www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 5 May 2017, Page No. 21371-21377
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i5.31

A Machine Learning Approach for Early Prediction of Breast Cancer

Younus Ahmad Malla, Mohammad Ubaidullah Bokari
Student CS/IT, BFIT Dehradun,
Email:unismalla4@gmail.com
Chairman, Deptt. of Computer Science AMU Aligarh)
Email:mubokhari@gmail.com

Abstract
Nowadays by the rapid digitization of the data in the Healthcare sector has resulted in the collection of mountains amount of data in various Electronic Health Records (EHR). As the data is the biggest asset in the modern age, whose proper utilization in the Healthcare sector can lead to the discovery of the dreadful diseases very well in time which in turn will provide high quality of care to patients and at less expenditure. Breast Cancer is a primary cause of death in women whose precise detection of Breast Cancer is important in early stages. Precise results can be achieved through data mining algorithms. Developing a machine learning models that can help us in prediction the disease can play a vital role in early prediction. These Machine learning methods can be used to classify between healthy people & people with different disease. In the given project the light is been thrown on the same disease by using certain selected machine learning algorithms in WEKA tool and a corresponding evaluation of the selected Machine learning algorithms in terms of accuracy is also performed so as to select the best classifier for the early diagnosis of the said disease with better accuracy results In this paper three different types of models were implemented on the Breast Cancer dataset as Naïve Bayes, Logistic Regression and Random Forest. Out of the three Random Forest lead the top by having accuracy of 98% and sensitivity 99% followed by Logistic Regression with accuracy of 96% and sensitivity 98% and finally with Naive Bayes with accuracy of 91% and sensitivity 94%.

Keywords: UCI Repository, ARFF, Machine Learning, Data Mining

1. INTRODUCTION

Breast cancer is one of the most common dreadful diseases in the worldwide which if not diagnosed well in time leads to death. A large number of people in the world nowadays are the victims of the same in huge numbers. As per United States statistics for the year 2014 there will be an estimated 232,670 females and 2,360 males' new cases of breast cancer. Among them 40,000 females and 430 males was death [1] women are expected to die of the disease. India is also not lacking behind as being the victim of this dreadful disease. From last couple of years we are now witnessing more and more number of patients being diagnosed with breast cancer. For the years 2015, as per Statistics of Breast Cancer in India 1,55,000 new cases of breast cancer has been registered in which about 76000 women India[2] are expected to die of the disease.So here arises a need to make a shift from the traditional cure based systems to the prevention based which is possible only if we aggressively work for early detection and prediction of the same. Earlier the detection of Breast cancer, the better treatment will be prescribed to the patient.

To accurately and reliably diagnosis it fires the need from a paradigm shift from the traditional treatment based methods to more sophisticated event driven precision treatment.

Data mining is powerful tool to manage this task. Data mining [3] which is also known as knowledge discovery that extract and discover hidden features in data warehouse. The use of data mining widely used in prediction mainly in medical fields, which provides perfect data analysis tools to discover previously unknown, valid association in large datasets. Data mining is the integration of the set of tools and techniques from different fields like statistics, Mathematical algorithm and machine learning methods. The tremendous use of powerful computers with automated tools, storage and retrieval of large value of medical data are being collected and are being made available for medical research purpose.

The KDD[4] process comprise of a few steps leading from raw data collection to some form of new information an able to predict the outcome of a disease using historical cases stored with in data set.

- **Data cleaning**:-in this step the noise and inconsistent date are removed from the collection.
- **Data integration**:-in this step several sources of data are combined.
- **Data selection**:-in this step data relevant to the analysis task are retrieve from the data collection.
- **Data transformation**:- in this step, data is transformed into forms appropriate for the mining procedure.
- **Data mining**:-in this step intelligent techniques are applied to extract patterns.
- **Pattern evolution**:-in this step interesting data patterns are evaluated.
- **Knowledge representation**:- in this step discovery knowledge is represented.

We are using three different types of classification algorithms Naive Bayes, Logistic Regression and Random Forest along with 10 fold cross validation techniques to compare performance of these classification algorithms and show the best prediction classifier. The effect of certain Weka filters on the final output accuracy is also been carried out and presented in the form of graph. The rest of the paper is as follows: Section II gives a brief introduction about the Data mining algorithms which are used in carrying out this work. Then in the Section III the tool Weka is discussed, its functions and the compatible issues are discussed.

