases. <b>Hata! Yer işareti tanımlanmamış.</b>	
Figure 5.5: Graphical representation of average performance from both cases. ....	32
Figure 5.6: Weka main user interface. ....	33
Figure 5.7: The explore of Weka main window. ....	34
Figure 5.8: The database recalling process into Weka. ....	35
Figure 5.9: The analytical calculations of the database as it loads into Weka.....	36
Figure 5.10: Accuracy of classifying of Benign cases in terms of TP and FP. ....	38
Figure 5.11: Accuracy of classifying of Benign cases in terms of Precision and Re call. ....	38
Figure 5.12: Accuracy of classifying of Benign cases in terms of F measure and ROC area.....	39
Figure 5.13: Class-wise Performance of Malignant Cases – Weka Evaluation. ....	39
Figure 5.14: Accuracy of classifying of Malignant cases in terms of TP and FP.....	40
Figure 5.15: Accuracy of classifying of Malignant cases in terms of Precision and Re call.	40

Figure 5.16: Accuracy of classifying of Malignant cases in terms of F measure and ROC area.  
..... 41

Figure 5.17: Accuracy of classifying of both cases in terms of TP and FP in Weka. .... 42

Figure 5.18: Accuracy of classifying of both cases in terms of Precision and Re call in Weka.  
..... 43

Figure 5.19: Accuracy of classifying of both cases in terms of F measure and ROC area in Weka. .... 43



## LIST OF TABLES

Table 5.1: Performance NaiveBayes method.....	27
Table 5.2: Performance of J48 Based Method.....	27
Table 5.3: Performance of RBF-NN Based Method.....	28
Table 5.4: Comparison of the produced outcomes. ....	29
Table 5.5: The performance metric of Malignant case. ....	30
Table 5.6: The average performance of both cases information.....	31
Table 5.7: Class-wise Performance of Benign Cases in Weka Evaluation.....	37
Table 5.8: Class-wise Performance of Both Cases – Weka Evaluation. ....	41

## LIST OF ABBREVIATIONS

ANN	Artificial neural networks
PCA	Principal Component's analysis
LTE	Long term evaluation
DM	Data mining
PPV	Positive prediction value
ESDM mining	Efficiency scalability data
PM	Performance metrics
VM	Validation metrics
AUC	Area under the curve
BC	Breast cancer
RM	Research methodology
GA	Genetic algorithm
PR	Pattern recognition
ROC	Receiver operating characteristic

# 1 INTRODUCTION

## 1.1 OVERVIEW

In large organizations where database is recorded, the process of extracting the useful information from the big database is posing great advantage. The process of knowledge extraction is termed as KDD: Knowledge Discovery in Database. The databases are usually maintained for year of information recording where bulky data is generated, this data includes large information in different time lines where standalone procedure must be started for extraction of knowledge from this bulky data. This process usually involves the following:

- a. Understand the purposes of this database and which field it is about, this is termed as business understanding section.
- b. General data understanding: by referring the objectives of data mining, the data attributes can be studied and data structures can be analysed with accuracy.
- c. Data preparation: where databases can be arranged to be transformable into proper formatting that suites data analyser. The null cells in databases and other unnecessary associations might get omitted in this process, since that, the data preparation is usually the biggest time consuming action.
- d. Data exploration: the procedure to study the data after filtering and preparation process in order to form the hypothesis and plan the algorithms of mining.
- e. Mining of data: establishment of algorithms and methods to extract the hidden patterns in the database. The methods and roles are applied to the database and they are expected to perform patterns extraction which can be judged to obtain mining accuracy.
- f. Results judgments: this is termed to the evaluation of data mining outcomes in order to produce the level of matching with the practical situations.

- g. Results monitoring: for deployment of the obtained results in the practical fields, results should be monitored to find out the malfunctions like errors. This will be very important to interpret the results later.
- h. Project documentation: by inserting and filing all steps and procedures performed at all stages of the project and making it ready for future development. This process is necessary in all data mining project for efficient Knowledge Discovery in Database.

## **1.2 PROBLEM FORMULATION**

Classification is popular terminology deployed in all data mining projects, actually it is an algorithm to perform essential tasks in data mining projects. Many algorithms are associated with classification in data mining the supervised learning algorithms are draws extra attention in this field. The noticeable impact of data mining researches lies on their ability for drawing the same performance on data variation, as data base content increased; data mining algorithms must stand for tolerating this variation. This concept is known as data mining scalability where data base volume is increasing with time which is opposed to the algorithm regime, scalability is key feature of data mining technologies. The cancer data can form two-dimensional database as it has too many features that is recorded for long time in cancer investigation process which is the only reference for forming the cancer database. Due to their large volume. Clustering and data classification are essential point for data mining due to heavy volume of data.

## **1.3 AIM OF THE WORK**

In the work I aim to see the good biomarkers of cancer, It was clear that Glucose, Insulin, HOMA, Leptin, Age, Resistin and BMI are used to indicate the occurrence of breast cancer. I have used algorithms in two tools. Each algorithm will perform in a way to indicate the best biomarker. According to the figures and tables displayed later we can help the doctors to reveal

the illness in early stages. My goal was to check the performance of each algorithm and to detect Cancer in the most efficient ways.

#### **1.4 PREVIOUS WORK**

Many researches have been detecting similar areas and collecting data to detect breast cancer. In 2008 serum levels of the tissue polypeptide specific antigen, specific cancer antigen and insulin like growth factor were introduced as indicators, [2] In 2013 BMI, CA15-3, Leptin and the ratio between Leptin and Adiponectin were used as biomarkers for breast cancer.

In 2015 [3] Serum, irisin levels were found to indicate breast cancer with 62.7% sensitivity and 91.1% specificity. [4] Dalamaga et al reported serum resistin as a predictor of postmenopausal breast cancer and found an AUC value of 0.72, 95% CI [0.64, 0.79]. In 2015, a similar analysis was performed for Leptin, Resistin and Visfatin,

#### **1.5 PROPOSED WORK**

Classes in here are maintained by using a supervised classification technique which trains itself to form the classes. Human must guide the said algorithm for performing the classification tasks such as categorizing the data in to classes, for example when patients are undergone some treatments, their responses for this treatment must be recorder and classified as enhanced groups with respect to time. In following, we have listed some technique that is usually exploited for classification purpose, in further actions we are going to examine selected techniques of classification and validate their scalability by conducting a detailed survey. The classification techniques can be listed herein:

1. Decision tree
2. Neural network
3. Bayesian network
4. K-Nearest Neighbour algorithm
5. DNF rules
6. Genetic algorithm



## 7. Fuzzy and Rough sets

For all the above method, the selection of proper method can only be decided after understanding the actual requirements of the application. The classes density and database volume are the main measures for selection of proper classification method. Furthermore, classification algorithm need to be customized for meeting the requirement of particular volume and classes of the project.

In this project, we are going to establish a software for classification of cancer data by using MATLAB running in environments of windows operation system. Weka function will be ordered in MATLAB routine for classification methods implementations. The performance of classification methods will be analyzed extensively using Weka function in order to obtain efficient paradigm for cancer data clustering. Using of multi-dimensional databases of cancer, we can test the exact performance of the classification system. Multiple tests will be used on the paradigm for deriving the level of performance on practical situation.

### 1.6 THESIS ORGANIZATION

- Chapter one: introduction, the importance of data mining, background and motivation, thesis aim, research questions, The problem statements, Thesis Structure.
- Chapter two: Background study.
- Chapter three is about the research and Literature review
- Chapter four: Practical model and coding work, the methodology and performance metrics.
- Chapter five: Results and discussion.

Conclusion and future work are listed at the end of the thesis with the references and all-important data are tabled in the appendix

## 2 BACKGROUND STUDY

### 2.1 ROLES IN DATA MINING CLASSIFICATION

Two goals are realized from the tasks of classification algorithms which are either aimed to predict a result or to generate a descriptive approach on the collected data. The approach of knowledge discovery in database is involving several keys (steps) in which may be mixing for achieving particular requirements. The knowledge discovery in database usually takes place with the following steps:

- i. Descriptive approach: is set for providing a summary about database description, It is consisting of the following particles:
  1. Summarization of data: by using methods alike Association of Rules algorithm, data summarization can be obtained to provide the description about the database cells.
  2. Clustering or segmentation: this process is found for segregating the database contents into clusters (groups) that involve the similar nature contents. This task is achievable by using available clustering algorithms and closeting methods.
  3. Deviation and changing capturing: the changing in consecutive data behaviours can be detected in this step of processing.
  4. Modelling of dependency: the detection and categorizing the database structure causality.
- ii. Prediction approach: by using the present existent data fields to decide what is the other contents which is nor present, this procedure includes two point such as:
  1. Classification: that used to predict of the outline of categorized groups state of structure.
  2. Regression: for data with variables numerical continuity.

## 2.2 MODEL CONSTRUCTION

The structure of classification model is consisting of the following steps:

1. For the predefined classes, each class is supposed to have a known sample;
2. By using the attributes labelling, the class of known sample is determined;
3. The construction of the classification model infrastructure by training the set of class samples.
4. Each classification model can be represented by the correspondence rules of classification and the tree of decision with mathematical representation.

In data classification, set of methods are shown big interest in similar approaches. As databases are made through gathering enormous volume of information, the task of data clustering and mining the knowledge from the said data is derived through difficulties due to their inconstant volume. Three main procedures are taking place in the knowledge discovery from this big data: the first stage is about data classification which considered as corner stone of data mining project, it used to generate a unified groups that follows particular roles and used as a guidance to generate the upcoming classes (groups), as an example to describe the classification is disease classification where data can be subdivided into smaller groups where different groups of symptoms are forming different diseases. Most traditionally known techniques for classification can be applied to achieve the sub-groups more likely, the neural networks used for learning the classification rules and predict the upcoming groups

This is followed by rules association and eventually it showed go through analysis of the sequence. With presence of large number of algorithms and methods that performs classification tasks, the competition among these methods is made base on their drawn performance in algorithm run time, execution time, latency (time to wait before algorithm generates the results). In the henceforth, the methods of classification such as decision tree, neural network and extra, will be discussed in details.

## 2.3 CLASSIFICATION METHODS

Many methods are existed for performing the classification in data, statistical analysis method is one of those techniques that employed for detection the patterns of the data burst of data by applying of statistical calculations alike linear methods. The terms SPSS and SAS are stands for statistical classification techniques varieties. Artificial neural networks are used to obtain the hidden patterns associated with the data, it uses the same mechanism of human brain of mapping the data between its layers; neural networks proved noticeable performance in most of applications where artificial intelligence is required. Another concept of optimization which is used for optimizing the outcomes based on the nature behaviors, this is called as genetic algorithm where the process of combination and mutation is obtained the better performance from the outcomes. The nearest neighbor classification is another method to obtain the classes from the big data, it creates a set of classes rules and searches the similar (matchings) from all data and arrange the same inside the sub-class. The concept of rules segregation of the useful rules in data is termed as rules indication which is considered as another approach of data classification. Ultimately, classification can be made on the bases of visualization of data sets and on this biases the classes can be made.

The data abstraction is a step taken by several classification algorithms prior to performing any classification work. In order to do that, several approaches are coming into image where the data is generalized and classes are briefly defined. Three variations can be achieved based on the level of data abstraction.

- Minimum level of abstraction
- Intermediate level of abstraction
- High level of abstraction

Level of abstraction is revealing the required definition about the data set in each cell where the classification rules and process can depend on this abstraction to perform the further procedures. The low level of abstraction may degrade the classification process as it provides a very few information about the said data whereas the high level of abstraction may cause a serious malfunction in the classification by increasing the processor load due to computational complexity and high calculation power requirements.

## 2.4 LEARNING IN CLASSIFICATION RULES

The learning of classification rules is all about searching those rules that groups of data followed for forming that group. The practical situation more likely, producing the rules of classification is a tough job due to their large volume and diversity that forms the decision trees. The difficulties of finding the said rules of classification is existing due to the complex computations of the same. Algorithms such as hunt method are usually used for producing the said rules of classification for a given situation, the mechanism of the hunt algorithm can be explained by the following example:

For a known cases of training set,  $R$  where  $R = \{K_1, K_2, K_3, K_k\}$ ; and  $K$  is the training case. The decision tree formulation may take place as the following:

First assumption: if the tree set involves single case of more and all related to the same class identified by  $K_i$ , then the decision tree may be generated with single leaf that stands for the class  $K_i$ .

Second assumption: the decision tree is set with single leaf for no class condition, in this case the information tabulated in the leaf may only stands for the other decision tree class.

Third assumption: when multiple cases are taking place and related to mixing classes, in this condition, first procedure to be taken is establishment of test and this test may yield an outcome such as  $\{o_1, o_2, o_3, o_n\}$  in this case the decision tree might be segregated into several subsets. This tree is identifying the tests by having of multiple nodes that represent each test individually.

The decision tree is used to identify the classes for a given data by partitioning the rules of different nature and forming a sun sets that builds the classes. For particular volume of data, classification process can be performed in two stages:

1. Some part of the said data are taken for training purpose;
2. The remaining portion of the data is used for test and validation purpose.

The test or validation is the only measure of the classification accuracy of the data which is revealing the percentage of classification success. Two factors can be raised due to level of abstraction where the high level of abstraction is causing disorder to the accuracy of classification whereas the lower level of abstraction may lead to classes scattering and difficulty of the interpretation of concise semantic. Taking about the implementation of decision tree, the process of implementation is depending on some attributes associated with the data. The splitting of attributes is one approach proposed to form the decision tree which depends on the relationship between the consecutive values in the data.

## **2.5 REGRESSION TREE AND CLASSIFICATION**

One of the best means to formulate the tree of decisions, the classification and decision tree (CADT) method uses a binary representation for segregation the data in every node of the tree base on the functions of different attributes. The best splitting determination can be taken by using the term gini index, this procedure is used to generate the nodes by splitting of the data. It happens firstly by generating of two node that act as root nodes in which generate the upcoming nodes. Leaf node can be labeled as so if no information could be obtained for splitting the data.

Ultimately, at the stage of completion of tree grown, the nodes that labeled as leaf nodes are remaining. The leaf of decision tree is used for training set assignments at the full tree of decisions. Class with its error rate can be gained using of the remaining leaf, the rate of error is representation of the classification fails in this particular leaf (node). The error taken from all leaves are summed and weighted which is producing the total error of the network.

