American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)
ISSN (Print) 2313-4410, ISSN (Online) 2313-4402

http://asrjetsjournal.org/

# Classification of Breast Cancer Using Data Mining

Farah Sardouk[a]*, Dr. Adil Deniz Duru[b], Dr. Oğuz Bayat[c]

[a]*MSc. Department Electrical Engineering, Altinbaş University, Istanbul, Turkey*
[b]*Assist. Prof. Department of Physical Education and Sports, Marmara University, Istanbul, Turkey*
[c]*Assoc. Prof. Department of Electrical Engineering, Altinbaş University, Istanbul, Turkey*
[a]*Email: farahsardouk@ogr,altinbas.edu.tr*
[b]*Email: deniz.duru@marmara.edu.tr*
[c]*Email: oguz.bayat@altinbas.edu.tr*

**Abstract**

According to the publications of leading health organization in the world, the World Health Organization (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 sickness may begin at early stages for combating purpose. 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 had resulted in considerable results that may help for breast cancer aid. In this research, Coimbra dataset are collected and studied according to 10 predictors. We used these predictors to estimate if the 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.

***Keywords*:** ANN; Artificial Neural Network; BMI; KDD; k-fold cross validation; PPV; WHO.

## 1. Introduction

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 beneficial 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. The data about particular observations are forming a multi-dimensional database, the same is abstracted with data outline. In this study, we have discussed the possibility of classification process on this database.

---

* Corresponding author.

Assuming the existence of reasonable observation in this database, the randomness of the observations is causing a drawback while clustering the data. In attempts to tackle this difficulty, a rudest decision rule algorithm is proposed based on the truncation principle. However, the performance of the classification is enhanced to be double than it was before [1]. 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 the measure the performance on clustering is underlie by the fact the more none-similar objects (points) under the same cluster is directly lead to performance disorder. In other words, 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 of 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 [2]. 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. 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 is increasing the 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 (4). 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. 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 such as (Decision tree, neural network, Bayesian network, K-Nearest Neighbor algorithm, DNF rules, Genetic algorithm, Fuzzy and Rough sets) 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.

## 2. PCA Algorithm

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 ith elements, more likely, S(i) where i=1,2,3,4…n. the principle components can be recognized from this linear data by using the following formula.

$$\boldsymbol{f(S,V) = (sV)V^T + u} \qquad (1)$$

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 matric 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.

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

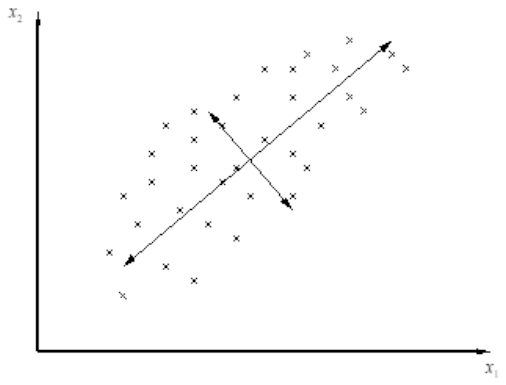


**Figure 1:** The outcomes of the principal component analysis algorithm.

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

### 3. Performance Metrics

To trigger the efficiency of the clustering algorithm many set of measurements' agreements are made. In the process 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. The following concepts are used for measuring the performance of classification:

**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. To get away with the confusion matric, the figure below is demonstrating the sample content of the matrix.

**Table 1:** The matrix of confusion in performance assessment of classifiers

| Hypothesis | | |
|---|---|---|
| + | - | Actual Class |
| a | b | + |
| c | d | - |

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 is also misclassified, and the result of classification is reversed 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, the classification process is 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 a 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.

$$\text{Sensitivity=Recall= } TP*(FN+TP)^{-0.5} \quad (3)$$

3) Specificity: similarly, like the sensitivity, when the database is involving a set of positive and negative values, and however, the classification process is 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.

$$\text{Specificity= } TN*(FP+TN)^{-0.5} \quad (4)$$

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.

Accuracy= (TN+TP)*(FN+TP+TN+FP)^-0.5

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

$$PPV=TP*(FP+TP) \quad (5)$$

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. Simulated model

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''. This Data set is originally prepared by the Faculty of Medicine of the University of Coimbra. It has 10 predictors. These predictors can potentially be used as a biomarker of breast cancer. The User Interface Showing the Multidimensional Data Projected in Virtual 2D Space is given in the figure below:

