Implementation of Data Mining Techniques in Breast-Cancer by  
WEKA TOOL

Md. Ashikuzzaman1, Md. Ishtiak Harun Shovon2, Md. Shahrier Sarder3, Fahim Talha4

Student ID: 2015-1-60-0351, 2015-1-60-1222, 2015-1-60-0813, 2014-3-60-0044

***1234****Undergraduate Student, Computer Science & Engineering, EWU, Dhaka-1212, Bangladesh*

Email: 1ashikuzzamannoyon@yahoo.com, [2shovon.ishtiak@gmail.com](mailto:2shovon.ishtiak@gmail.com), [3shahriersarder@gmail.com](mailto:3shahriersarder@gmail.com), 4fahimtalha@gmail.com

***Abstract: -* Data mining (DM) comprises the core algorithms that enable to gain fundamental insights and knowledge from massive data. In fact, data mining is a part of a larger knowledge discovery process. One of the new researches in data mining application involves analyzing Breast cancer, which are the deadliest disease and most common of all cancers in the leading cause of cancer deaths in women worldwide. Among the various DM techniques, classification plays a vital role in DM research. Breast cancer diagnosis and prognosis are two medical applications pose a great challenge to the researchers in medical field.[1] This survey work analyses the various review and technical articles on breast cancer diagnosis. The main goal of this research is to explore the overview of the current research being carried out using the data mining techniques to enhance the breast cancer diagnosis. Here, we used Naive Bayes, Logistic Regression, Random Forest and Decision Tree to analyze the Breast Cancer data from WEKA.**

***Keyword: Disease, Breast Cancer, Data Mining, breast cancer*** [***diagnosis***](http://ieeexplore.ieee.org/search/searchresult.jsp?matchBoolean=true&queryText=%22Index%20Terms%22:.QT.breast%20cancer%20diagnosis.QT.&newsearch=true)***,,***[***classification algorithms***](http://ieeexplore.ieee.org/search/searchresult.jsp?matchBoolean=true&queryText=%22Index%20Terms%22:.QT.classification%20algorithms.QT.&newsearch=true)***,***[***ID3***](http://ieeexplore.ieee.org/search/searchresult.jsp?matchBoolean=true&queryText=%22Index%20Terms%22:.QT.ID3.QT.&newsearch=true)***Supervised Learning, Unsupervised Learning, Naïve, Logistic, Random Forest, Decision Tree. , Prediction, WEKA.***

I. INTRODUCTION

Data mining is considered to deal with huge amounts of data which are kept in the database, to locate required information and facts. Breast cancer is an uncontrolled growth of breast  
cells. It refers to a malignant tumor that has  
developed from cells in the breast. Breast Cancer constitutes a major public health issue globally with over 1 million new cases diagnosed annually, resulting in over 400,000 annual deaths and about 4.4 million women living with the disease. It also affects one in eight women during their lives. It is the commonest site specific malignancy affecting women and the most common cause of cancer mortality in women worldwide. It is also found in  
men but not very common. [2] Here, we will apply some algorithms such as Nave Bayes, Logistic Regression, Random Forest and Decision tree to find accuracy of those Breast Cancer data from **weka**. In this paper our main objective is to find accuracy of those models and to find association. In our implementation we will discuss about this model in detail. Cancer occurs as a result of mutations, or abnormal changes, in the genes responsible for regulating the growth of cells and keeping them healthy. [3]That changed cell gains the ability to keep dividing without control or order, producing more cells just like it and forming a tumor. A tumor can be benign (not dangerous to health) or malignant (has the potential to be dangerous)

II. Related Work

In their work we used different data mining technique for diagnosis & the prognosis of breast cancer with the main parameter of male and female gene behavior, they take gene expression data set of 311 instance to test and validate model and major the performance. They prove classification data mining algorithm provide more optimum outcome of this paper focuses on diagnosing of early Brest cancer in women using SVM, Tree Boost and Tree Forest data mining classification technique. We focus on SEER public use-data to predict Breast Cancer. They use pre-classification method and find a possible solution to discover their formation of Breast Cancer. We built a classification model using various intelligent techniques such as ANN(Artificial Neural Network),Unsupervised Artificial Neural Network, Statistical technique and decision tree. Experimental results show a testing accuracy of 97.73% from which the efficiency of the ensemble model was highlighted.