## **2.6 NAÏVE K-MEAN ALGORITHM**

The local optimal solution can be used to obtain a solution for k-means problem locally by applying the simple iterative approach. The algorithm of such agenda is known as K-means algorithm which is known by its large variants so we will need to be more specific while describing of this algorithm, more likely, we are going to discuss about the Naïve k-means algorithm in the further sections. The database is partitioned into several sets called as k-sets by applying the algorithm of k-mean naïve algorithm this sub sets is usually called as points that related to the same central set, those points are approaching to the center more than

approaching to any other set. Portioning of space is happening randomly at the initial part of the procedure where the centroid is taking randomly to some points in the working space. A number of iterations are made in form of set called iterations set where in each iteration, the process of generating a fresh centroid is taking place. The steps of each iteration is performing according to the assumption of the iteration is nominated as (i) and the centroid in each iteration is nominated as  $CO(i)$ , however, the following procedure is needed to be followed:

1. Obtaining the database location where each points from the database is assigned to that location more likely, finding what so called the centroid of the points in the database. The Euclidean distance is used for portioning the points and set the centroid.
2. The location of the obtained centroid is to be updated and relocate to the point  $CO(i+1)$ , the same point nominated as  $CO(i+1)$  is produced by taking the average value of the given centroids in particular dataset. The same is known as the new location of old centroid.

Now, if the new location of partition that is considered as an updated location can determine the level of algorithm convergence. However, if the value of new location dose not much effecting the value of old location, then the same is making the algorithm known as converged algorithm. In other word, when the algorithm is said converged, that means the centroid of old points location  $CO(i)$  and the centroid of new (updated) location  $CO(i+1)$  are both identical. The K-means algorithm is considered as one of attractive clustering algorithms spatially when the property of convergence is applicable on the same. This algorithm is established in order to eliminate the points distortion in the cluster by obtaining the mean location from all the locations and then the probability of error (distortion) will be minimized.

### **3 LITERATURE SURVEY**

At [1], a data about particular observations are formed a multi-dimensional database, the same is abstracted with data outline, in this study, the author have discussed the possibility of classification process on this database. Assuming the existence of reasonable observation in this data base, the randomness of the observations is caused a drawback while clustering the data, in order to tackle this difficulty a rudest decision rule algorithm is proposed on the bases

of truncation principle; however, the performance of the classification is enhanced to be double than it was before.

At [2], the performance of clustering can be measured by using the linear algebra, it is usually made a good metric of performance of the clustering of text. The quality of clustering was described in this article. The linear algebra method measures the performance on clustering is underlie by the fact the more non-similar objects (points) under the same cluster is directly lead to performance disorder, in other word, the differences in the particular cluster points is proportional to the quality of clustering. Measuring the clustering overlapping across the entire clusters is counted by summation the individual metric of each cluster in the project. In some cases where random clustering is taking over the project, the metric of clustering quality is difficult to be obtained, due to that, some statistical calculations are proved a noticeable earthiness in this case, the standard deviation and mean square values are the most usable approaches in this condition. Such concept of clustering performance is known as standardized clustering metric.

At [3], the authors of this study proposed a new approach of clustering that is called a hierarchical clustering. The hierarchical clustering is proposed by forming a function called objective function that is directly dependent of Bayesian analysis. This model can be outlined as portioning the data into several clusters where each cluster forming a hierarchical clustered data. The outcomes of this study are resultant of evenly distributed features (attributes) amongst the all (most) clusters and then the level of scattering is reduced. Each sub-cluster in this complex is represented a point (node) that form the entire hierarchical structure, the features are uniformly distributed among those nodes.

At [4], the writers of this study were in process that is highlighting the importance of artificial neural network on mapping and clustering approaches. They have planned using the concept of artificial neural network as mapping technology for data science that exploit their ability of self-learning and organizing. A kohonen self-organized mapping approach (CSOM) is proposed by those authors as mapper in clustering of multi extension maps. This article involves presenting of kohonen self organized mapping (CSOM) followed by reviewing the internal technologies that deal with clustering project more likely detailing the way of



portioning of cluster and ways of naming the cluster and ultimately the means of generalization (abstraction) of the clusters.

At [5], the clusters in database of multidimensional are detectable by means of virtual framework that interactively used in this study. The most relevant problem in the field of data mining is about finding the closely related points in the databases of multidimensional structure in which forming the groups (clusters), however, performing such task is not completely easy as said by the authors. In this study, detecting the said clusters are more likely done by using virtualization technology where the database is viewed as animated objects that are adjusted using the operations of zooming, and selection and probing of clusters, however, the study resulted a smooth way for detecting the clusters by depending on the display or the outlooks of the clusters where the user (admin) can virtually selecting it.

6:39 At [6], another clustering technique is presented in this study where the clusters are recognized by using the method or random walls. The authors relay on the computational tasks and power of computations for detecting a particular pattern in the database of huge contains. The computational analysis is standing for the clustering analysis technologies, in presence of many techniques of clustering, the main drawback of the traditional clustering analyzers is their weak accuracy when special and big databases are used. For the above introductory, the presenting of classical cluster analyzing is found to be none worthy and then this study has proposed to find what so called a deterministic analyzer for analyzing the spatial data. The advantages of this technology are about grouping the data into several clusters that involves the similarly attributed points and then some properties can be obtained from the known shaping generated groups. From that noise in each cluster can be overcome and data outliers which discard the decomposition nature of the data, one more advantage of this method, is low time requirement and smaller space fitness.

At [7], a new clustering technique is proposed in this study which is called as fuzzy relational clustering of low complex nature. This technique is using the methods of Robust Fuzzy C-Medoids and another method called Fuzzy C-Medoids for performing the relational clustering of fuzzy. This method is basically work on the relational argument in the clusters where it is forming a c shape of the data and call it as C-Medoid cluster, these clusters are then prepared for the quality preferences selections more like it is done by fining the level of similarity in

each cluster, in that the minimum level of dissimilarity is always preferred. The use of both above methods of fuzzy c-Medoids are shown that the second method (Fuzzy C-Medoid) is drawn a noticeable performance as compared to the Relational Fuzzy C Medoids algorithm. This algorithm is deployed as clustering technique in most of the web-based applications such as clustering of the access data and clustering of the web documents.

At [8], the authors of this articles have proposed using the gene of the database for formulation of decision tree using the supervised learning that harvesting the expressions of genes in the data. The method is beginning for processing the data by building up clusters in hierarchy outline and then obtaining the variance in each hieratical cluster, the overall variance is found by average clustering variances of all clusters. The measure of performance can be obtained in this study by applying the same metrics into several classes aiming of finding the measures such as time of survival and metrics related to more than one class. The proposed method in this article is suggesting a way to obtain the genes that have strong impact on their class (cluster) and the other genes that impacting (associated) with other clusters (classes). The database used in this study was about the censer and lymphoma diseases which contain very large records from the both diseases data, the eight kind of cancer records are illustrated by clustering the database in this study.

At [9], the microarray genes can be clustered using of Rasch model clustering, the same is proposed in this study. The Rasch method of clustering is established with two kinds of statistical formulation that look after the problem of genes profile expression. In those two formulations, the first formulation is set to determine whether the observed phenotype is related to the same detected cluster genes expression profile, the prediction of future can be also made using this formulation model. The other model of formulation is used for calculating the status of recognition to the genes that under express or over express level in a sample of tissue cell for known type of cells tissue. This paradigm is resulting the clustering outcomes of leukemia's cancer in sixty type (variants) of this cancer. The results of cancer classification are made more efficient comparable with those results of previous experiments, however, the outcomes are also shown that patients are responding to different drugs, ninety drugs were illustrated in the same regard. Results shown the under and over levels of expressions in the genes profiles. The probabilistic model of the Rasch clustering method is yielded the pattern of genes expression in the phenotype status.

At [10], for known database a clustering is proposed based on the prediction of the resampling method that is used for knowing the clusters available in the said database. In the problems of biomedical and health sciences related issues, the microarray is used to obtain the classes (clusters) and address most of the problems in classification such as the issues related to the classification of diseases like cancer. Cancer related features can be identified using the statistical of genes profile expression, this statistic is employed for recognition of many tumor classification approaches. Using the method of profile of genes expression is yielding several advantages such as reducing the total number of classes (clusters) in big database, it also assigns different relevant samples from different clusters to the produced cluster, and it also used for assessment the cluster confidence by measuring the confidence of each sample (point) in this cluster. The problems of large clusters number obtained from the database is tackled by using the resampling method that based on the prediction. In order to validate the performance of the resampling method with prediction foundations, the data of clustering is compared with four different approaches which were published to identify the cancer disease from the database, the result of the comparison have shown extended accuracy of the clustering done using the prediction based resampling method, more likely, the results shows good performance at the time this approach was published.

10:15 At [11], for spatial databases concerning about the biomedical problems such as cancer related database which is describing the cancer variants with large data records, however, the database can be clustered into several clusters using the method called spatial access method (SAM). The use of spatial access method is yielding two cluster information in both data and their indexes. The authors of this articles are suggested the method of corner transformation and they argued that such method can preserve the property of clustering more likely those object of same location and size, the method of corner transformation reservation can place such objects at the same location and size in the transformable matrix.

At [12], this study is demonstrating the k-means algorithm with enabled kernel, the Mercer kernels are used in arbitrary cluster learning method that using k-means method with kernel enabling for cluster boundary learning. The centroid of these clusters are defined as combination of linear points of cluster in the space of higher dimensions, the formulation of these points are performed by k means algorithm with enabled mercer kernel. The contribution

of this approach lies on their capability to address the clustering process with very high dimensional space where points of cluster can be gathered for different clustering natures.

At [13], the authors of this study are performed a project that deals with market competing, the designed paradigm is looking after the growth of the said competing and virtualizing the competition and tracking it. The authors have integrating an approach to perform the tracking and virtualization of markets paradigm. The said approach is established to show the process of the long term nature that participate the competing in market trades. The field of problem as it said as market, is representing of the competing of scientific publication field where cluster are required to be formed as co-cited and cited clusters of the scientific publications. The resultant paradigm is reflecting the percentage of citation increment in scientific field where the clusters are varying between the cited and none cited groups.

At [14], this study is established to present three approaches for mining of clusters and interpretation of the index to describe the clusters. The proposed paradigm is emphasized on the production of results with sufficient interpretations and possibility of virtualization. The presentation of the data that inward to the paradigm is done as first step of the process, however, the data is mapped in self organized group that make the virtualization easy after removing the outer and noise components and filling the missing values from the database cells. After filtering and preprocessing of the inputted data, the fuzzy clusters are established after the preprocessing stage and based on the similarity between the data points. As any classical classification approach, this study ended with descriptive of the clusters which enables the analyst (admin) to get parabolic with quantitative details about the project.

At [15], this study proposed a repeated measurement to obtain the clustering of genes expression data. The repeated measurements are integrated with several clustering algorithms where the writers of this approach have presented. The aim of presentation of several algorithms that supports the expression measurements is to get popular with the advantages of the same, where the expression measurement-based clustering algorithms is a topic used for enhancing the performance of clustering model and optimizing the clusters stability. The superior outcomes are achievable as described by the authors by using the infinite mixture model as a backbone of the clustering paradigm.

At [16], the unsupervised clusters of fuzzy data that depends on the cluster merging method of similarity driven is proposed in this study for truckles the variations in the clustering. Beginning with over specified count of clusters existed in particular data, the approach that used to merge the similarity cluster pairs on the bases of the driven of similarity merging criterion. The matrix similarity of fuzzy clusters is used to produce the index of similarity between the clusters. The merging of points in particular clusters is done using the adaptive threshold method. The object function that generally modified is employed for paradigm based fuzzy classification. The principle components of clustering, the p-norms distance calculation are included in this method. Whereas the database can be used for obtaining the principle components in automatic fashion. The genes expressions data sets are used to identify the performance of the unsupervised learning clustering algorithm with help of artificial data sets in several experiments.

At [17], the extended fuzzy that form the clustering algorithms are used in this study, the authors are stating in this article that marketing and finance are field that processed in current days successfully by using the fuzzy clustering. In the face of the noticeable success of the fuzzy clustering algorithms is many of fields spatially those of economic interests, the fuzzy logic clustering is so far having leak of performance in none small number of applications. Two objective function extensions are proposed for dealing with the clustering problems of fuzzy clustering. For simplifying the said clustering issues of this method, the first step is to maximize the size of the points in particular cluster to be in hypervolume side, the size of the new point will be produced automatically at the time the point is getting cluster in the particular class. The second approach is optimization related process where the actual effectiveness of the said cluster is obtained and the similar clusters are getting meagered. In the known database, the begging with estimation of big number of clusters, during the clustering, the similar nature of two clusters may lead to merge them into single cluster for enhancing the data partitioning performance.

At [18], the writers of this article were demonstrated an approach to perform the clustering with worthiness by using the genetic algorithm for re-describing of subjects in the clusters descriptions. The information retrieving effectiveness and efficiency can be enhanced using the queries-based retrieval information system as said by the authors. The static description of web document is used in this study to perform the clusters of this document using the similarity

to produce groups of similar clusters. This method is directly depending on clustering from the user perspectives. For the system of queries, the clusters that related to co-relevant documents are drawing increased similarity in the descriptions, hence, the relevant queries matching of this document is become more efficient. The clustering algorithm accuracy can be increased with increasing the similarity index of the clusters of the big data.

At [19], the document of hypermedia is being clustered and discussed in this study using of adaptive clustering algorithm, however, the approach is consisted of two adaptive algorithms more likely the neural network algorithm and the genetic algorithm to be used for clustering the document of hypermedia. By clustering the said hypermedia document, the data is reformed as clustered nodes and information can be obtained with ease from the said nodes where user can study and make beneficial use of this data. This study is concern of abstraction the links that user might be following to reach the node of information, in the same time, authors made no efforts to manipulate the data within those clusters. The user vision that identify the unique relationships among the nodes is highlighted in this study. Ultimately, the document of hypermedia is standing with no change after performing the content of this article where users only can have personalized keys and indexes to the information contained in this document.

## 4 METHODOLOGY

### 4.1 PRINCIPAL COMPONENT ANALYSIS (PCA)

One of the classical methods for performing the clustering in big data is using the principal component analysis PCA, however, this algorithm is working on the visualization bases where the data points in the database are plotted in three-dimensional axis where the analyst can virtually distinguish the data pattern. For the data set of S, the linear data points are listed in this vector and contained as  $i^{\text{th}}$  elements, more likely,  $S(i)$  where  $i=1,2,3,4\dots n$ . the principle components can be recognized from this linear data by using the following formula.

$$f(S, V) = (sV)V^T + u \quad 1$$

In the equation 1:

The function of equation (1) is representing the vector value function  $f(S, V)$

The term “u” is representing the average value of the data points fallen in the dataset S

The term V is the orthogonality matrix that produced from the d by m matrixes

The term sV is called as mapping vector where the value of s is projected in low dimensional domain.

The estimation of V matrix projection can be given in the equation 2, where this represents the principle component analysis function of the above vectors.

$$R(u, V) = \frac{i}{n} \sum_{i=1}^{i \geq n} |s(i) - f(s(i), V)|^2 \quad 2$$

However, the principal component analysis can yield the results as virtual representation as demonstrated in the figure below.