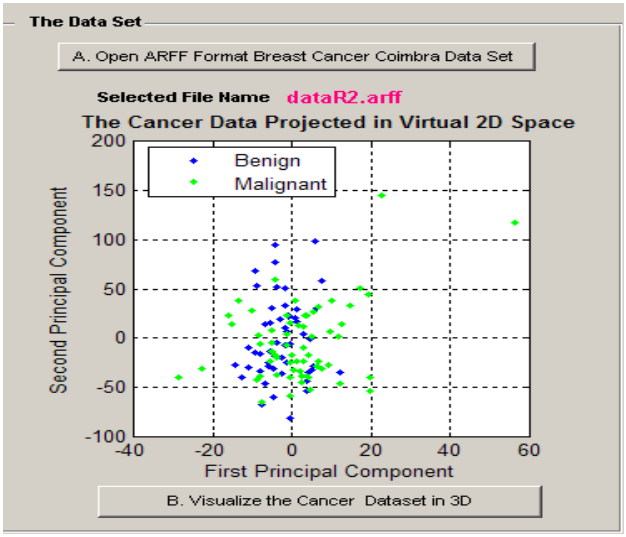


**Figure 2:** The cancer data projected in virtual 2D space

The figure below is showing the multidimensional data projected in virtual 3D Space:

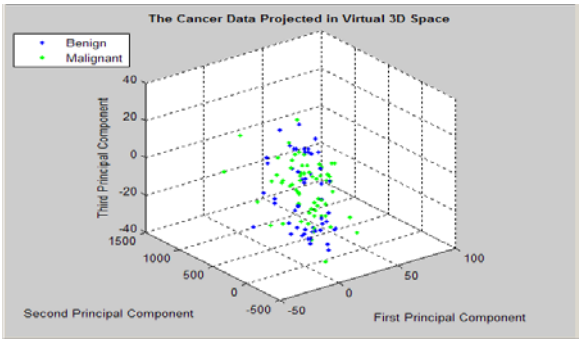


**Figure 3:** The cancer data projected in virtual 2D space

## 5. Performance of Classification

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

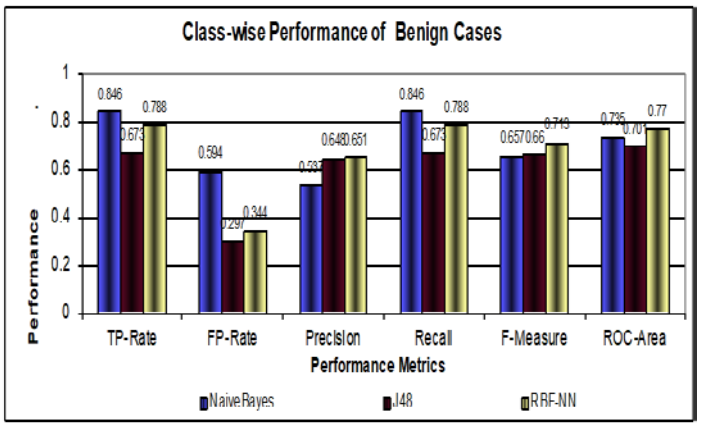


**Figure 4:** The class performance of Benign cases.

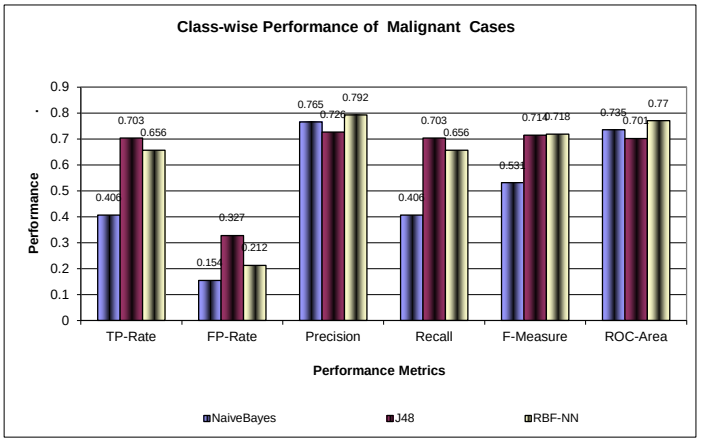


**Figure 5:** Average Performance of Classification with Both Cases

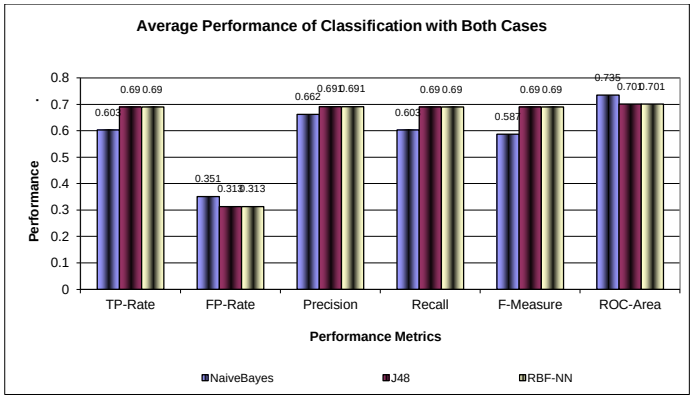


**Figure 6:** Graphical representation of average performance from both cases.

Weka is data mining software that uses a collection of machine learning algorithms (6). These algorithms can be applied directly to the data or called from the Java code. It also used for applying the classification on the cancer data and hence the following results are obtained.

