INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
*(Volume 5, Issue 2)*
Available online at: www.ijariit.com

# Prognostication of breast cancer using data mining and machine learning

*Shriya D. Narvekar*
*shriyanarvekar563@gmail.com*
*VIVA Institute of Technology, Virar, Maharashtra*

*Ameya Patil*
*ameyapatil97@gmail.com*
*VIVA Institute of Technology, Virar, Maharashtra*

*Jagruti Patil*
*patiljagruti136@gmail.com*
*VIVA Institute of Technology, Virar, Maharashtra*

*Saniket Kudoo*
*Saniketkudoo@viva-technology.org*
*VIVA Institute of Technology, Virar, Maharashtra*

**ABSTRACT**

***In this study, WPBC that is Wisconsin Prognostic Breast Cancer (original) dataset to find an efficient predictor algorithm, to predict the recurring and non-recurring nature of breast cancer. Breast cancer is the most common disease found among the women, it is difficult for the physicians to know the exact reason behind breast cancer, and they need a smart system for predicting the illness on time before it is too late to be treated. It is one of the crucial reasons for death among females all over the world. The cancer tumor is generally categorized into benign and malignant tumors. Using machine learning and data mining techniques it can easily identify the cancer cells, provide benefits to the medical system. Use of classification algorithm C5.0 and XGBOOST sums up to have a considerably better result, improving the performance of the system. İn Classification, the calculation separates the information into unmistakable gatherings. It fundamentally pursues two stages of preparing first to learn and after that to a group.***

***Keywords***— *C5.0, Confusion matrix, F1 score, WPBC dataset, Xgboost, Breast cancer*

## 1. INTRODUCTION

Breast cancer is the most common cancer which is mostly found in women around the globe. Women around 140 countries of 184 worldwide are diagnosed with breast cancer. Cancer is a tumor that can cause death or can be deleterious to women health. Since 2008 there has been increasing of more than 20% in the number of breast cancer incidences worldwide. İt is the second leading cause of death amongst the women after lung cancer. Treatments of breast cancer basically include surgery, radiations, chemotherapy and hormone therapy based on the intensity of the disease.

Data Mining is a process of arranging informational indexes to distinguish examples and relationship to take care of the issue through information examination. It is a procedure to extricate usable information from a larger dataset. It investigates information designs in vast numbers by utilizing at least one programming. Data Mining is likewise known for the KDD procedure that is Knowledge Discovery in Data. It holds incredible potential for the human services industry which empowers wellbeing frameworks to systematically utilization of information.

Machine learning can be of two types supervised or unsupervised. In supervised learning, labelled information is given to the algorithm to classify the data. It makes the system learn without any explicit programs. The learning algorithms deduce the functions to predict the values and obtain results.

## 2. RELATED WORK

In [2], various techniques and review on breast cancer diagnosis and prognosis problems are examined. The prognostic problem is mainly analysed using C4.5 and its accuracy is higher in comparison to other classification techniques applied for the same.

In [4], C5 is a classifier which classifies the data in less time compare to another classifier. For generating a decision tree the memory usage is minimum and it also improves the accuracy. This paper developed a system on the bases of C5 algorithm which provides Feature selection, Cross-validation and reduced error pruning facilities.

 

In [1], highlights the performance of different clustering algorithms and classification algorithms on the dataset. The result shows that classification algorithms are a better predictor than clustering algorithms. The decision tree (C5.0) and SVM show the best predictor with 81% accuracy.

In [5], the experimental results in WBC dataset with a fusion between MLP and J48 (decision tree) is superior to the other classifier. The Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository is used, to distinguish malignant from benign tumors.

## 3. PROPOSED SYSTEM

The proposed system is intended to build a prediction of breast cancer which has to be carried out with very minimal time constraint which has been the significant drawbacks of most of the existing system. Wisconsin Prognostic Breast cancer (WPBC original) dataset, have 699 instances of 10 attributes. The proposed system shows that Xgboost gives a higher accuracy in accordance with the other algorithm.