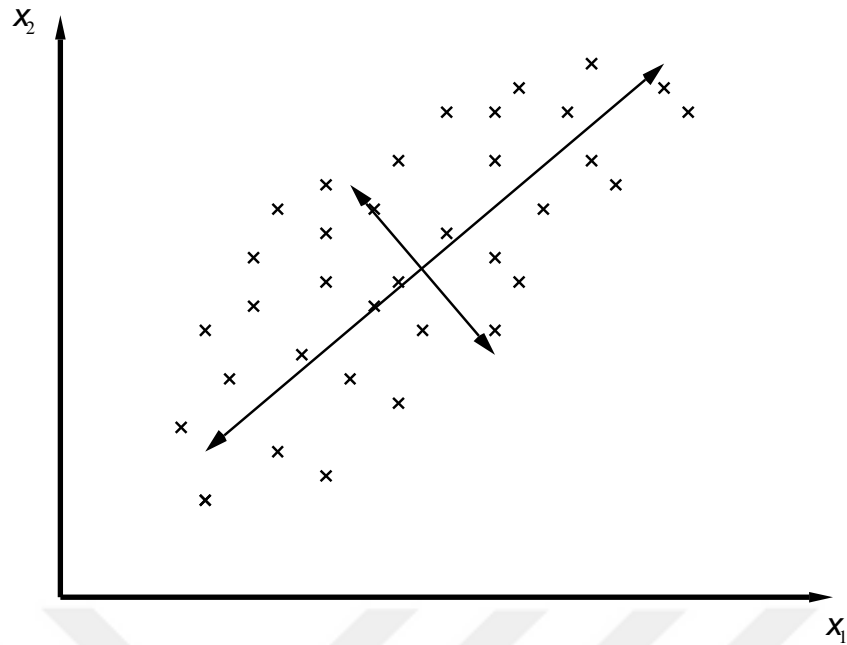


Figure 4.1: The outcomes of the principal component analysis algorithm.

## 4.2 PERFORMANCE METRICS

In order to find the efficiency of the clustering algorithm, generally set of measurements agreements are made to determine the performance metrics. Two variables are provided at each time clustering process is into the image: the dataset which is recorded on the bases of some event in the life and in return, the clusters that is resulted from the clustering are also produced. The characteristics of the data that is being classified are directly affecting the classification performance. In the process of performance assessment, two procedures are made such as the performance measures and validation tests. The performance measure can be done by conduction of the following actions: specificity metrics, sensitivity metrics, precision metrics, accuracy metrics, error rate and F-score metrics. Whereas the validation test can be conducted by performing the following procedures: random sub sampling, bootstrap technique, K-fold validation cross and holdout.

### 4.2.1 performance measures

1. Confusion matrix: the performance of classifier is usually partitioned out to be form as matrix called as confusion matrix. The errors that made by the classifier are usually listed in the confusion matrix that is virtualizing all errors and abstracting of it to be



more readable. In order to get away with the confusion matrix, the figure below is demonstrating the sample content of the matrix.

Hypothesis		Actual Class
+	-	
<b>a</b>	<b>b</b>	+
<b>c</b>	<b>d</b>	-

Figure 4.2: The matrix of confusion in performance assessment of classifiers.

The nomination “a” is termed to the classification of positive values of the data and the process where no errors are made in this kind of classification.

The nomination “b” is termed the classification of negative values in the dataset where the process are made with misclassification events.

The nomination “c” is the negative data that also misclassified and the result of classification is revered as positive.

The nomination “d” is the results of classification of the negative values and found the same negative without making any misclassification or error.

2. The Recall or Sensitivity: when the database is involving a set of positive and negative values, and however, the classification process are taken place with percentage of error, hence some positive values are classified as positive without error and others are classified with error. The sensitivity is the measure to identify the percentage of error in the classification in terms of negatives misclassified as positives in the classification results. The recall is another terminology of the same point (sensitivity) and it can be calculated using the following formula.

$$\text{Sensitivity}=\text{Recall}=\frac{TP}{FN+TP} \quad 3$$

3. Specificity: similarly alike the sensitivity, when the database is involving a set of positive and negative values, and however, the classification process are taken place with percentage of error, hence some positive values are classified as positive without error and others are classified with error. The sensitivity is the measure to identify the percentage of error in the classification in terms of positives misclassified as negatives in the classification results. The recall is another terminology of the same point (sensitivity) and it can be calculated using the following formula.

$$\text{Specificity}=\frac{TN}{FP+TN} \quad 4$$

4. Accuracy: the classification process is usually happening with errors where the positives or negatives values are classified into negatives and positives respectively. The overall percentage of error happening y accumulating all sensitivity and specificity error are called as accuracy, the same can be represented as the following formula.

$$\text{Accuracy}=\frac{TN+TP}{FN+TP+TN+FP} \quad 5$$

5. Positive prediction of the value (precision): it is called in short as PPV and is determined as per the following formula.

$$\text{PPV}=\frac{TP}{FP+TP} \quad 6$$

6. F-score: it also called as F-measure and can be determined by using the following formula.

$$\text{F-score}=\frac{2 * (\text{Recall} * \text{precision})}{\text{Recall} + \text{precision}} \quad 7$$

7. The error rate: which is producible using the following formula.

$$\text{Rate of Error}=\frac{FN+FP}{FN+FP+TP+TN} \quad 8$$

#### 4.2.2 Validation measure

The K-fold cross validation method is used as validation technique in this project where it based on the method of holdout. However, for particular dataset, the K-fold method is firstly segregating the dataset into ks groups where the first group can be nominated as GR(1) and last group can be nominated as GR(k). The K-fold involves using the holdout method on each Kth group and validating the results obtained from this method with the results of entering the GR(n+1) into the holdout method. More likely, the groups are used for testing and validation consecutively.

### 4.3 SIMULATION TOOLS

The programmable objects in some simulators (platforms) are given more flexibility for accommodating the requirements of data science problems. The Matlab software is used to encode all the algorithms and to generate the graphical user interface as in the figure below so that the analyst can easily login to the same and apply the algorithms on any dataset.

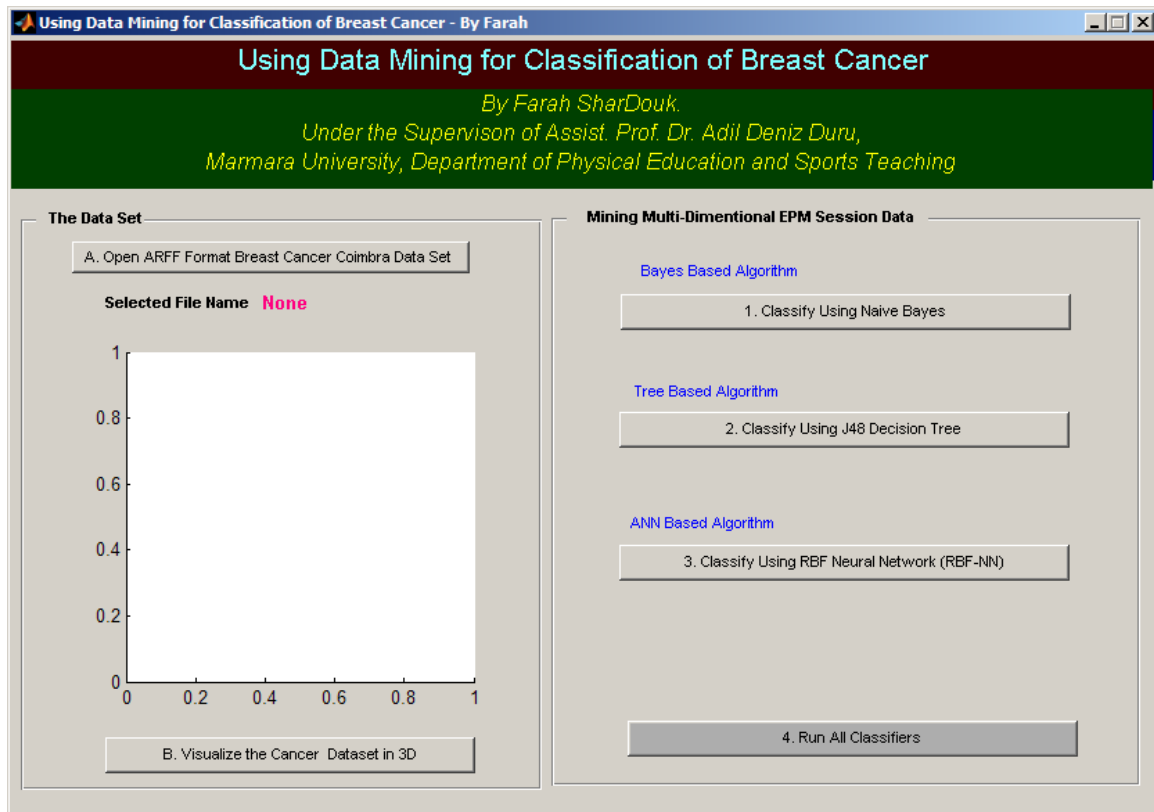


Figure 4.3: The main frontend of paradigm as in Matlab platform.

The figure is declared that data is recalled first from the local machine to the Matlab workspace and first it is plotted using the concept of principle component analysis PCA. Furthermore the algorithms pf our proposal are applied on the said data respectively.

## 5 RESULT AND DISCUSSION

The proposed system has been implemented and tested in Matlab and Weka under Windows Operating System. The Performance of the classification algorithm was tested with the breast cancer database called “Breast Cancer Coimbra Data Set “About Breast Cancer Coimbra Data Set This Data set is originally prepared by the Faculty of Medicine of the University of Coimbra and also by the University Hospital Centre of Coimbra.

### 5.1 DATASET CHARACTERISTICS

Type	Multivariate
Number of Instances	116
Area	Life
Attribute Characteristics	Integer
Number of Attributes	10
Date Donated to UCI Repository	2018-03-06
Missing Values	None

- There are 10 predictors, all quantitative, and a binary dependent variable, indicating the presence or absence of breast cancer.
- The predictors are anthropometric data and parameters which can be gathered in routine blood analysis.
- Prediction models based on these predictors, if accurate, can potentially be used as a biomarker of breast cancer.

## 5.2 DATASET INFORMATION

Quantitative Attributes:

Age (years)

BMI (kg/m<sup>2</sup>)

Glucose (mg/dL)

Insulin ( $\mu$ U/mL)

HOMA

Leptin (ng/mL)

Adiponectin ( $\mu$ g/mL)

Resistin (ng/mL)

MCP-1(pg/dL)

Labels:

1=Healthy controls (Benigne)

2=Patients (Malignant)

### 5.3 THE MAIN INTERFACE

The User Interface Showing the Multidimensional Data Projected in Virtual 2D Space is given in figure below:

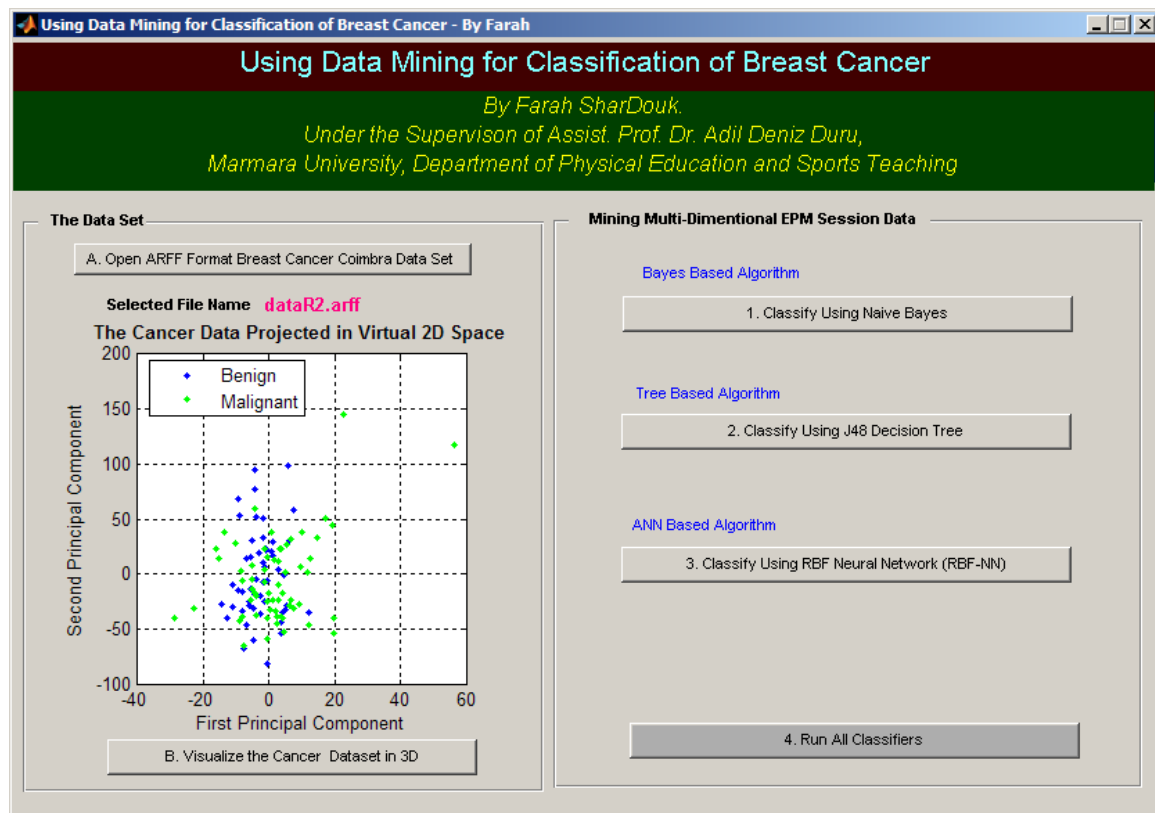


Figure 5.1: The graphical user interface of the paradigm.