II. ALGORITHMS USED

In this paper three different types of models were implemented on the Breast Cancer dataset in order to throw a light on the corresponding performance of these machine learning techniques which are:

1. Naïve Bayes
2. Logistic Regression
3. Random Forest

- **Naïve Bayes**: Naïve Bayesian [5] classifier is primarily based on Bayes theorem with independence assumptions between predictors. A naïve Bayesian model is simple to construct, without a complex iterative parameter estimation which make it specially useful for extremely large datasets. Regardless of its simplicity, the naïve Bayesian classifier regularly dies relatively nicely and is broadly used because it regularly out performs greater sophisticated type methods.
- **Logistic Regression:** Logistic Regression [6] predicts the probability of an outcome that can only have two values (i.e., a dichotomy).the predictions is based on one or several predictors (numerical and categorical). Logistic regression is used to describe records and to explain the relationship between one or more independent variables that determine an outcome. It is measured with dichotomous variables (in which there are only two possible outcomes).
- **Random Forest:** Random Forest[7] is a machine learning Algorithm also called bagging algorithm based on decision trees. It is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees.

III. TOOLS AND TECHNIQUES

In this research WEKA [8] (The Waikato Environment for Knowledge Analysis) for running several algorithms has been chosen. Weka is a collection of machine learning algorithms for solving real world data mining issues. The algorithms can either be applied directly to a data set or called from your own java code. Weka contains modules for data preprocessing, classification, clustering and association rule extraction.

- **Data preprocessing:-** An essential step in the data mining process is data preprocessing. Data preprocessing techniques are used in various stages to remove errors from the input data so that it cannot effect in[09] experimental results. One of the demanding situations that face the knowledge discovery process in medical database is poor data quality. For this reason a great effort was laid in preparing information carefully to acquire accurate and correct consequences.
- Data mining stages:-Stage of data mining is break down into three phases. Every phase all the algorithms were implemented to examine the health datasets. The testing method adopted for this research was parentage split that train on a percentage of the dataset, cross validate on it and test on the remaining percentage. Sixty six percent (66%) of the health dataset which were randomly

selected was used to train the dataset using all the classifiers. The validation was carried out using ten folds of the training sets. The models were now applied to unseen or new dataset which was made up of thirty four percent (34%) of randomly selected records of the datasets. Thereafter interesting patterns representing knowledge were identified.

Performance matrix:

In this paper the overall performance measures which might be used for evaluation are accuracy, sensitivity and specificity. A confusion matrix[10] is a table that is used to measure the performance of classification model.

The following table displays a 2×2 confusion matrix for two classes (Positive and Negative)

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="2">Predicted</td><td></td><td></td></tr>
<tr><td>Negative</td><td>Positive</td><td>Positive Predictive value</td><td>a/(a+b)</td></tr>
<tr><td rowspan="2">Actual</td><td>Negative</td><td>A</td><td>B</td><td>Negative Predictive value</td><td>d/(c+d)</td></tr>
<tr><td>Positive</td><td>C</td><td>D</td><td colspan="2" rowspan="3">Accuracy=(a+d)/(a+b+c+d)</td></tr>
<tr><td colspan="2" rowspan="2"></td><td>Sensitivity</td><td>Specificity</td></tr>
<tr><td>a/(a+c)</td><td>d/(b+d)</td></tr>
</table>

Fig.1: Description of Confusion Matrix

- **Accuracy**: The proportion of the total number of predictions that were correct.
- **Positive predictive value**: The proportion of positive cases that were correctly identified.
- **Negative predictive value**: The proportion of negative cases that were correctly identified.
- **Sensitivity**: The proportion of actual positive cases which are correctly identified.
- **Specificity**: The proportion of actual negative cases which are correctly identified.

IV. EXPERIMENTAL RESULTS

In this research the selected data mining algorithms is been applied on the breast cancer data set. The dataset is obtained from UCI [11] repository which is an open source repository containing large amounts of data of different diseases**.** Large number of data sets is available for research use. The Breast cancer dataset contains 286 instances and 10 attributes. The features considered in this dataset consist of diverse components inclusive age, area of tumor, and so forth. All the attributes of the said dataset are of the type nominal whose description is given below in the table.