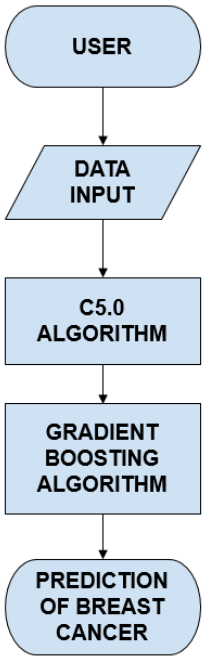


**Fig. 1: Proposed System Flow**

In figure 1. The system flow of the proposed system is represented. The user provides the data input including personal information, the data inputs include the tumour information. The Wisconsin Prognostic Breast Cancer (WPBC) dataset is considered in the proposed system. Where the data input is applied first with C5.0 and then with Xgboost.

### 3.1 Dataset

The Wisconsin breast cancer database is used to analyse, that have been collected by the UCI Machine Learning Repository. There are total of 699 records in this database. Each record in the database has nine attributes. The ten attributes detailed in Table 1 are graded on an interval scale from a normal state of 1–10, with 10 being the most abnormal state. The selected attributes are Uniformity of Cell Size, Mitoses, Clump thickness, Bare Nuclei, Normal Nucleoli, Single Epithelial cell size, Uniformity of Shape, Marginal adhesion, Bland Chromatin and Class.

### 3.2 Classification Algorithm

The classification algorithm divides the data into distinct groups. It basically follows two steps of processing first to learn and then to classify. In this study, the methods used are C5.0, Xgboost. This machine learning algorithms are applied to the WPBC dataset. The results obtained are compared with each other.

### 3.3 Confusion Matrix

The confusion matrix is nothing but a simple way to describe the performance of the classification algorithm. Table 2. Shows the generalized confusion matrix.

**Table 1: Confusion matrix (C5.0)**

| | **Predicted Positive** | **Predicted Negative** |
|---|---|---|
| Actual Positive | 113 | 5 |
| Actual Negative | 0 | 57 |

**Table 2: Confusion matrix (Xgboost)**

| | **Predicted Positive** | **Predicted Negative** |
|---|---|---|
| Actual Positive | 102 | 2 |
| Actual Negative | 2 | 69 |

 

*True Positive*: This means that the number of people who are sick and are categorized as sick according to the algorithm.
*False Positive*: This means that the number of people who are healthy and are categorized as sick according to the algorithm.
*False Negative*: This means that the number of people who are sick and are categorized as healthy according to the algorithm.
*True Negative:* This means that the number of people who are healthy and are categorized as healthy according to the algorithm.
*Precision*: It can be termed as predicting the actual positive value of the predicted value of the model.

$$Precision = \frac{True\ Positive}{Total\ Predicted\ Positive} \quad (1)$$

*Recall:* It can be termed as predicting the positive value in the model from the labelled positive value.

$$Recall = \frac{True\ Positive}{Total\ Actual\ Positive} \quad (2)$$

*F1 score:* It can be termed as a measure to create a balance between the uneven class distribution and the Precision and Recall.

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (3)$$

Table 3 shows the report for Xgboost and Table 4. Shows the report for C5.0.

**Table 3: Report for Xgboost**

| | Precision | Recall | F1 score |
|---|---|---|---|
| Benign | 0.96 | 0.98 | 0.97 |
| Malignant | 0.97 | 0.92 | 0.95 |

**Table 4: Report for C5.0**

| | Precision | Recall | F1 score |
|---|---|---|---|
| Benign | 0.96 | 0.94 | 0.95 |
| Malignant | 0.89 | 0.93 | 0.91 |

## 4. EXPERIMENTAL RESULTS

The application is implemented in the Python Programming language. The dataset is been divide into two set that is a training set of (524, 11) and testing set of (175, 11). The results obtained are in terms of the percentage of the accuracy. The Xgboost gives comparatively higher accuracy than the C5.0 algorithm. The obtained accuracy of C5.0 is 97.14% and Xgboost is 97.71%. İt was successfully classified whether the given tumour is benign or malignant. Figure. 2. Shows the comparative analysis of C5.0, Xgboost and Logistic Regression.