Matlab Console output Showing First 10 sample records as the following:

The Selected Cancer Dataset is : dataR2.arff

The Total Number of Records in the Dataset :116

The Total Number of Attributes in the Dataset :10 (including 1 class label)

## 5.4 DATA 3D VIRTUALIZATION

The New Figure Window Showing the Multidimensional Data Projected in Virtual 3D Space is given in figure below:

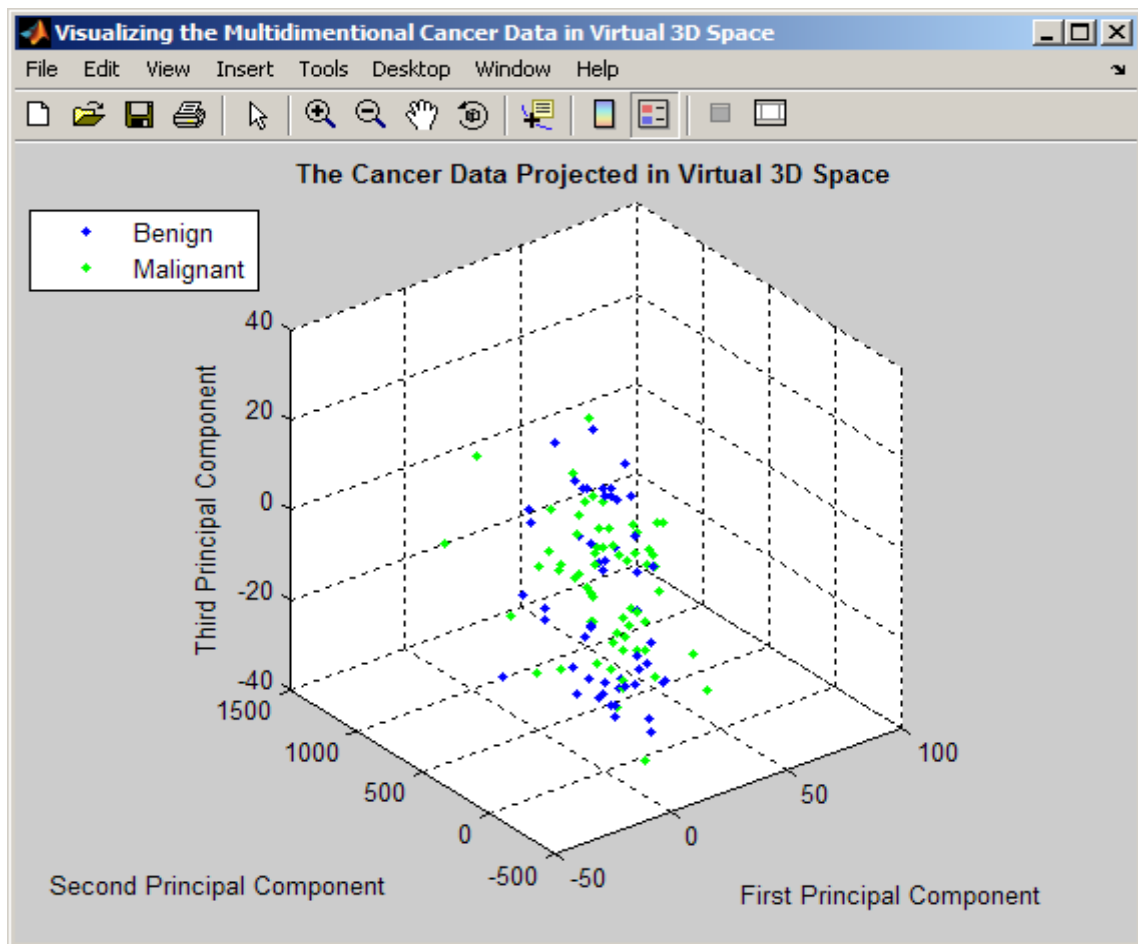


Figure 5.2: The 3D virtualization of project's data.

## 5.5 PERFORMANCE OF CLASSIFICATION

The Console Output Showing the Classification Performance of the three Classification Algorithms as the following table:

Table 5.1: Performance NaiveBayes method.

Performance of Naïve Bayes Based Method							
Time Consumed For Training and Testing 0.07 sec							
	TP rate	FP rate	Precision	Recall	F metric	ROC	Class
	0.846	0.594	0.537	0.846	0.657	0.753	1
	0.406	0.154	0.765	0.406	0.531	0.753	2
Avg.	0.603	0.351	0.662	0.603	0.587	0.735	--

Table 5.2: Performance of J48 Based Method.

Performance of J48 Based Method							
Time Consumed For Training and Testing 0.09 sec							



	TP rate	FP rate	Precision	Recall	F metric	ROC	Class
	0.673	0.297	0.648	0.673	0.66	0.701	1
	0.703	0.327	0.726	0.703	0.714	0.701	2
Avg.	0.69	0.313	0.691	0.69	0.69	0.701	--

Table 5.3: Performance of RBF-NN Based Method.

Performance of RBF-NN Based Method							
Time Consumed For Training and Testing 0.34 sec							
	TP rate	FP rate	Precision	Recall	F metric	ROC	Class
	0.788	0.344	0.651	0.788	0.713	0.77	1
	0.656	0.212	0.792	0.656	0.718	0.77	2

Avg.	0.716	0.271	0.729	0.716	0.716	0.77	--
------	-------	-------	-------	-------	-------	------	----

## 5.6 ANALYSIS RESULTS

The following results are from k-fold validation with k=10.

### 5.6.1 Class-wise Performance of Benign Cases

The following table shows the class-wise performance of Benign cases in terms of different metrics.

Table 5.4: Comparison of the produced outcomes.

Algorithm	TP-Rate	FP-Rate	Precision	Recall	F-Measure	ROC-Area
Naïve Bayes	0.846	0.594	0.537	0.846	0.657	0.735
J48	0.673	0.297	0.648	0.673	0.66	0.701
RBF-NN	0.788	0.344	0.651	0.788	0.713	0.77

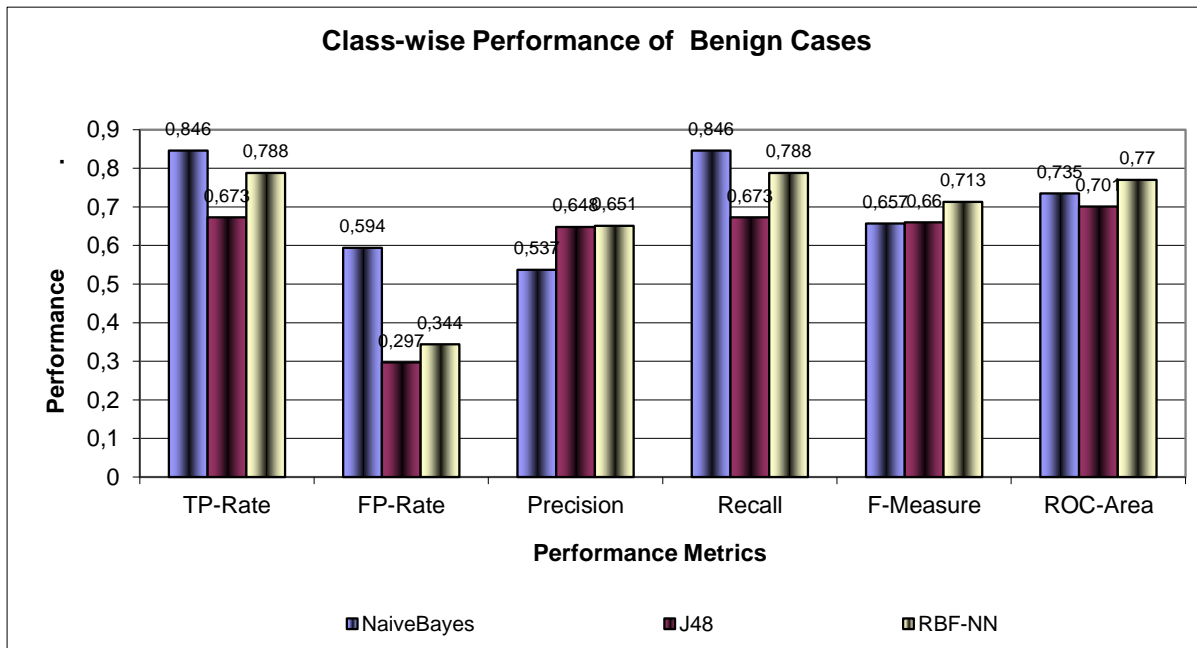


Figure 5.3: The class performance of Benign cases.

### 5.6.2 Class-wise Performance of Malignant Cases

The following table shows the class-wise performance of Malignant cases in terms of different metrics.

Table 5.5: The performance metric of Malignant case.

Algorithm	TP-Rate	FP-Rate	Precision	Recall	F-Measure	ROC-Area
Naïve Bayes	0.406	0.154	0.765	0.406	0.531	0.735
J48	0.703	0.327	0.726	0.703	0.714	0.701

RBF-NN	0.656	0.212	0.792	0.656	0.718	0.77
--------	-------	-------	-------	-------	-------	------

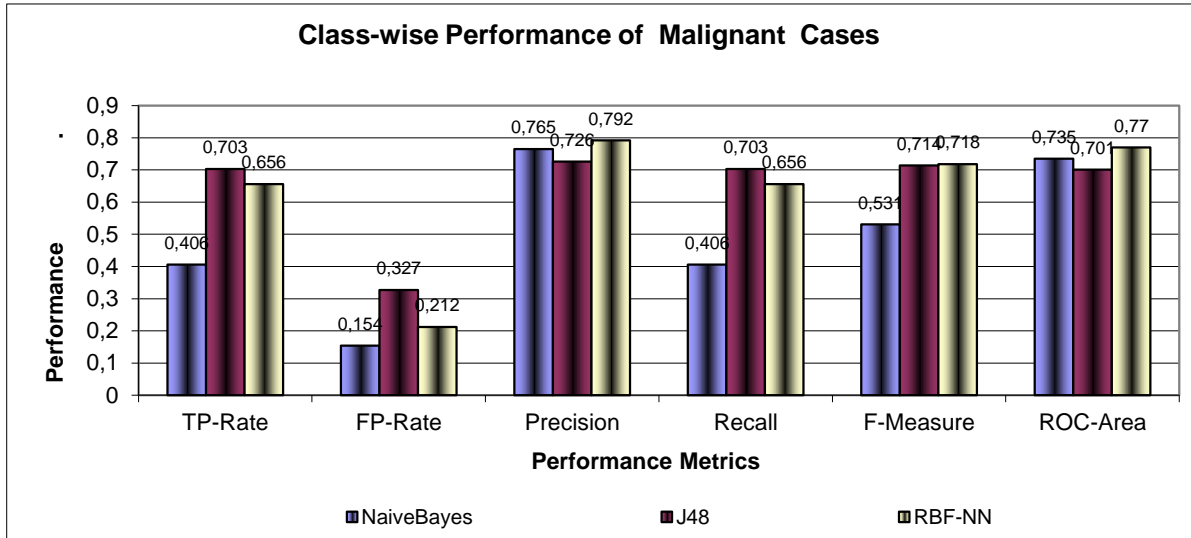


Figure 5.4: Average Performance of Classification with Both Cases

The following table shows the average performance of both cases in terms of different metrics.

Table 5.6: The average performance of both cases information.

Algorithm	TP-Rate	FP-Rate	Precision	Recall	F-Measure	ROC-Area
Naïve Bayes	0.603	0.351	0.662	0.603	0.587	0.735
J48	0.69	0.313	0.691	0.69	0.69	0.701

RBF-NN	0.716	0.271	0.729	0.716	0.716	0.77
--------	-------	-------	-------	-------	-------	------

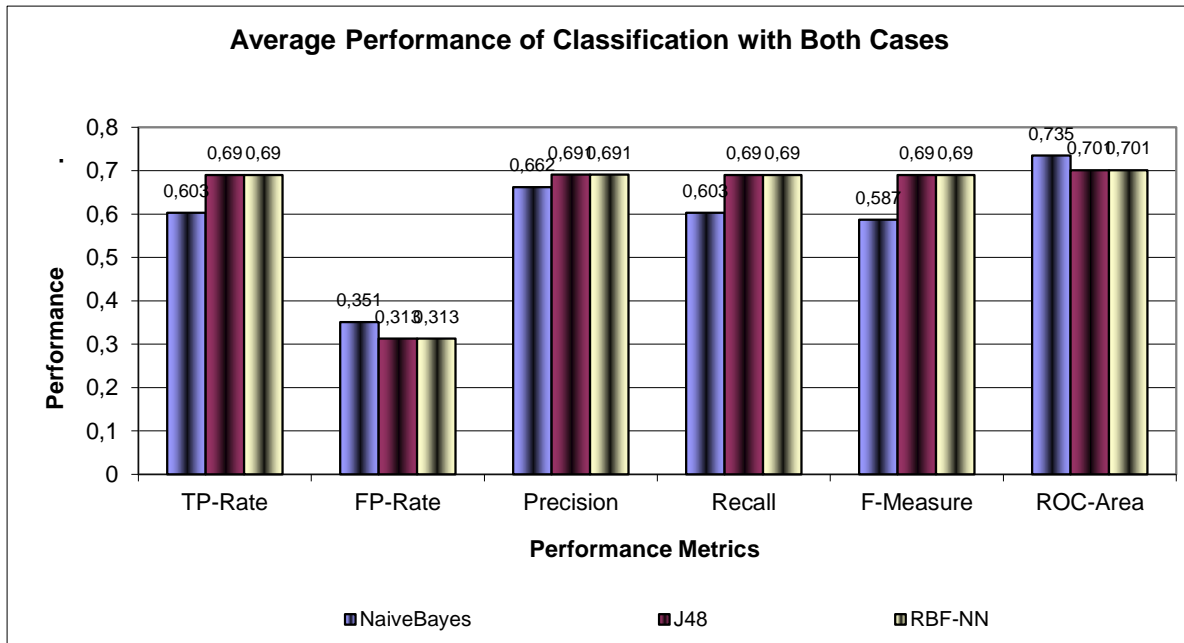


Figure 5.5: Graphical representation of average performance from both cases.

## 5.7 CLASSIFICATION OF CANCER DATA USING WEKA TOOL

Weka is data mining software that uses a collection of machine learning algorithms. These algorithms can be applied directly to the data or called from the Java code. The Weka Explorer is illustrated in below figures and contains a total of six tabs.

- 1) Preprocess: This allows us to choose the data file.
- 2) Classify: This allows us to apply and experiment with different algorithms on preprocessed data files.
- 3) Cluster: This allows us to apply different clustering tools, which identify clusters within the data file.

4) Association: This allows us to apply association rules, which identify the association within the data.

5) Select attributes: These allow us to see the changes on the inclusion and exclusion of attributes from the experiment.

6) Visualize: This allows us to see the possible visualization produced on the data set in a 2D format, in scatter plot and bar graph output.

The user cannot move between the different tabs until the initial preprocessing of the data set has been completed.

Preprocessing: Data preprocessing is a must. There are three ways to inject the data for preprocessing.



Figure 5.6: Weka main user interface.