**III Causes of breast cancer  
A. Genetic factors**1.**Gender**: Breast cancer occurs nearly 100 times  
more often in women than in men.  
2.**Age**: Two out of three women with invasive  
cancer is diagnosed after age 55.  
3.**Race:** Breast cancer is diagnosed more often in  
women and it occurs in every races.  
4.**Family history and genetic factors:** If a relative  
has been diagnosed with breast or ovarian  
cancer, such person has a higher risk of being  
diagnosed before the age of 50.  
5.**Personal health history:** if a person has been  
diagnosed with breast cancer in one breast, there  
is an increased risk of been diagnosed with  
breast cancer in the other breast in the future.  
6.**Menstrual and reproductive history:** Early  
menstruation (before age 12), late menopause  
(after age 55), having your first child at an older  
age, or never having given birth can also  
increase the risk of breast cancer.  
7. **Certain genome:** this is cause by mutation in  
certain genes and can be determined through  
genetic test as individual with certain gene  
mutation can pass it onto their children.  
8. **Dense breast tissue:** this can increase the risk of  
having breast cancer and makes lumps harder to  
detect.

IV. PROPOSED METHOD

For more accurate results we are using data mining technique i.e. classification. Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. Apply Classification technique on the database using Weka tool.[5] This step will provide us with accurately predict the values of Breast Cancer. This process consist of various consecutive stages communicating with each other as data preprocessing, data partitioning and classification rule mining. We would like to analyze the breast cancer data collected from Weka tool and try to predict the disease risk. For analyzing here we applied some algorithms to find the accuracy. We applied the following algorithms:

1. Nave Bayes

2. Logistic Regression

3. Random Forest

4. Decision tree

## V. IMPLEMENTATION

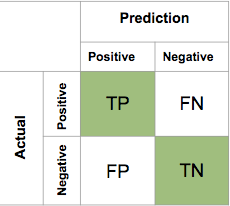
1. ***Cross Validation Score***

For cross-validation, the scores are summed and then divided by the number of cases, so here the score is an average case likelihood. It is demonstrate approval procedure for surveying how the consequence of examination will sum up to free dataset. In spite of the fact that we will probably anticipate certain hazard so we utilize cross approval to gauge the precision. In cross approval, each record is utilized a similar number of times for preparing information and precisely once to test. We utilized 10-crease cross approval technique that has been precluding the likelihood of over fitting the information. Along these lines, CV score of accuracy, precision, recall and F measure have been considered for lifting the execution of our model.

### **Accuracy, Precision, Recall and F measure**

Performance of a model is measured by reflection of well observed actual events. In this paper we used confusion matrix which gives a relation between actual and predicted class. From this confusion matrix we have measured the Accuracy, Precision and Recall.

Here true positive and true negative are the correctly predicted observation but the term false positive and false negative can be confusing. False negatives are observations where the actual event was positive. The way to think about it is that the terms refer to the observations and not the actual events.



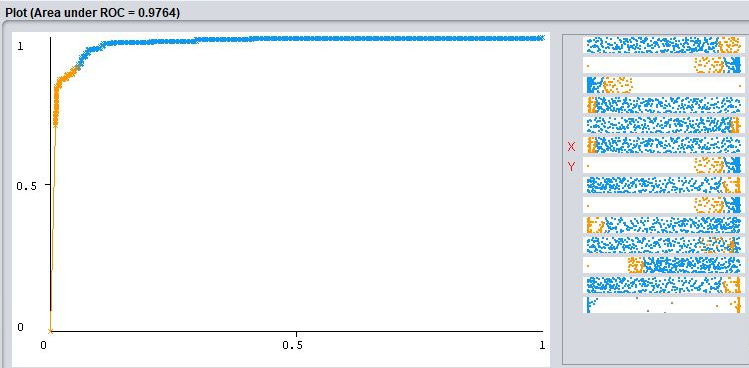
## VI. NAIVE BAYES

Naive bayes is a model that assign class level to problem instance. This classification uses theorem that describe the probability of an events based on prior knowledge of condition that should be related to the events. We used it because it requires a small number of training data to estimate the parameters necessary for classification.