In this paper, the performance measures which are used for comparison are : accuracy, sensitivity and specificity. A distinguished confusion matrix is obtained to calculate the three measures. Confusion matrix is a matrix representation of the classification results. the upper left cell denotes the number of samples classifies as true while they were true (i.e., true positives), and lower right cell denotes the number of samples classified as false while they were actually false (i.e., true false). The other two cells (lower left cell and upper right cell) denote the number of samples misclassified. Specifically, the lower left cell denoting the number of samples classified as false while they actually were true (i.e., false negatives), and the upper right cell denoting the number of samples classified as true while they actually were false (i.e., false positives).Once the confusion matrixes were constructed, the accuracy, sensitivity and specificity are easily calculated as: sensitivity = TP/(TP. + FN); specificity =TN/(TN + FP). Accuracy = (TP + TN)/(TP + FP + TN + FN); where TP, TN, FP and FN denotes true positives, true negatives, false positives and false negatives, respectively. 10-fold cross validation is used here to minimize the bias produced by random

sampling of the training and test data samples. Extensive tests on numerous datasets, with different learning strategies, have shown that 10 is about the right number of folds to get the best estimate of error, and there is also some theoretical evidence that backs this up [12]. More Matrices include used are as:

- **Time**: This is referred to as the time required [12] to complete training or modeling of a dataset. It is represented in seconds
- **Kappa Statistic**: A measure of the degree of nonrandom agreement between observers or measurements of the same categorical variable.
- **Mean Absolute Error**: Mean absolute error is the average of the difference between predicted and the actual value in all test cases; it is the average prediction error.
- **Mean Squared Error**: Mean-squared error is one of the most commonly used measures of success for numeric prediction. This value is computed by taking the average of the squared differences between each computed value and its corresponding correct value. The mean-squared error is simply the square root of the mean- squared-error. The mean-squared error gives the error value the same dimensionality as the actual and predicted values**.**
- **Root relative squared error**: Relative squared error is the total squared error made relative to what the error would have been if the prediction had been the average of the absolute value. As with the root mean-squared error, the square root of the relative squared error is taken to give it the same dimensions as the predicted value.
- **Relative Absolute Error**: Relative Absolute Error is the total absolute error made relative to what the error would have been if the prediction simply had been the average of the actual values. Every model was evaluated based on the above measures discussed below.

The results were achieved using average value of 10 fold cross-validation for each algorithm. We found that in case of Breast Cancer Disease the Random forest achieved classification accuracy of 98.951% with a sensitivity of 74.40% and a specificity of 48.11% while as Naive Bayes achieved a classification accuracy of 91.2587% with a sensitivity of 95.27% and a specificity of 61.90**%** and Logistic Regression achieved a classification accuracy of 98.2527% with a sensitivity of 99.1% and a specificity of 94.11 . Table 3 shows the complete set of results in a tabular format. The detailed prediction results of the validation datasets are presented in form of confusion matrixes.

Attribute name	Description
Age	Patient age in year's
Menopause	The period in a women's life when menstruation ceases
Tumor-size	Patient's tumor-size in her breast
Inv-nodes	Node size in main portion of the breast
Node-caps	Node is present or not in cap of the breast
Deg-malig	Stage of breast cancer
Breast	Left breast or Right breast or both breast
Breast-quad	Portion of the breast for example left-up, left-low, right-up, right-low, central
Irradiat	Preset or not (YES/NO)
Class	No-recurrence-events, recurrence-events (Reduce the risk of breast cancer)

Table 1: Description of Breast Cancer Dataset

Performance matrix	Naïve Bayes	Logistic regression	Random Forest
Time	0	0.17	0.05
Kappa Statistics	0.704	0.9399	0.9628
MAE	0.1054	0.0175	0.948
RMSA	0.2475	0.1322	0.1616
RAE(%)	36.3333%	6.0259%	32.6616%
RRSE(%)	65.1687%	34.8108%	42.5449%
Accuracy=(TP+TN)/(TP+FP+TN+FN)	71.6783%	68.8811%	69.5804%
Sensitivity=TP/TP+FN	36.01%	74.03%	75.30%
Specificity=TN/TN+FP	70.07%	66.10%	48.11%

Table 2: Comparison of various parameters on Breast cancer data set before applying filter