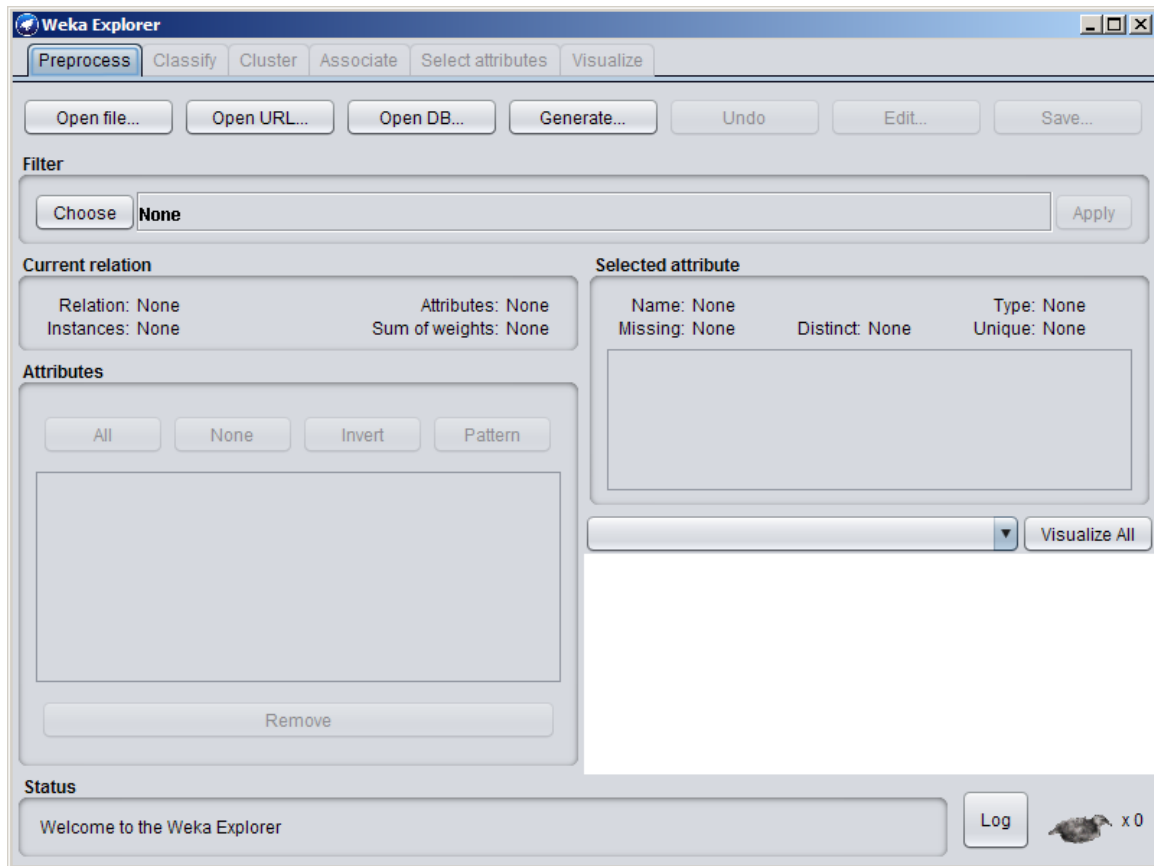


Figure 5.7: The explore of Weka main window.

A screen for selecting a file from the local machine to be preprocessed is shown in the following figure. After loading the data in Explorer, we can refine the data by selecting different options. We can also select or remove the attributes as per our need and even apply filters on data to refine the result. The process of the same is demonstrated at the figure below.

Open File – enables the user to select the file from the local machine

Open URL – enables the user to select the data file from different locations

Open Database – enables users to retrieve a data file from a database source

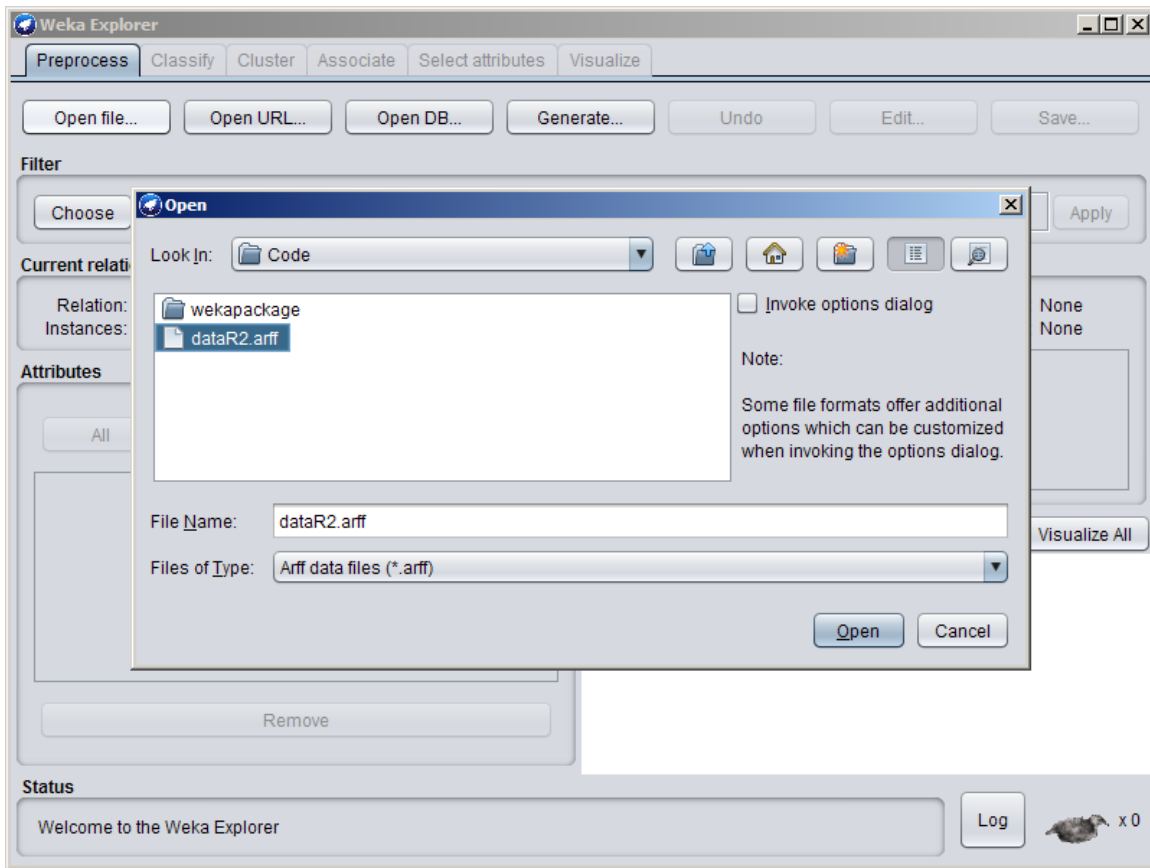


Figure 5.8: The database recalling process into Weka.

As database is loaded into the system, the following statistical will be derived by Weka, as demonstrating in figure below.



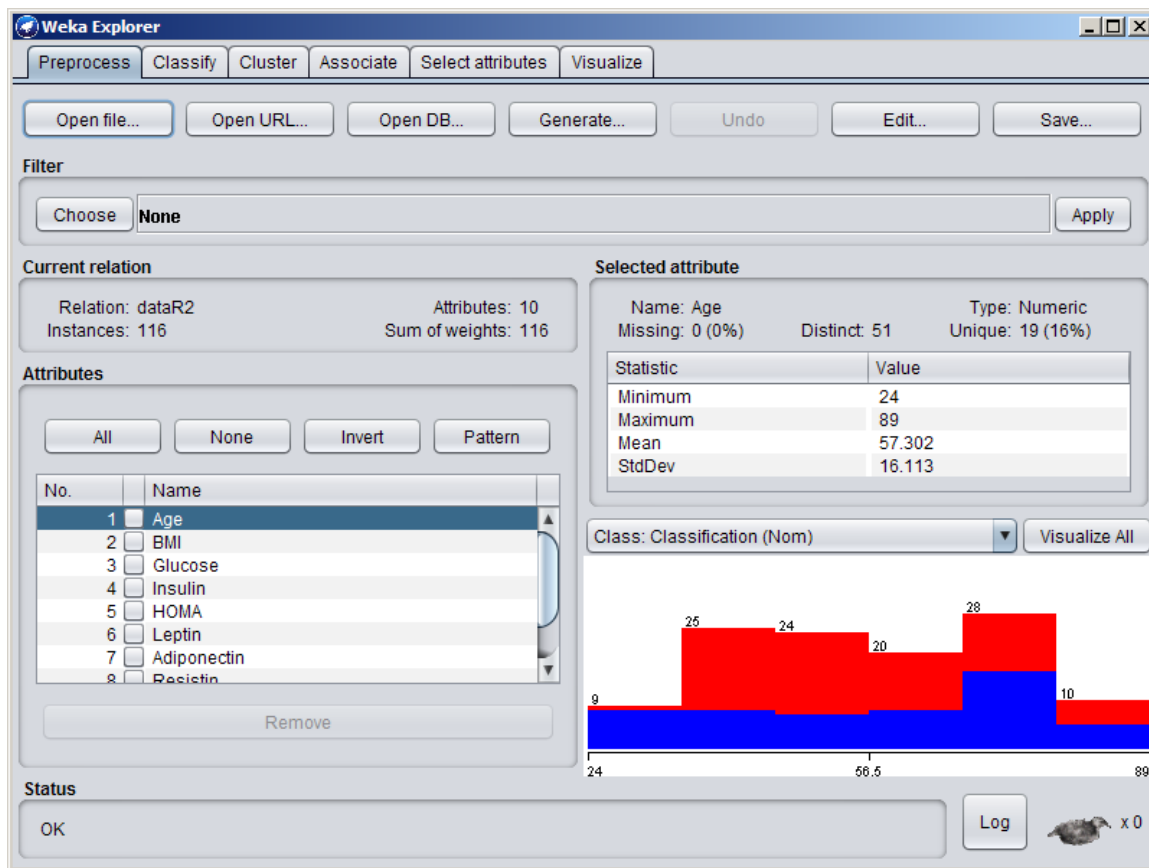


Figure 5.9: The analytical calculations of the database as it loads into Weka.

## 5.8 CLASSIFICATION PERFORMANCE USING WEKA

### 5.8.1 Class-wise Performance of Benign Cases in Weka

The following table shows the class-wise performance of Benign cases in terms of different metrics.

Table 5.7: Class-wise Performance of Benign Cases in Weka Evaluation.

Algorithm	TP-Rate	FP-Rate	Precision	Recall	F-Measure	ROC-Area
AdaBoost M1	0.712	0.219	0.725	0.712	0.718	0.796
Classification Via Regression	0.769	0.250	0.714	0.769	0.741	0.792
Random Forest	0.673	0.203	0.729	0.673	0.700	0.816
Jrip	0.654	0.234	0.694	0.654	0.673	0.715
RBFNN	0.788	0.344	0.651	0.788	0.713	0.770
J48	0.673	0.297	0.648	0.673	0.660	0.701

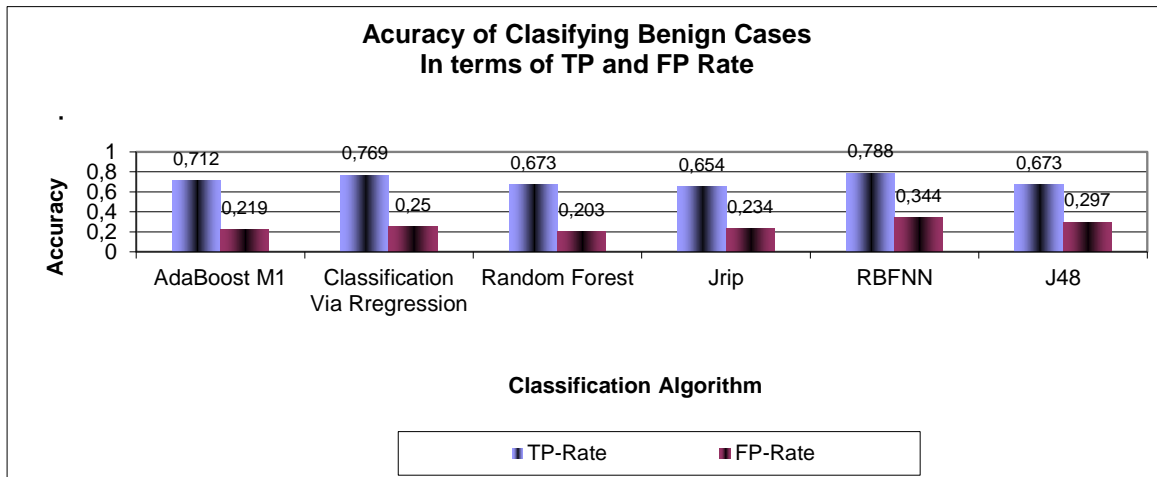


Figure 5.10: Accuracy of classifying of Benign cases in terms of TP and FP.

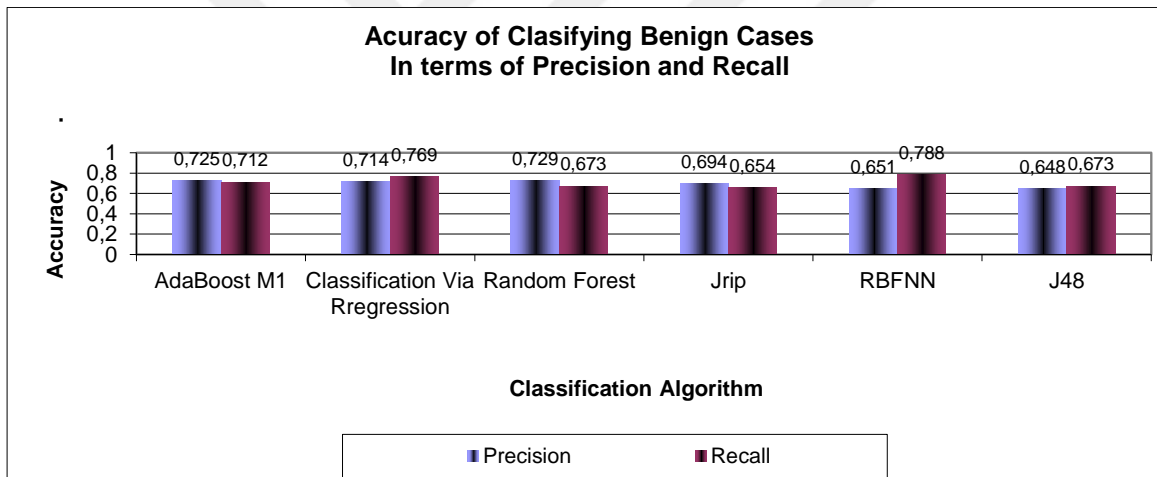


Figure 5.11: Accuracy of classifying of Benign cases in terms of Precision and Re call.

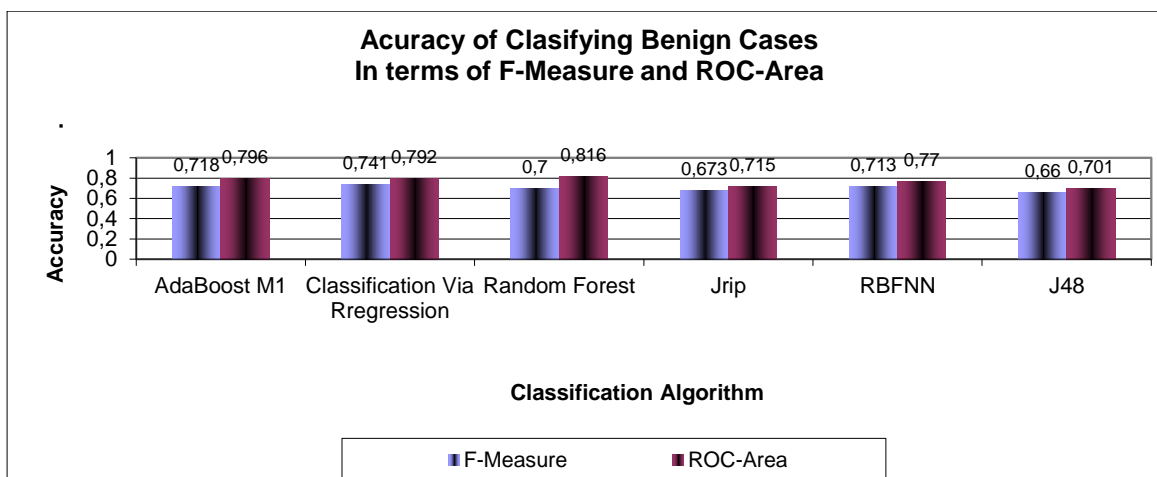


Figure 5.12: Accuracy of classifying of Benign cases in terms of F measure and ROC area.