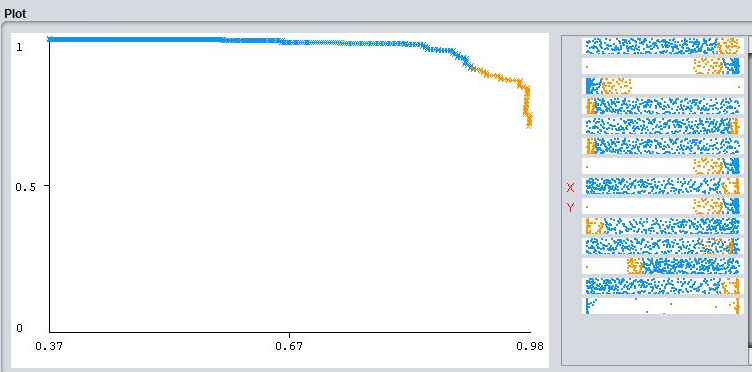
**Table 1: Performance measure – Naïve Bayes**

|  |  |
| --- | --- |
| Evaluating Measure Naïve Bayes | 10-fold Cross Validation |
| Accuracy | 92.615 % |
| Precision | 0.926 |
| Recall | 0.926 |
| F measure | 0.926 |
| Error | 7.381 % |

**Receiver Operating Characteristics (ROC) graph for tested negative class:**



**Precision –Recall curve:**



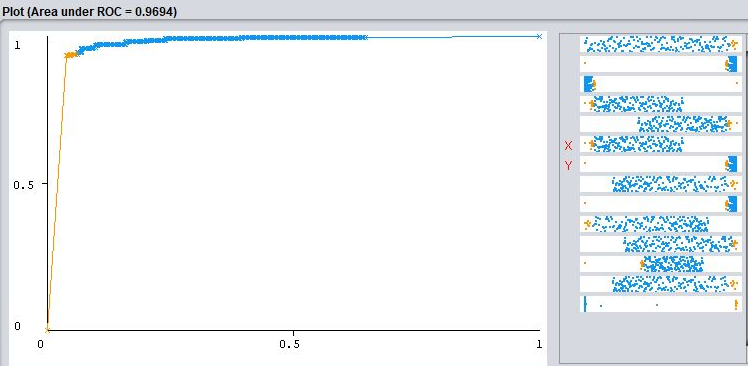
## VII. LOGISTIC REGRESSION

Logistic regression is used to predict the risk of developing a given disease based on observed characteristics of the patient. Thus using this model we can find the correlation.

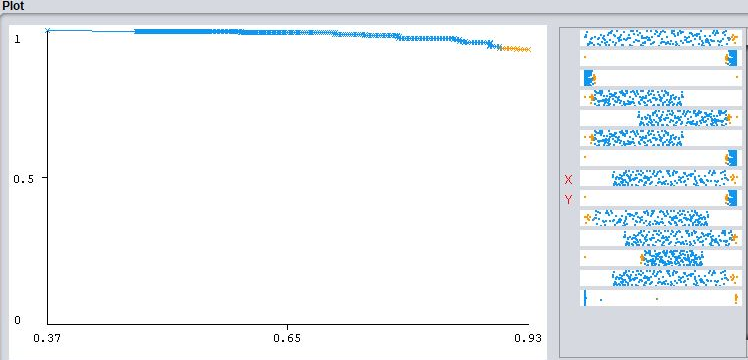
**Table 2: Performance measure – Logistic Regression:**

|  |  |
| --- | --- |
| Evaluating Measure  Logistic Regression | 10-fold Cross Validation |
| Accuracy | 94.024% |
| Precision | 0.941 |
| Recall | 0.940 |
| F measure | 0.940 |
| Error | 5.975 % |