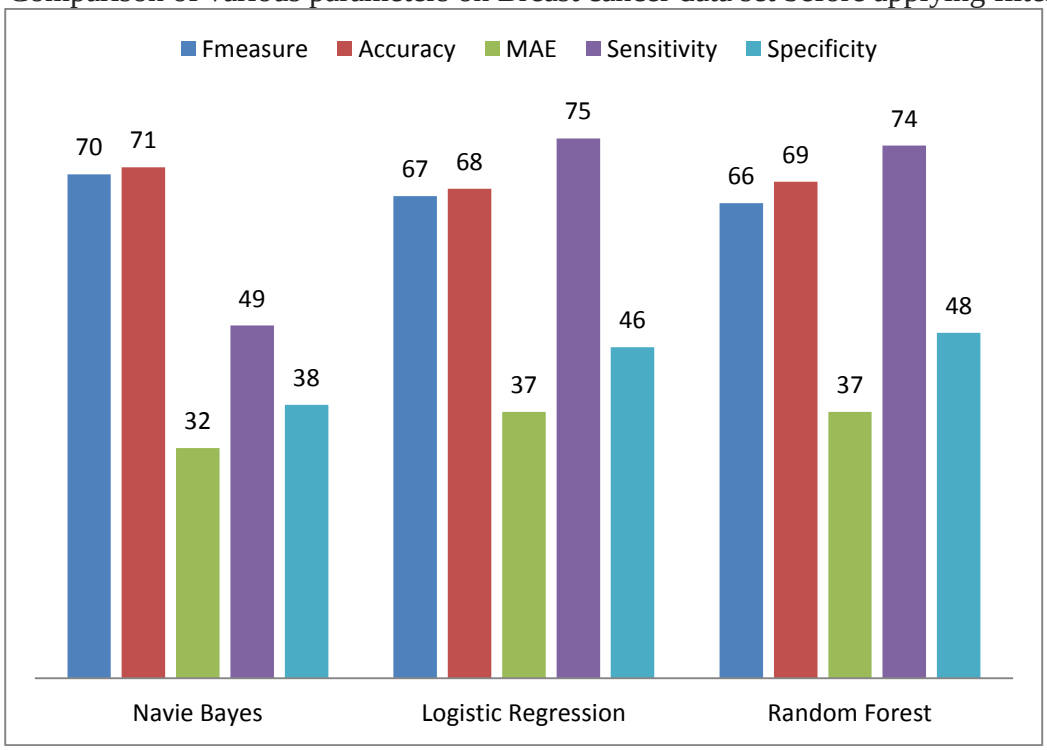

Fig.2: Comparison of various parameters on Breast cancer data set before applying filter

Performance matrix	Naïve Bayes	Logistic regression	Random Forest
Time	0	0.16	0.31
Kappa Statistics	0.2857	0.1979	0.1736
MAE	0.3272	0.37	0.3727
RMSA	0.4534	0.4631	0.4613

RAE(%)	**78.2086%**	**88.4196%**	**89.0857%**
RRSE%)	**99.1872%**	**101.309%**	**100.9171%**
Accuracy=(TP+TN)/(TP+FP+TN+FN)	**91.2587%**	**98.2527%**	**98.951%**
Sensitivity=TP/TP+FN	**95.27%**	**99.1%**	**74.40%**
Specificity=TN/TN+FP	**61.90%**	**94.11%**	**48.11%**

Table 3: Comparison of various parameters on Breast cancer data set after applying filter