### 5.8.2 Class-wise Performance of Malignant Cases in Weka

The following table shows the class-wise performance of Malignant cases in terms of different metrics.

Figure 5.13: Class-wise Performance of Malignant Cases – Weka Evaluation.

Algorithm	TP-Rate	FP-Rate	Precision	Recall	F-Measure	ROC-Area
AdaBoost M1	0.781	0.288	0.769	0.781	0.775	0.796
Classification Via Regression	0.75	0.231	0.8	0.75	0.774	0.792
Random Forest	0.797	0.327	0.75	0.797	0.773	0.816
Jrip	0.766	0.346	0.731	0.766	0.748	0.715
RBFNN	0.656	0.212	0.792	0.656	0.718	0.77
J48	0.703	0.327	0.726	0.703	0.714	0.701

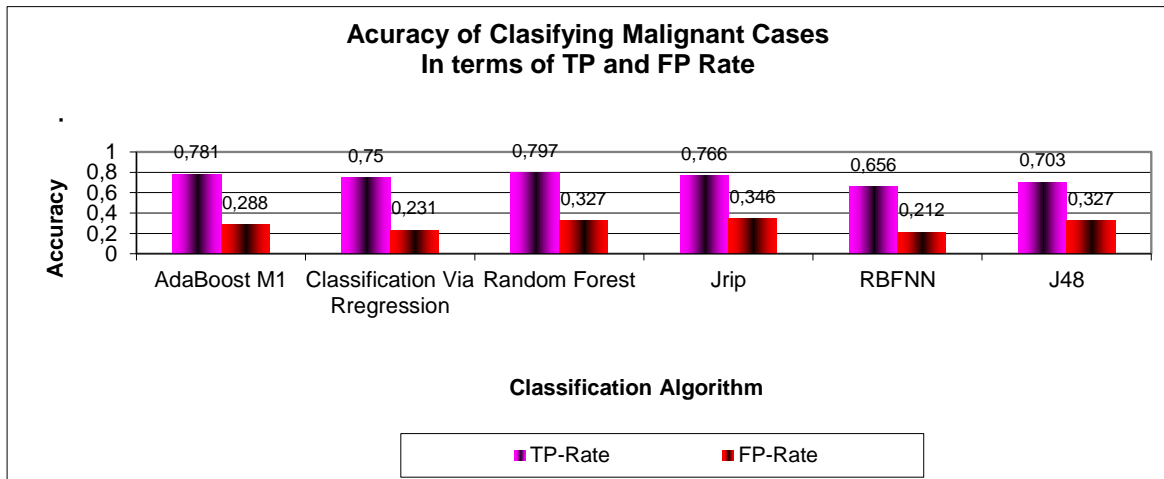


Figure 5.14: Accuracy of classifying of Malignant cases in terms of TP and FP.

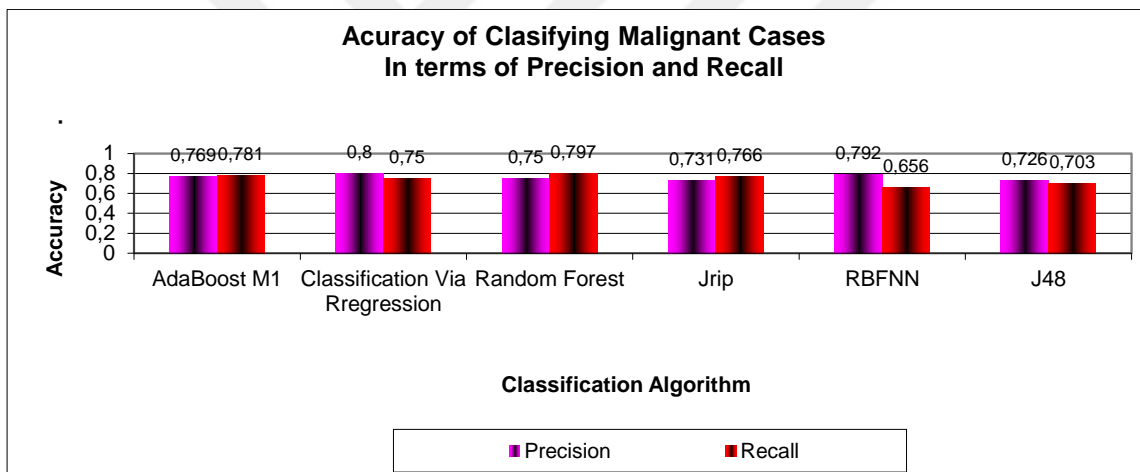


Figure 5.15: Accuracy of classifying of Malignant cases in terms of Precision and Re call.

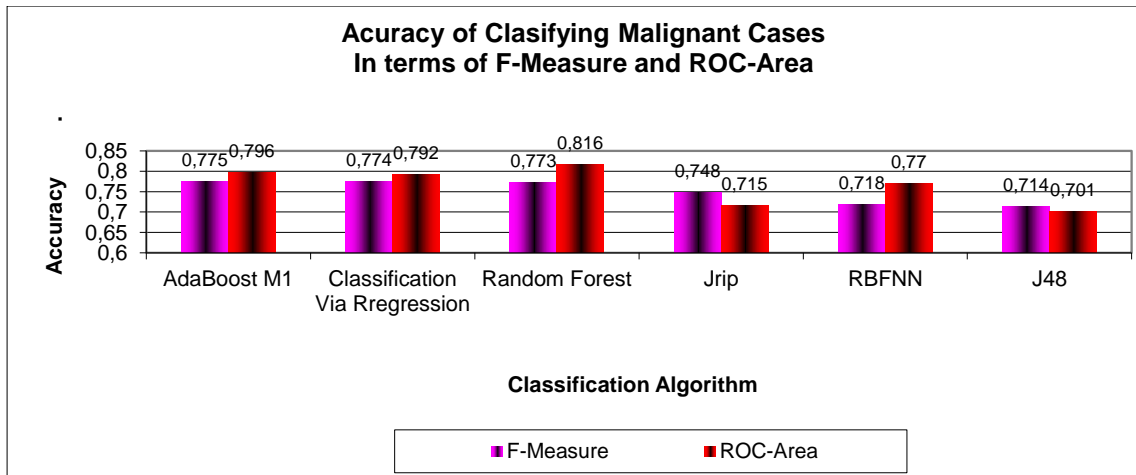


Figure 5.16: Accuracy of classifying of Malignant cases in terms of F measure and ROC area.

### 5.8.3 Average Performance of Classification with Both Cases of Weka

The following table shows the average performance of both cases in terms of different metrics.

Table 5.8: Class-wise Performance of Both Cases – Weka Evaluation.

Algorithm	TP-Rate	FP-Rate	Precision	Recall	F-Measure	ROC-Area
AdaBoost M1	0.75	0.257	0.75	0.75	0.75	0.796

Classification Via Regression	0.759	0.239	0.762	0.759	0.759	0.792
Random Forest	0.741	0.271	0.741	0.741	0.74	0.816
Jrip	0.716	0.296	0.715	0.716	0.715	0.715
RBFNN	0.716	0.271	0.729	0.716	0.716	0.77
J48	0.69	0.313	0.691	0.69	0.69	0.701

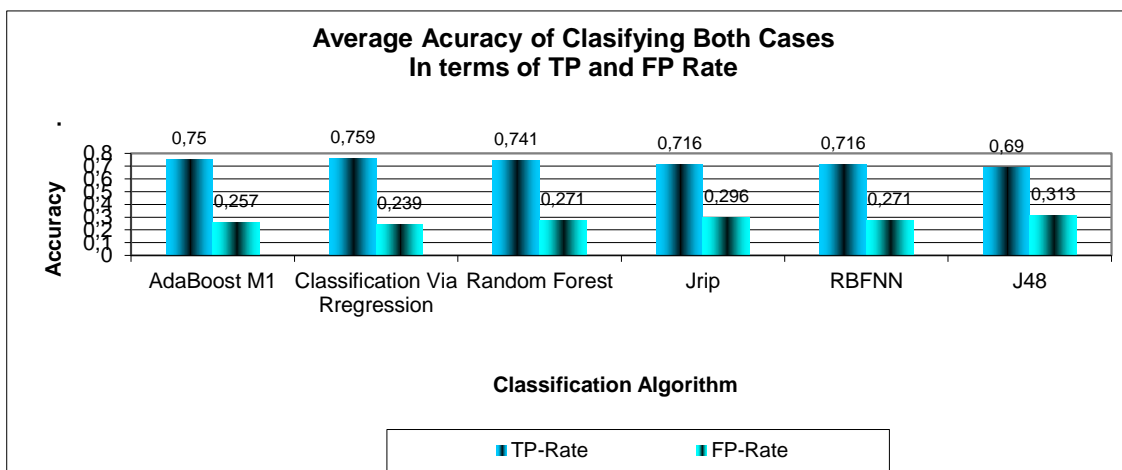


Figure 5.17: Accuracy of classifying of both cases in terms of TP and FP in Weka.

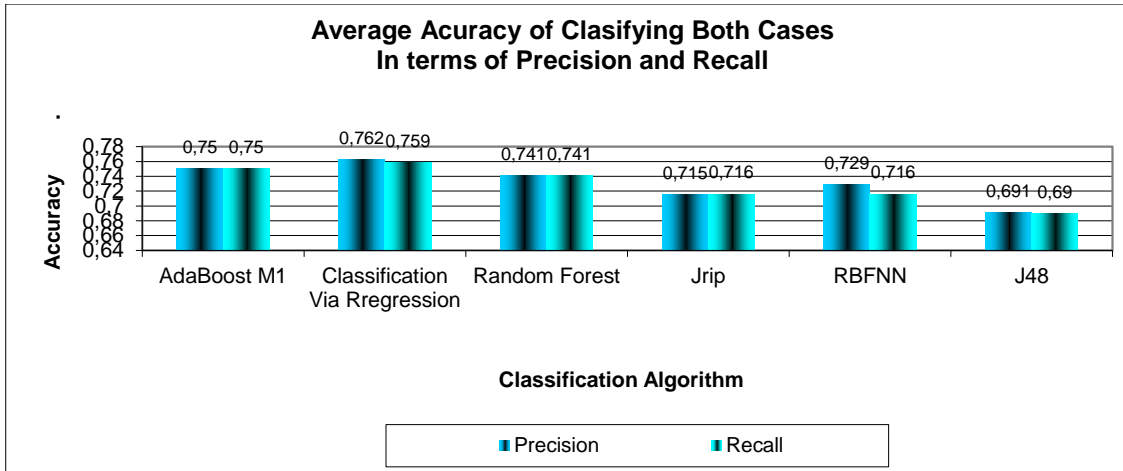


Figure 5.18: Accuracy of classifying of both cases in terms of Precision and Re call in Weka.

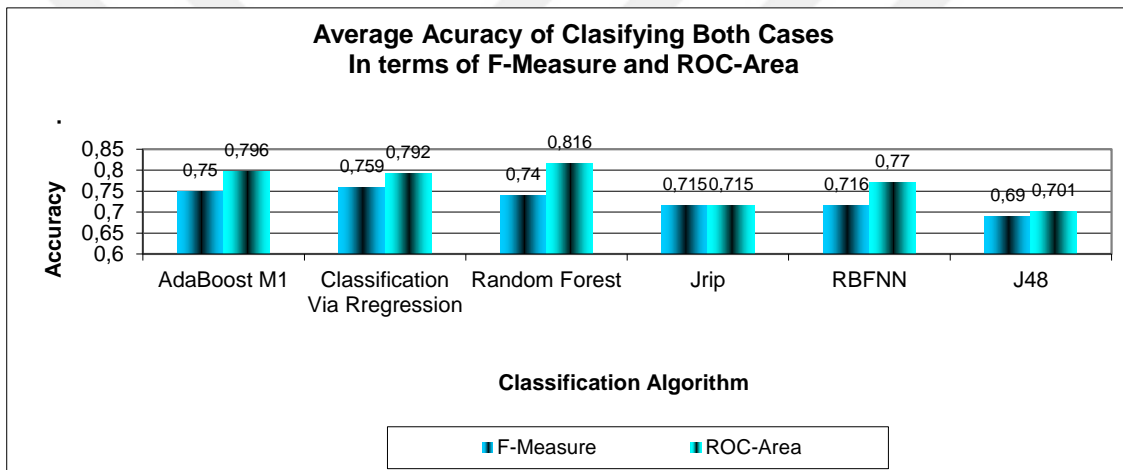


Figure 5.19: Accuracy of classifying of both cases in terms of F measure and ROC area in Weka.



## 6 DISCUSSION

This project is designed to classify the cancer data using several clustering techniques. However, in this project we have chosen the Matlab software as platform to simulate out proposed methods. The cancer data are recorded and sources in order to be classified for revealing the patterns associated with the said data. Different kind of cancers are planned to be obtained after applying the proposed method on the said dataset. Matlab is used for building the algorithms and importing the dataset, in other word, a graphical user interface have been made to be as frontend of the Back office procedure. The user need not to vary any kind of code if he is willing to use the designed software in this project. User may deal easily and smoothly with graphical user interface made in Matlab, import the dataset and apply the designated procedure on the said data. The use of several clustering technologies lead to conclusion more likely the algorithms of NAÏVE BAYES, J48 and RBF\_NN are used to cluster the breast cancer data and the performance of each method is recorded under both performance measure such as: specificity metrics, sensitivity metrics, precision metrics, accuracy metrics, error rate and F-score metrics; and validation measure such as k-fold cross validation. For latency and program run time, we found that NaiveBayes method is yielding the shortest time and producing the results in faster possible period that is equal to 0.07 seconds, whereas the other techniques such J48 and RBF-NN are resulting a latency of 0.09 seconds and 0.34 seconds respectively. As far as we surveyed, the previous works using this cancer dataset were only evaluated few, frequently used classifications such as C4.5 and SVM and little bit improved version of them for achieving good results. And of course, in some previous works, with the improved version of SVM and other techniques achieved very good results. But, there are LOT of classification algorithms available to test on this dataset. Even there are some very old classification algorithms such as AdaBoost, that were not at all addressed in most of the earlier works. Without testing the performance of such classical methods on this particular dataset, previous researchers just chose the state of the art algorithms like "SVM" only by belief and tend to improve them. For example, if AdaBoost will give best performance than any other classification algorithm, they why previous authors decided to select some other poor performing algorithm such as "SVM" or "RBF" (on this particular dataset) to design their improved model?