**ROC curve for Logistic Regression Model:**



**Precision –Recall curve:**

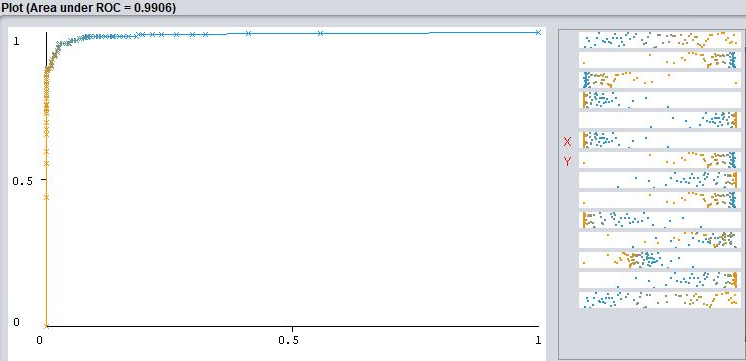


## VIII. RANDOM FOREST

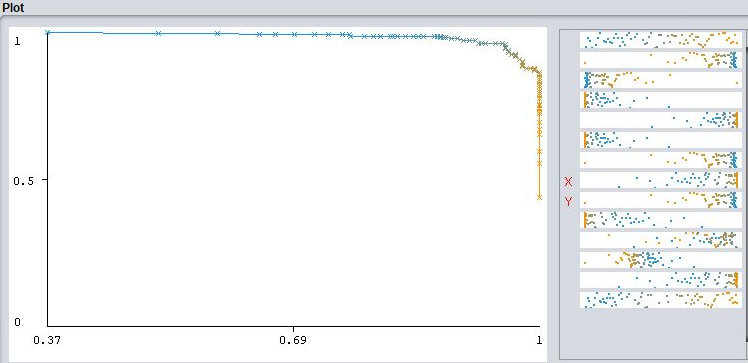
Random forest operated by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classification or prediction of the individual trees.

|  |  |
| --- | --- |
| Evaluating Measure  Random Forest | 10-fold Cross Validation |
| Accuracy | 95.958% |
| Precision | 0.960 |
| Recall | 0.960 |
| F measure | 0.958 |
| Error | 4.042% |

**ROC curve for Logistic Regression Model:**



**Precision –Recall curve:**



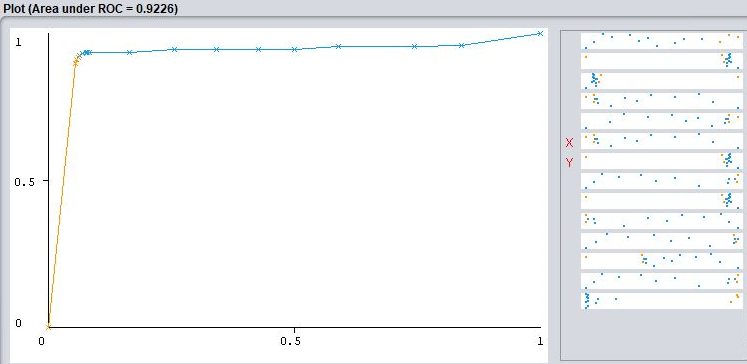
## IX. DECISION TREE

We used decision tree method to illustrate every possible outcome of a decision. It was built to predict the target column after splitting the data set.

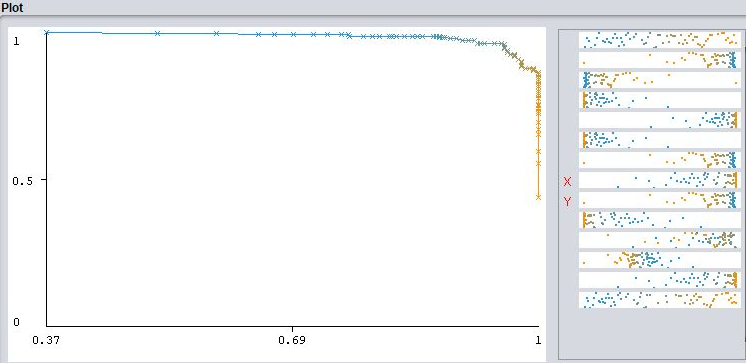
**Table 4: Performance measure – Decision Tree:**

|  |  |
| --- | --- |
| Evaluating Measure  Decision Tree | 10-fold Cross Validation |
| Accuracy | 92.970% |
| Precision | 0.930 |
| Recall | 0.930 |
| F measure | 0.930 |
| Error | 7.029% |