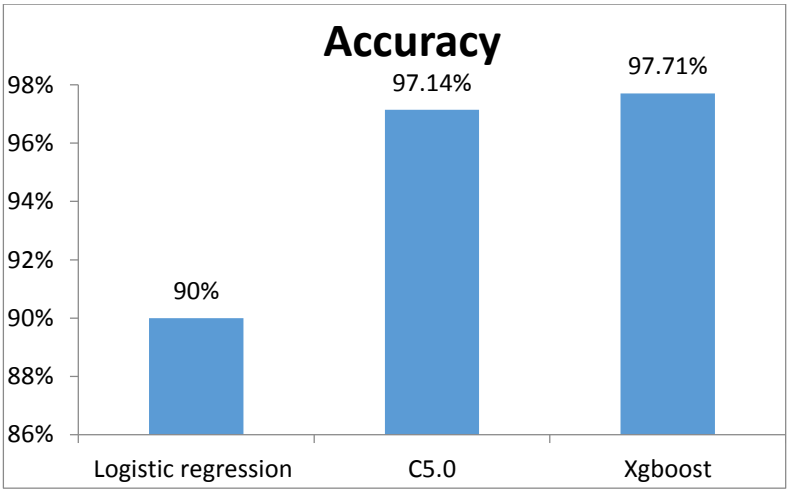


**Fig. 2: Comparison of Logistic Regression, C5.0, Xgboost**

## 4. CONCLUSION

Various data mining techniques are used to predict and identify breast cancer in patients. In this paper, we have used two classification methods for prediction. We have focused on the accuracy and the lowest computing time. The experimental results in WPBC dataset show that the Xgboost proves to give a comparative better accuracy than the other machine learning algorithms. An overall accuracy from the system was found out to be 97.71%.

## 5. REFERENCES

[1] U.Ojha, Dr S.Goel, "A Study On Prediction Of Breast Cancer Recurrence Using Data Mining Techniques", IEEE,2017
[2] B.Padmapriya, T.Velmurugan, A.Dsouza, N.Kazi", Survey on Breast Cancer Analysis Using Data Mining Techniques", IEEE

 

[3] K.Arutchelvan, Dr.R.Periyasamy, "Cancer Prediction System Using Data Mining Techniques", International Research Journal of Engineering and Technology,2014
[4] R.Pandya, J.Pandya, "C5.0 Algorithm To Improved Decision Tree With Feature Selection And Reduced Error Pruning', International Journal of Computer Applications,2015
[5] G. I. Salama, M.B.Abdelhalim, MagdyAbd-elghanyZeid, "Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers", International Journal of Computer and Information Technology
[6] R. Raj Bharath, L. Sankari, "Predicting Breast Cancer using Random Forest and Logistic Regression", IJESC,2017
[7] M.Rana, P. Chandorkar, A.Dsouza, N.Kazi, "Breast Cancer Diagnosis And Recurrence Prediction Using Machine Learning Techniques", International Journal of Research in Engineering and Technology
[8] M.Azmi, Zaihism, "Breast Cancer Prediction Based On Backpropagation Algorithm", IEEE
[9] Mr C.Shah, Dr.A. Jivani, "Comparison of Data Mining Classification Algorithms for Breast Cancer Prediction", IEEE
[10] Siyabend Turgut, Mustafa Dagtekin, Tolga Ensari," Microarray Breast Cancer Data Classification Using Machine Learning Methods", IEEE
[11] https://www.expertsystem.com/machine-learning-definition, Last Accessed on 27th Feb 2019

 