The scope of this project is only to do an extensive evaluation on several algorithms from each family of classification algorithms and discover which algorithm will really give better performance on this particular dataset. Our intention is not to compete any of the existing "state-of-the-art models" - instead we decide to find the best, standard classification algorithm which is much suitable for classifying the dataset in hand.

So, we have decided to select the best two algorithms from each family of algorithms for the comparisons. We selected the best two classifiers from the following the family of algorithms: Bayes (Bayes Variants), functions (Neural Network based Algorithms), lazy (lazy classifiers), meta (Meta-Classifiers), rules (Rule-based Classifiers), Trees(Tree-Based Classifiers)

The results of SVM is not presented because it didn't produce good results among the compared Neural Network Based Algorithm. RBF and MLP are the top best-performing algorithms among the all Neural Network Based Algorithms.

The average top-performing algorithms (in terms of precision) from each family are: Classification Via Rregression (0.762), AdaBoost (0.75), Random Forest (0.741), RBFNN (0.729), Jrip (0.715) and J48 (0.691)

The average top performing results in the other papers are different. For example RF(0.743), SVM (0.714), DT(0.686) where the best in the paer studying the performance evaluation of machine learning methods for breast cancer prediction[35].

Furthermore, Using TP-rate. FP-rate, Precision, Recall, F-measure, MCC, ROC area in my validation was not used with no any other paper studying the same data like in [36] and in [37]. Most of the papers mentioned didn't do k-fold validation so I think their results are inferior to my results.

After an extensive evaluation, we found that, the very old, historical algorithms called "Ada-Boost" and "classification via regression" gave very good results than all the compared algorithms including some of the so-called "state of the art neural network based algorithms. Why these two algorithms are able to give very good results on this dataset? this is a big research question that can be answered only through another exclusive research on this two algorithms. The characteristics of "Ada-Boost" and "classification via regression" should be analysed and incorporated in the future design of classification of medical datasets. A future work may address these issues.

## 7 Appendix I - The Matlab Code

### The Main User Interface Code Developed for this Project

```
function varargout = ClassificationOfCancerData(varargin)

% CLASSIFICATIONOFCANCERDATA M-file for ClassificationOfCancerData.fig

% CLASSIFICATIONOFCANCERDATA, by itself, creates a new

% CLASSIFICATIONOFCANCERDATA or raises the existing

% singleton*.

%

% H = CLASSIFICATIONOFCANCERDATA returns the handle to a new

% CLASSIFICATIONOFCANCERDATA or the handle to

% the existing singleton*.

%

% CLASSIFICATIONOFCANCERDATA('CALLBACK',hObject,eventData,handles,...)

% calls the local

% function named CALLBACK in CLASSIFICATIONOFCANCERDATA.M with the given

% input arguments.

% CLASSIFICATIONOFCANCERDATA('Property','Value',...) creates a new

% CLASSIFICATIONOFCANCERDATA or raises the
```

```
% existing singleton*. Starting from the left, property value pairs are

% applied to the GUI before ClassificationOfCancerData_OpeningFunction gets called. An

% unrecognized property name or invalid value makes property application

% stop. All inputs are passed to ClassificationOfCancerData_OpeningFcn via varargin.

%

% *See GUI Options on GUIDE's Tools menu. Choose "GUI allows only one

% instance to run (singleton)".

%

% See also: GUIDE, GUIDATA, GUIHANDLES

% Copyright 2002-2003 The MathWorks, Inc.

% Edit the above text to modify the response to help ClassificationOfCancerData

% Last Modified by GUIDE v2.5 13-Aug-2018 09:58:05

% This Matlab Code Developed for Classification of Cancer Data

% Using Classification Models Through Matlab-Weka Interfacing Techniques

%

% A Project "Using Data Mining for Classification of Breast Cancer"

%
```

```

% By Farah SharDouk.

% Under the Supervision of

% Assist. Prof. Dr. Adil Deniz Duru,

% Marmara University, Department of Physical Education and Sports Teaching

% Istanbul, Turkey.

% Begin initialization code - DO NOT EDIT

gui_Singleton = 1;

gui_State = struct('gui_Name', mfilename, ...

'gui_Singleton', gui_Singleton, ...

'gui_OpeningFcn', @ClassificationOfCancerData_OpeningFcn, ...

'gui_OutputFcn', @ClassificationOfCancerData_OutputFcn, ...

'gui_LayoutFcn', [] , ...

'gui_Callback', []);

if nargin && ischar(varargin{1})

gui_State.gui_Callback = str2func(varargin{1});

end

if nargout

```

```
[varargout{1:nargout}] = gui_mainfcn(gui_State, varargin{:});
```

```
else
```

```
gui_mainfcn(gui_State, varargin{:});
```

```
end
```

```
% End initialization code - DO NOT EDIT
```

```
% *****
```

```
% --- Executes just before ClassificationOfCancerData is made visible.
```

```
function ClassificationOfCancerData_OpeningFcn(hObject, eventdata, handles, varargin)
```

```
% This function has no output args, see OutputFcn.
```

```
% hObject handle to figure
```

```
% eventdata reserved - to be defined in a future version of MATLAB
```

```
% handles structure with handles and user data (see GUIDATA)
```

```
% varargin command line arguments to ClassificationOfCancerData (see VARARGIN)
```

```
% Choose default command line output for ClassificationOfCancerData
```

```
handles.output = hObject;
```

```
% Update handles structure
```

```

guidata(hObject, handles);

% UIWAIT makes ClassificationOfCancerData wait for user response (see UIRESUME)

% uiwait(handles.figure1);

javaaddpath([pwd '\wekackage\weka.jar']);

% *****

% --- Outputs from this function are returned to the command line.

function varargout = ClassificationOfCancerData_OutputFcn(hObject, eventdata, handles)

% varargout cell array for returning output args (see VARARGOUT);

% hObject handle to figure

% eventdata reserved - to be defined in a future version of MATLAB

% handles structure with handles and user data (see GUIDATA)

% Get default command line output from handles structure

varargout{1} = handles.output;

% *****

% --- Executes on button press in OpenARFFFormatBreastCancerCoimbraDataSet.

function OpenARFFFormatBreastCancerCoimbraDataSet_Callback(hObject, eventdata,
handles)

```

```
% hObject handle to OpenARFFFormatBreastCancerCoimbraDataSet (see GCBO)
```

```
% eventdata reserved - to be defined in a future version of MATLAB
```

```
% handles structure with handles and user data (see GUIDATA)
```

```
global ARFF_CancerDataFileName
```

```
global t_ARFF_CancerDataFileName
```

```
global DataAttributes
```

```
global DataLabels
```

```
[ARFF_CancerDataFileName,pfname] = uigetfile('*.arff','Select Input Cancer Dataset');
```

```
t_ARFF_CancerDataFileName=['P_' ARFF_CancerDataFileName];
```

```
fprintf('\n\nThe Selected Cancer Dataset is : %s ', ARFF_CancerDataFileName);
```

```
set(handles.SelectedFileName,'string',ARFF_CancerDataFileName);
```

```
[dataName,attributeName, attributeType, tdata]= arffread(ARFF_CancerDataFileName);
```

```
DataAttributes = tdata(:,1:end-1);
```

```
DataLabels = tdata(:,end);
```

```
axes(handles.axes1);
```

```
hold on
```

```
PCAGraph(DataAttributes,2,DataLabels)
```



```

title('\bfThe Cancer Data Projected in Virtual 2D Space')

xlabel('First Principal Component');

ylabel('Second Principal Component');

hold off

% *****

function PCAGraph(X,dims,labels)

%PCAGRAPH Plots data projected onto its first 2 principal components

% PCAGraph(X,dims,labels), where X = data, dims = no. of components

% to plot (2 or 3) and labels = class label of each entity (optional).

[U,S,V]=svd(X);

W=diag(S);

Xcen=X-repmat(mean(X),length(X),1); %centre data

Y=Xcen * V';

hold off ;

%plot each class with different marker

for i = 1:max(labels)

subset=(labels==i);

```

```

if dims == 2

plot(Y(subset,1),Y(subset,2),PlotColour(i),'MarkerSize',5);hold on

else

plot3(Y(subset,1),Y(subset,5),Y(subset,9),PlotColour(i),'MarkerSize',5);hold on

end

end

axis square; grid on

function c=PlotColour(index,lineflag)

%PLOTCOLOUR Returns colour and marker string for PLOT

% c=PLOTCOLOUR(index) where index is an integer, return

% next marker string in sequence

list=['b.':'g.':'k.':'r.':'m.':'y.':'b.':'r.':'k.':'c.'];

index=mod(index,size(list,1));

if index==0, index=size(list,1);end

c=list(index,:);

if exist('lineflag') & lineflag == 1

c=c(1);

```

```

end

% --- Executes on button press in VisualizetheCancerDatasetIn3D.

function VisualizetheCancerDatasetIn3D_Callback(hObject, eventdata, handles)

% hObject handle to VisualizetheCancerDatasetIn3D (see GCBO)

% eventdata reserved - to be defined in a future version of MATLAB

% handles structure with handles and user data (see GUIDATA)

global DataAttributes

global DataLabels

figure('name', 'Visualizing the Multidimensional Cancer Data in Virtual 3D Space',
'numbertitle', 'off');

PCAGraph(DataAttributes,3,DataLabels)

title('\bfThe Cancer Data Projected in Virtual 3D Space')

xlabel('First Principal Component');

ylabel('Second Principal Component');

zlabel('Third Principal Component');

% *****

function [dataName,attributeName, attributeType, data]= arffread(fileName)

% ARFFREAD Reads arff formatted file.

```

```
%  
  
% USAGE:  
  
% [dataName,attributeName, attributeType, data] = arffRead(fileName)  
  
%  
  
% INPUT:  
  
% fileName: file name to be read  
  
%  
  
% OUTPUT:  
  
% dataName: relation name of the arff file  
  
% attributeName: attribute name of attribute as cell array  
  
% { 1 by nAttr }  
  
% attributeType: attribute type of attribute as cell array  
  
% { 1 by nAttr }  
  
% data: data (nInstan by nAttr)  
  
%  
  
% Ported from the original code of Durga Lal Shrestha at Mathworks.com  
  
if nargin < 1,
```

```

error('No input arguments!');

end

if nargin > 1,

error('Too many input arguments!');

end

% read whole string

wholeData = textread(fileName,'%s','delimiter','\n','whitespace','');

atRelation = '@relation';

atAttribute = '@attribute';

atData = '@data';

noOfLines = size(wholeData,1);

k=0;

%% Finding data name

for i=1:noOfLines

k = findstr(wholeData{i},atRelation);

if k ~= 0;

lineAtRelation = i;

```

```

[token,dataName] = strtok(wholeData{lineAtRelation});

break

end

end

% Check whether dataName has whitespaces

tf = isspace(dataName);

tf = find(tf ==1);

if size(tf,2)>1

    dataName = dataName(2:tf(2)-1);

else

    dataName = dataName(2:size(dataName,2));

end

% Check whether dataName has semicolons or others

% First convert to ascii code and note that quotation mark is 39 is ascii

ascDataName = double(dataName);

if ascDataName(1) == 39

    ascDataName = ascDataName(2:end);

```

```
end
```

```
if ascDataName(end) == 39
```

```
ascDataName = ascDataName(1:end-1);
```

```
end
```

```
% Convert back to characters
```

```
dataName = char(ascDataName);
```

```
%% Finding attribute name
```

```
lineAtAttribute =[];
```

```
k=0;
```

```
l=0;
```

```
j = 0;
```

```
for i=lineAtRelation+1:noOfLines
```

```
k = findstr(wholeData{i},atAttribute);
```

```
if k ~= 0;
```

```
lineAtAttribute =[lineAtAttribute i];
```

```
[chopped,remainder] = strtok(wholeData{i});
```

```
[attrName,remAttrType] = strtok(remainder);
```

```
[attrType,rem] = strtok(remAttrType);
```

```
j=j+1;
```

```
attrVector{j} = attrName;
```

```
attrTypeVector{j} = attrType;
```

```
end
```

```
l = findstr(wholeData{i},atData);
```

```
if l ~= 0;
```

```
lineAtData = i;
```

```
break
```

```
end
```

```
end
```

```
% Finding whether data is tab formatted or csv and the position of data
```

```
k = [];
```

```
for i=lineAtData+1:noOfLines
```

```
str = wholeData{i};
```

```
if ~isempty(str) & ~strcmp(str,'%')
```

```
k = findstr(wholeData{i},',');
```



```
if ~isempty(k);

    dataFormat = 'comma' ;

    lineData = i;

    break

else

    dataFormat = 'tabOrSpace' ;

    lineData = i;

    break

end

end

end

%% Reading formatted data

%nRowSkip=lineData-1;

nColSkip = 0;

%dataName = dataName;

attributeName = attrVector ;

attributeType = attrTypeVector;
```

```

% {

% You have to convert " marks for each var to write in arff file

if strcmp(dataFormat,'comma')

data = csvread(fileName,nRowSkip);

elseif strcmp(dataFormat,'tabOrSpace') | strcmp(dataFormat,'tab')...

| strcmp(dataFormat,'Space')

data = dlmread(fileName,'\t',nRowSkip,nColSkip); % Space delimiter

dataFormat = 'space';

if size(data,2)~=size(attributeName,2)

data = dlmread(fileName,'\t',nRowSkip,nColSkip); % tab delimiter

dataFormat = 'tab';

end

if size(data,2)~=size(attributeName,2)

error('arff file is not tab or comma delemited!');

end

end

% }

```

```

strData = wholeData(lineData:end);

for i = 1:size(strData,1)

    data(i,:) = str2num(strData{i});

end

% *****

function arffwrite(fname,data)

% fname This is file name without extension

% data is m x n where m are the instances and n-1 are the features. The last

% column is the class as integer

% Ported from the original code of Muhammad at Mathworks.com.

% *****

sss=size(data,2)-1;

filename1=strcat(fname,'.arff');

out1 = fopen (filename1, 'w+');

aa1=strcat('@relation',{ ' },fname,'-weka.filters.unsupervised.attribute.NumericToNominal-
Rlast');

fprintf (out1, '%s\n', char(aa1));

for jj=1:sss

```

```

fprintf (out1, '@attribute %s numeric\n',num2str(jj));

end

n_classes=max(unique(data(:,end)));

txt1=strcat('@attribute',{' },num2str(sss+1),{' {'});

for ii=0:n_classes

txt1=strcat(txt1,num2str(ii),{' });

end

txt1=strcat(txt1,{' }');

fprintf (out1, '%s\n\n',char(txt1));

fprintf (out1, '@data\n');

fclose(out1);

dlmwrite (filename1, data, '-append' );

% *****

% --- Executes on button press in ClassifyUsingNaiveBayes.

function ClassifyUsingNaiveBayes_Callback(hObject, eventdata, handles)

% hObject handle to ClassifyUsingNaiveBayes (see GCBO)

% eventdata reserved - to be defined in a future version of MATLAB

```

```

% handles structure with handles and user data (see GUIDATA)

global ARFF_CancerDataFileName

if isempty(ARFF_CancerDataFileName) | strcmp(ARFF_CancerDataFileName, 'None')

    msgbox('Select a ARFF Format Cancer Data File and then run');

return;

end

%%% importing of classes

import java.util.*

import java.lang.*

import weka.classifiers.*

import weka.classifiers.functions.*

import weka.classifiers.trees.*

import java.io.*

import weka.core.*

import java.lang.*

%%% Reading and configuring data set

reader = javaObject('java.io.FileReader',ARFF_CancerDataFileName);

```

```

dataset = javaObject('weka.core.Instances',reader);

dataset.setClassIndex(dataset.numAttributes() - 1);

%%% getting the parameters of an arff file

relationName = char(dataset.relationName);

numAttr = dataset.numAttributes;

numInst = dataset.numInstances;

StartTime=cputime;

classifier = javaObject('weka.classifiers.bayes.NaiveBayes');

%classifier.setOptions(weka.core.Utils.splitOptions('-K 0 -M 1.0 -V 0.001 -S 1'));

%classifier.setOptions(weka.core.Utils.splitOptions('-K 0 -M 1.0 -S 1'));

%%% string for setting the options

v1 = String('-t');

v2 = String(ARFF_CancerDataFileName);

v3 = String('-x'); %%% if we want cross-validation

v4 = String('10'); %%% with number of folds

v5 = String('-v'); %%% Outputs no statistics for the training data.

v6 = String('-i'); %%% Outputs information-retrieval statistics for two-class problems.