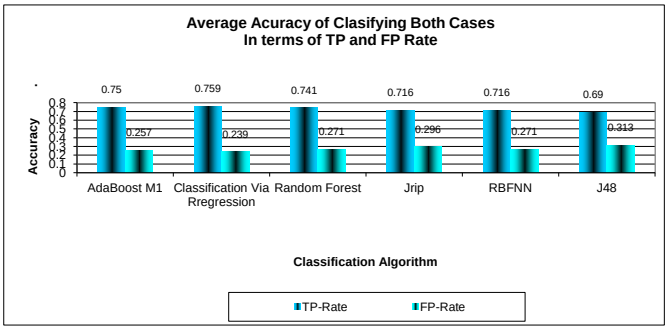


**Figure 7:** Accuracy of classifying of both cases in terms of TP and FP in Weka.

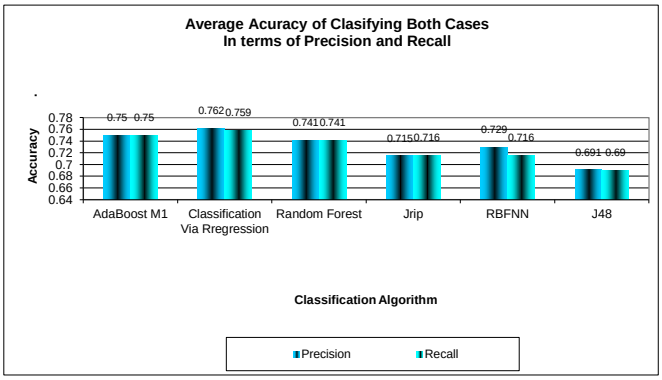


**Figure 8:** Accuracy of classifying of both cases in terms of Precision and Recall in Weka.

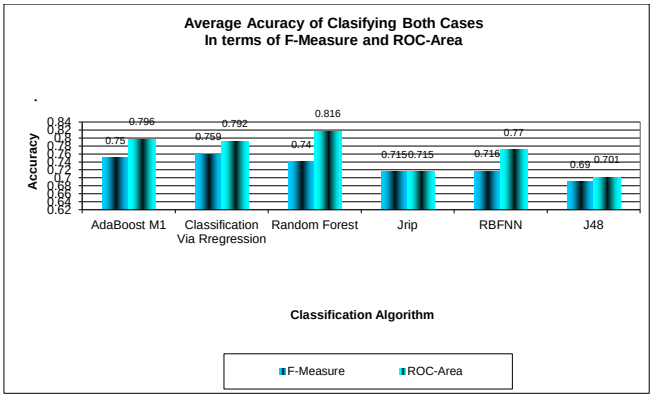


**Figure 9:** Accuracy of classifying of both cases in terms of F measure and ROC area in Weka.

## 6. Average Performance of Classification with Both Cases of Weka (Benign and Malignant)

The following table shows the class-wise performance of Malignant cases in terms of 6 different algorithms.

**Table 2:** Figure **Error! No text of specified style in document.**.1: 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 |

## 7. Discussion

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.

## 8. Conclusion

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[8]. 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 [9] and in [10]. Most of the papers mentioned didn't do k-fold validation so I think their results are inferior to my results.

## References

[1]. Umesh D R ; B Ramachandra 'Association rule mining based predicting breast cancer recurrence on SEER breast cancer data'.

[2]. Galal, G., Cook, D.J., Holder, L.B. "Improving Scalability in a Scientific Discovery System by exploiting Parallelism", Proceedings KDD '97.

[3]. Holsheimer, M., Kersten, M., Mannila, H., Toivonen, H. "A Perspective on Databases and Data Mining", Proceedings KDD '95.

[4]. William B. Schwartz, M.D., Ramesh S. Patil, Ph.D., and Peter Szolovits, Ph.D. "Artificial Intelligence in Medicine'', Volume 34, Issue 2, June 2005, Pages 113-127.

[4]. Ultrasound in Medicine & Biology, Volume 29, Issue 5, May 2003, Pages 679-686.

[5]. Ian Witten Eibe Frank Mark Hall Christopher Pal, Data Mining: Practical Machine Learning Tools and Techniques, Fourth Edition, Pages 160, 161.

[6]. Chen D, Chi HC, Jing D, Chun LD. Citation retrieval in digital libraries' International Conference on Systems, Man, and Cybernetics, 1999; 105-109.

[7]. 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.

[8]. Crisóstomo, J., Matafome, P., Santos-Silva, D. et al. Endocrine (2016) 53: 433.

[9]. Miguel Patrício,José Pereira, Joana Crisóstomo, Paulo Matafome, Manuel Gomes, Raquel Seiça and Francisco Caramelo, "Using Resistin, glucose, age and BMI to predict the presence of breast cancer''.