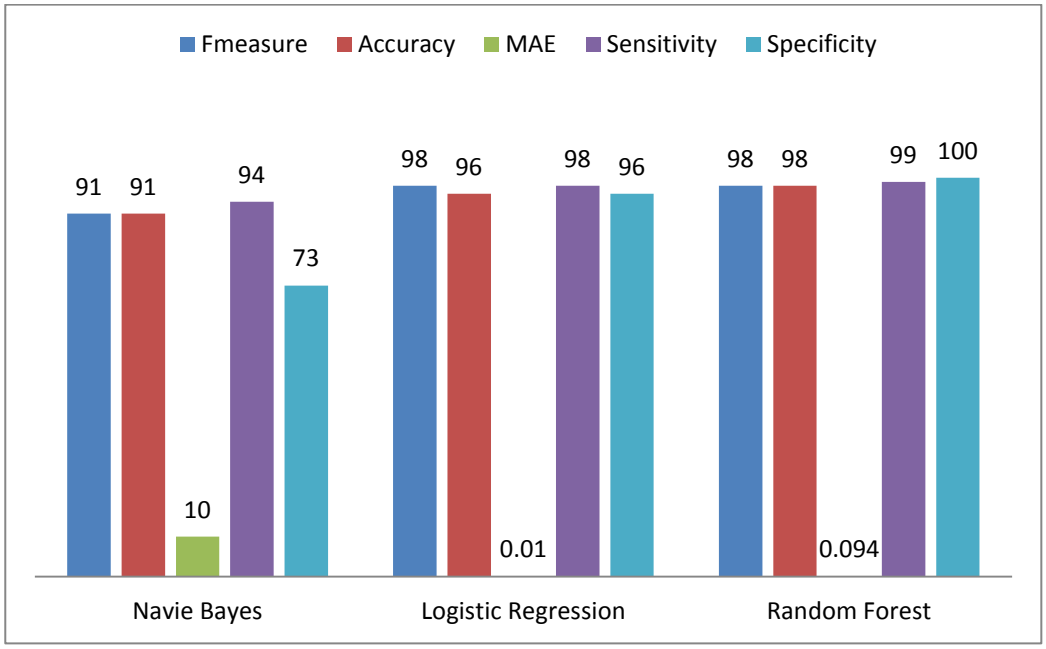

Fig. 3: Comparison of various parameters on Breast cancer data set after applying filter

V. CONCLUSION AND FUTURE DIRECTIONS

Breast Cancer which is second most common disease in the world wide. In this paper an analytical evaluation of certain selected machine learning algorithms is carried on the Breast Cancer dataset by using the open source tool WEKA. Some preprocessing is also done on the input dataset by applying certain WEKA in build filters and its overall effect on the accuracy of the prediction was also noted down. The results showed that the Random forest from the Decision Trees lead the best accuracy with filters by having accuracy of 69% before applying filter and 98% after applying filter. Similarly Logistic Regression got the second rank with 96% and without fffilters68% and finally it was Naive Bayes with 91% and without filters 71%In future different datasets can be analyzed also by using different combination of the data mining algorithms. Also a carefully feature selection can be used to extract the finest features which will increase the prediction accuracy in addition to increasing speed.

REFRENCES

1. J.Joshi and J.Patl, "Diagnosis and Prognosis Breast Cancer using Classification Rules", *International journal of Engineering Research and General Science*, Vol-2, Issue-,pp- 315-323,Nov 2014.
2. www.breastcancerindia.net. [Last visited 24-01-2017]

3. M. Durairaj and V. Ranjani, "Data Mining Applications In HealthCare Sector", *International Journal of Scientific and Technology Research*, Vol-2, Issue-, pp-29, 2013.
4. B.Srinivasan and K.Pavya, "A Study On Data Mining Prediction Techniques in HealthCare", *International Research Journal of Engineering and Technology*, Vol-3, Issue-03, pp-552-556,March 2016.
5. H.Ganesh and G.Annamary, "Comparative study of Data Mining Approaches for Parkinson's Diseases" *International Journal of Advanced Research in Computer Engineering and Technology*, Vol-3,Issue-9,pp-3062-3068,Sept.2014.
6. H.Yusuff, N.Mohamad, K.Nigh and S.Yahata "Breast Cancer Analysis Using Logistic Regression *"International Journal of Research And Applied Studies*, Vol-11,Issue-01 Jan 2012.
7. M.N. Ekgedawy "Prediction of Breast Cancer using Random Forest, Support Vector Machines and Naive Bayes", *International Journal of engineering and Computer Science* ,Vol-2, pp- 19885, 2017.
8. H.Written and E.Frank, "Data Mining: Practical Machine Learning Tools and Techniques" *Morgan Kaufmann publishers*, 1999. www.cs.waikato.ac.nz.
9. S.Kulkami and M. Bhagwat, "Predicting Breast Cancer Recurrence Using Data Mining Techniques*"* *International Journal of Computer Application*, Vol-122, ppt-26-31, July 2015.
10. K.Santtra,C.Joesphine and Christy, "Genetic Algorithm and Confusion Matrix for Document" *International Journal of Computers Science* Vol-9,Issue-1 pp-322-328,Jan 2012.
11. www.archieve.ics.uci.edu[Last visited, 18-02-2017].
12. T.A.Shaikh "A Prototype of Parkinson's and Primary Tumor Disease Prediction using Data Mining Techniques" *International Journal of Engineering Science Invention* Vol-3,Issue-4,pp-23-28,April 2014