```

```

%v7 = String('-k'); %%% Outputs information-theoretic statistics.

options = cat(1,v1,v2,v3,v4,v5,v6);

%% Creation of classifier Model

classifier.buildClassifier(dataset);

%% evaluation of the Classifier with options

output=weka.classifiers.Evaluation.evaluateModel(classifier,options);

ConsumedTime=cputime-StartTime;

PrintOutputMeasures('NaiveBayes',ConsumedTime, output);

% *****

% --- Executes on button press in ClassifyUsingJ48DecisionTree.

function ClassifyUsingJ48DecisionTree_Callback(hObject, eventdata, handles)

% hObject handle to ClassifyUsingJ48DecisionTree (see GCBO)

% eventdata reserved - to be defined in a future version of MATLAB

% handles structure with handles and user data (see GUIDATA)

global ARFF_CancerDataFileName

if isempty(ARFF_CancerDataFileName) | strcmp(ARFF_CancerDataFileName,'None')

msgbox('Select a ARFF Format Cancer Data File and then run');

```

```
return;end
```

```
%%% importing of classes
```

```
import java.util.*
```

```
import java.lang.*
```

```
import weka.classifiers.*
```

```
import weka.classifiers.functions.*
```

```
import weka.classifiers.trees.*
```

```
import java.io.*
```

```
import weka.core.*
```

```
import java.lang.*
```

```
%%% Reading and configuring data set
```

```
reader = javaObject('java.io.FileReader',ARFF_CancerDataFileName);
```

```
dataset = javaObject('weka.core.Instances',reader);
```

```
dataset.setClassIndex(dataset.numAttributes() - 1);
```

```
%%% getting the parameters of an arff file
```

```
relationName = char(dataset.relationName);
```

```
numAttr = dataset.numAttributes;
```



```

numInst = dataset.numInstances;

StartTime=cputime;

classifier = javaObject('weka.classifiers.trees.J48');

classifier.setOptions(weka.core.Utils.splitOptions('-c last -C 0.25 -M 2'));

%%% string for setting the options

v1 = String('-t');

v2 = String(ARFF_CancerDataFileName);

v3 = String('-x'); %%% if we want cross-validation

v4 = String('10'); %%% with number of folds

v5 = String('-v'); %%% Outputs no statistics for the training data.

v6 = String('-i'); %%% Outputs information-retrieval statistics for two-class problems.

%v7 = String('-k'); %%% Outputs information-theoretic statistics.

options = cat(1,v1,v2,v3,v4,v5,v6);

%%% Creation of classifier Model

classifier.buildClassifier(dataset);

%%% evaluation of the Classifier with options

output=weka.classifiers.Evaluation.evaluateModel(classifier,options);

```

```

ConsumedTime=cputime-StartTime;

PrintOutputMeasures('J48',ConsumedTime, output);

%*****
*

% --- Executes on button press in ClassifyUsingRBFNeuralNetwork.

function ClassifyUsingRBFNeuralNetwork_Callback(hObject, eventdata, handles)

% hObject handle to ClassifyUsingRBFNeuralNetwork (see GCBO)

% eventdata reserved - to be defined in a future version of MATLAB

% handles structure with handles and user data (see GUIDATA)

global ARFF_CancerDataFileName

if isempty(ARFF_CancerDataFileName) | strcmp(ARFF_CancerDataFileName,'None')

msgbox('Select a ARFF Format Cancer Data File and then run');

return;

end

%%% importing of classes

import java.util.*

import java.lang.*

import weka.classifiers.*

```

```

import weka.classifiers.functions.*

import weka.classifiers.trees.*

import java.io.*

import weka.core.*

import java.lang.*

%%% Reading and configuring data set

reader = javaObject('java.io.FileReader',ARFF_CancerDataFileName);

dataset = javaObject('weka.core.Instances',reader);

dataset.setClassIndex(dataset.numAttributes() - 1);

%%% getting the parameters of an arff file

relationName = char(dataset.relationName);

numAttr = dataset.numAttributes;

numInst = dataset.numInstances;

StartTime=cputime;

classifier = javaObject('weka.classifiers.functions.RBFNetwork');

classifier.setOptions(weka.core.Utils.splitOptions('-B 2 -S 1 -R 1.0E-8 -M -1 -W 0.1'));

```

```

%%% string for setting the options

v1 = String('-t');

v2 = String(ARFF_CancerDataFileName);

v3 = String('-x'); %%% if we want cross-validation

v4 = String('10'); %%% with number of folds

v5 = String('-v'); %%% Outputs no statistics for the training data.

v6 = String('-i'); %%% Outputs information-retrieval statistics for two-class problems.

%v7 = String('-k'); %%% Outputs information-theoretic statistics.

options = cat(1,v1,v2,v3,v4,v5,v6);

%%% Creation of classifier Model

classifier.buildClassifier(dataset);

%%% evaluation of the Classifier with options

output=weka.classifiers.Evaluation.evaluateModel(classifier,options);

ConsumedTime=cputime-StartTime;

PrintOutputMeasures('RBF-NN',ConsumedTime, output);

% *****

% --- Executes on button press in RunAllClassifiers.

```

```

function RunAllClassifiers_Callback(hObject, eventdata, handles)

% hObject handle to RunAllClassifiers (see GCBO)

% eventdata reserved - to be defined in a future version of MATLAB

% handles structure with handles and user data (see GUIDATA)

global ARFF_CancerDataFileName

if isempty(ARFF_CancerDataFileName) | strcmp(ARFF_CancerDataFileName, 'None')

msgbox('Select a ARFF Format Cancer Data File and then run');

return;

end

% *****

fprintf('\n-----');

hh=waitbar(1/3, ['Classifying Cancer Data with Naive Bayes Classifier']);

ClassifyUsingNaiveBayes_Callback;

waitbar(2/3, hh, ['Classifying Cancer Data with C4.5 Classifier' ]);

ClassifyUsingJ48DecisionTree_Callback

waitbar(3/3, hh, ['Classifying Cancer Data with Radial Basis Functions Neural Network
Classifier']);

ClassifyUsingRBFNeuralNetwork_Callback;

```

```

close(hh);

fprintf('\n-----');

% *****

% *****

function PrintOutputMeasures(classifierName, ConsumedTime,output)

%%% Removing Unnecessary lines From output

fprintf('\nPerformance of %s Based Method', classifierName);

fprintf('\nTime Consumed For Training and Testing %5.2f sec\n', ConsumedTime);

coutput=output.toCharArray';

Lines= regexp(coutput, '\n+', 'split');

% whos('Lines');

ss=size(Lines,2);

for i=ss-8:ss-5

fprintf('%s\n', Lines{i});

end

% *****

```

## 8 References

- [1]. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. *CA Cancer J Clin.* 2011/
- [2]. Hwa HL, et al. Prediction of breast cancer and lymph node metastatic status with tumor markers using logistic regression models. *J Eval Clin Pract.* 2008.
- [3]. Santillan-Benitez JG, et al. The tetrad BMI, Leptin, Leptin/Adiponectin (L/a) ratio and CA 15-3 are reliable biomarkers of breast cancer. *J Clin Lab Anal.* 2013.
- [4]. Dalamaga M, et al. Serum resistin: a biomarker of breast cancer in postmenopausal women? Association with clinicopathological characteristics, tumor markers, inflammatory and metabolic parameters. *Clin Biochem.* 2013.
- [5]. Assiri AM, Kamel HF. Evaluation of diagnostic and predictive value of serum adipokines: Leptin, resistin and visfatin in postmenopausal breast cancer. *Obes Res Clin Pract.* 2015.
- [6]. Patrício M, Caramelo F. Comment on "evaluation of diagnostic and predictive value of serum adipokines: Leptin, resistin and vistafin in postmenopausal breast cancer". 2016.
- [7]. Umesh D R ; B Ramachandra ‘Association rule mining based predicting breast cancer recurrence on SEER breast cancer data’.
- [8]. Galal, G., Cook, D.J., Holder, L.B. “Improving Scalability in a Scientific Discovery System by xploiting Parallelism”, *Proceedings KDD '97.*
- [9]. Holsheimer, M., Kersten, M., Mannila, H., Toivonen, H. “A Perspective on Databases and Data Mining”, *Proceedings KDD '95.*
- [10]. John, G.H., Lent, B. “SIPping from the Data Firehose”, *Proceedings KDD '97.*
- [11]. Toivonen, H. “Discovery of frequent patterns in large data collections”, *PhD Thesis, 1996.*
- [12]. Gabowski, H., Lossack and Weibkopf, “Automatic Classification and Creation of Classification Systems Using Methodologies of Knowledge Discovery in Databases,” Chapter 5 (pp. 127-144) of *Data Mining for Design and Manufacturing: Methods and Applications* edited by D. Braha, Kluwer Academic Publishers: New York.
- [13]. Porter, A. L., Kongthon, A., and Lu, J. C.) “Research Profiling – Improving the Literature Review: Illustrated for the Case of Data Mining of Large Datasets,”
- [14]. Teófilo Campos, “PCA for face recognition” Creativision research group, IME - USP – Brazil

- [15]. Jain K, Murty MN, Flynn PJ. Data clustering: a review, *ACM Computing Survey* 1999; 31: 264-323.
- [16]. Wang X, Zhao Y, Liu R, Jing Z. Knowledge-transfer analysis based on co-citation clustering. *Scientometrics* 2013; 97: 859-869.
- [17]. Jiang H, Lou W, Wang W. Three-Tier Clustering: An Online Citation Clustering System. *Advances in Web-age information management* 2001; 2118: 237-248.
- [18]. Kejzar N, Korenjak-Černe S, Batagelj V. Clustering of Distributions: A Case of Patent Citations. *J Classi* 2011; 28: 156-183.
- [19]. Aljaber B, Stokes N, Bailey J, Pei J. Document clustering of scientific texts using citation contexts. *Info Retri* 2010; 13: 101-131.
- [20]. Fu-Xin R, Xue-Qi CH, Hua-WS. Modeling the clustering in citation networks. *Stat Mech Appl* 2012; 391: 3533-3539.
- [21]. Yongjing L, Wenyuan Li, Keke CH, Ying L. A Document Clustering and Ranking System for Exploring MEDLINE Citations. *J Am Med Inform Assoc* 2007; 14: 651-661.
- [22]. Xiaoran XU, Zhi HD. BibClus: A Clustering Algorithm of Bibliographic Networks by Message Passing on Center Linkage Structure. *IEEE International Conference on Data Mining*, 2011; 864-873.
- [23]. Tomonari M, Atsuhiko T, Jun A. Citation data clustering for author name disambiguation, *ACM International conference on Scalable Information Systems*, 2007; 1-8.
- [24]. Frizo J, Wolfgang G, Bart DM. Dynamic hybrid clustering of bioinformatics by incorporating text mining and citation analysis. *ACM SIGKDD international conference on Knowledge discovery and data mining* 2007; 360-369.
- [25]. Bolelli L, Ertekin S, Giles LC. Clustering Scientific Literature Using Sparse Citation Graph Analysis, *Knowledge Discovery in Databases. Lect Not Comp Sci* 2006; 4213: 30-41.
- [26]. Vasileios K, Lyle U, Rajeev A. Online Clustering and Citation Analysis Using Streemer. *Publicly Accessible Penn database*, 2009.
- [27]. Hui H, Hongyuan Z, Giles CL. Name disambiguation in author citations using a K-way spectral clustering method. *ACM/IEEE-CS Joint Conference on Digital Libraries* 2005; 334-343.
- [28]. Fujita K, Kajikawa Y, Mori J, Sakata I. Detecting research fronts using different types of weighted citation networks. *International Conference on Technology Management for Emerging Technologies* 2012; 267-275.



- [29]. Michael JB II, Daniel MK, Jonathan LZ, James HF. Distance measures for dynamic citation networks. *Physica A* 2010; 389: 4201-4208.
- [30]. Parthasarathy G, Tomar DC. Sentiment analyzer: Analysis of journal citations from citation databases. *Confluence The Next Generation Information Technology Summit*, 2014.
- [31]. Egghe L, Rousseau R. BRS-Compactness in Networks: Theoretical Considerations Related to Cohesion in Citation Graphs, Collaboration Networks and the Internet. *Math Comp Mod* 2003; 37: 879-899.
- [32]. Judit BI, Mark L, Ayelet L. Some measures for comparing citation databases. *J Inform* 2007; 1: 26-34.
- [33]. Chyan Y, Szu-Hui WU, Lee J. A study of collaborative product commerce by co-citation analysis and social network analysis. *IEEE International Conference on Industrial Engineering and Engineering Management*, 2007; 24: 209-213.
- [34]. Chen D, Chi HC, Jing D, Chun LD. Citation retrieval in digital libraries. *IEEE International Conference on Systems, Man, and Cybernetics*, 1999; 105-109.
- [35]. Evaluation of Machine Learning Methods for Breast Cancer Prediction. *Applied and Computational Mathematics*. Vol. 7, No. 4, 2018, pp. 212-216. Yixuan Li, Zixuan Chen. Performance.
- [36]. Crisóstomo, J., Matafome, P., Santos-Silva, D. et al. *Endocrine* (2016) 53: 433.
- [37]. Miguel Patrício, José Pereira, Joana Crisóstomo, Paulo Matafome, Manuel Gomes, Raquel Seica and Francisco Caramelo. Using Resistin, glucose, age and BMI to predict the presence of breast cancer.