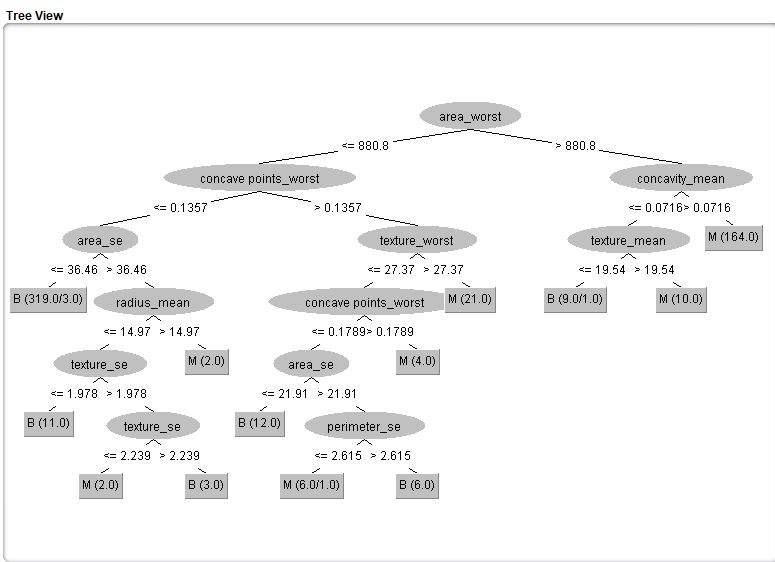
**ROC curve for Decision Tree:**



**Precision –Recall curve:**



**Decision Tree:**



# **X. EXPERIMENTAL RESULT**

Table5 Here, Nave Bayes classifier gave 92.61 %, Logistic Regression gave 94.024 %, Random Forest gave 95.957 % and Decision Tree gave 92.970 % accuracy. Random Forest gave the higher accuracy than others. But if we apply all these models at the same time with experiment and compare those results we can see naive bayes gives 93.042 % accuracy, logistic regression gives 95.028 % accuracy, random forest gives 96.329 % accuracy and decision tree gives 91.012 %. Therefore, we can say Random Forest gives the best performance for those Breast Cancer data.

**Accuracy Chart:**

## XI. CONCLUSION

This paper concludes with **Random Forest** Classifier gives the most balanced result with respect to accuracy of the Breast data while Decision tree gave the lowest accuracy. So **Random Forest** is proved to be perfect for this type of data. We can find the association between those attributes and tested negative class using random forest and also predict the risk accordingly. This model can be widely used.

XII.REFERENCES

***1. 1.J. Han, M. Kambar, "Data Mining: concepts and techniques" in , Morgan Kaufmann, 2006.***

***2. J. Ferlay, Globocan 2012 v1.0 Cancer Incidence and Mortality Worldwide: Iarc Cancerbase no. 11, 2014, [online] Available:*** [***http://globocan.iarc.fr***](http://globocan.iarc.fr)***.***

***3. D. Tomar, S. Agarwal, "A survey on Data Mining approaches for Healthcare",*International Journal of Bio-Science and Bio-Technology*, vol. 5, no. 5, pp. 241-266, 2013.***

***4. S. Demigha, "A Data Warehouse System to Help Assist Breast Cancer Screening in Diagnosis Education and Research", CSA the Second International Conference on Computer Science and its Applications, pp. 1-6, 2009.***

***5. D. Delen, G. Walker, A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods",*Artificial Intelligence in Medicine*, vol. 34, no. 2, pp. 113127, 2004.***