International Journal of Scientific Research and Engineering Development-– Volume 3 Issue 2, Mar-Apr 2020

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

# Breast Cancer Prediction Using Data Mining

Iffat Khan*, Ankita Gandhi**, Nandan Parmar***,Bindu Garg****

(*Computer Engineering, Bharati Vidyapeeth College Of Engineering, Pune India*

*Email: iffatk07@gmail.com

** Email: agandhi082@gmail.com

*** Email: parmar.nandan189@gmail.com

**** Email: brgarg@bvucoep.edu.in)

***Abstract* — Data mining consists of core algorithms that enable to gain fundamental insights and knowledge from large datasets and is also a part of larger knowledge discovery process. Breast cancer is that serious threat that is the second leading reason for deaths of women. Early detection and diagnosis of Breast Cancer plays an important role in the well being of women. Food habits, environmental pollution, hectic lifestyle and genetics are common factors that attribute to breast cancer. Number of studies have been undertaken in order to understand the prediction of breast cancer risk using data mining techniques. Hence, the goal of this research is focused on using two data mining techniques to predict breast cancer risks in women.**

***Keywords* — Accuracy, Benign, Breast Cancer, Classification, Confusion Matrix, Logistic Regression , Malign, Prediction, Random Forest.**

## I. INTRODUCTION

After skin cancer, breast cancer [1] is the most common cancer that is diagnosed in women . Breast cancer can develop in both men as well as women [2], but it is very common in women [3]. Breast cancer mainly affects the physical and the mental health of women. There are a number of factors that promote the cause of breast cancer. The early diagnosis of Breast Cancer can help in forecast and chance of survival significantly, as it can aid timely clinical treatment to the patients.

Data mining [4] is a field that is used to find interesting patterns, unknown relationships from large datasets. It is a powerful field having various techniques for the analysis of real world problems. It can convert the data that is raw into useful information in various research fields and finds the important patterns to predict future trends in medical field. In data mining, prediction is about identifying the data points based on the description of another related data value [5]. In this research we are using the logistic regression and random forest [6] for prediction of breast cancer. Classification [7] is a data mining function and its goal is to precisely predict the target class for each case in the data. Classification and data mining methods are an effective way to classify data. Especially in medical field, where those methods are widely used in diagnosis and analysis to make the decisions.

The data set used in this research is publicly available and was created by Dr. William H. Wolberg [8], physician at the University of Wisconsin Hospital at Madison, Wisconsin, USA.

The aim of this research is to observe which characteristics are most helpful in predicting cancerous and non cancerous and to see the general trends that can aid us in model selection and hyper parameter selection. The basic goal is to classify whether the breast cancer is dangerous or non dangerous.

## II. RELATED WORK

The Wisconsin University's breast cancer database was analyzed by using Naïve Bayes prediction algorithm and classification algorithm. So that various other algorithms can be used to predict and classify whether the tumor is harmful or not [9]. So the datasets were chosen randomly. So at the final Naïve Bayes classification algorithm was 10-15% was incorrectly classified and 85-95% was correctly classified.

Two different datasets from Wisconsin Breast Cancer have been evaluated by different data mining algorithms. The outcome was that Rotation Forest Model shows the highest accuracy (99.48%) and when it is compared with previous works, the new approach and methodologies have come with much more high performance as well as accuracy [10].

D S Jacob et al [2018] [11], gave a survey of breast cancer prediction using the data mining technology. It shows that the classification algorithm performs better compared to the clustering algorithms in predicting breast cancer.

Liu et al.[12] used decision table based predictive models for breast cancer survivability, concluding that the survival rate of the patients was 86.52%.

They have used the under- sampling C5 technique and bagging algorithm to take care of disparity problem, thus it is improving the predictive performance of breast cancer.

  

## III. PROPOSED RESEARCH

The Cancer in breasts starts to grow in the body when the cells present in the breast grow in the most unexpected manner. After the cells grow they can be observed with the help of X- Ray. In this research we have used the logistic regression and random forest for the purpose of prediction of breast cancer. Both of these are classification algorithms of Supervised Learning because in our dataset we have the outcome classes i.e Y has only two set of values either M (Malign) or B (Benign).

We used Jupiter notebook to work on this dataset. We observed that the dataset contained 569 rows and 32 columns. We used the column 'Diagnosis' to predict if the cancer is M or B. 1 means that the cancer is M and 0 means that the cancer is B. We identified that out of 569 people, 357 are B and 212 are labeled as M.

We then found out if there were any missing values or null values followed by splitting the data into testing and training data. Then for Feature Scaling we used the Standard Scaler method. After this we applied both of the algorithms.

### *1. Logistic Regression*

Logistic regression is used in biological sciences since the early twentieth century. Later it was used in many social science applications. Logistic Regression is used when the Dependent variable i.e the target is categorical. Thus it satisfies the condition of our dataset.

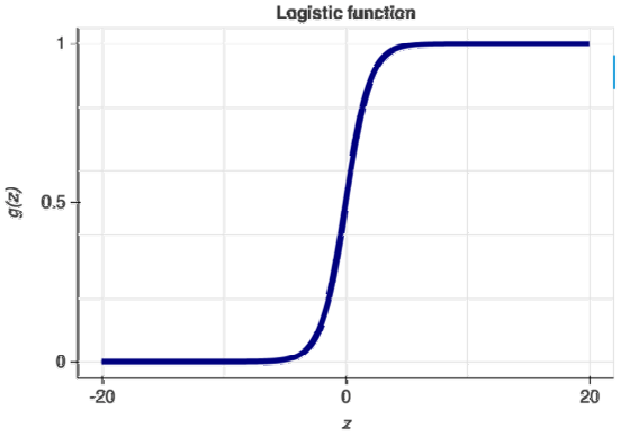


Fig 1: Logistic Regression

### *2. Random Forest*

Random Forest algorithm is a collections of various decision trees which come together to build a forest that is known as Random Forest. In this algorithm each of the node is split used the best node among other random nodes. It is not affected by any missing values as well as noise present in the input dataset. The stability of random forest is better as compared to the single decision tree and can handle the data minorities very well which makes it more efficient.

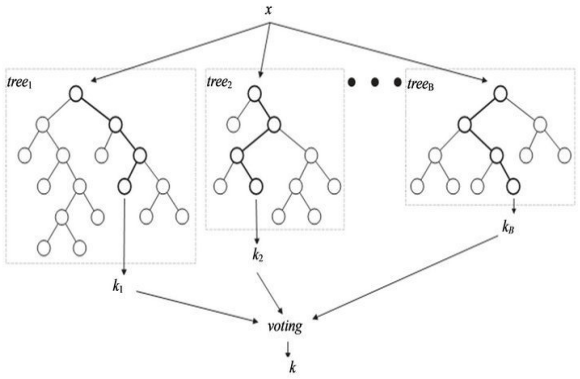


Fig 2: Random Forest

Using the above algorithm we classified our Breast Cancer Data into Malign or Benign. The accuracy achieved using the Logistic Regression algorithm is 95.8% and using Random Forest algorithm it is 98.6%.

To check the accuracy we imported the confusion_matrix method of the metrics class. We used the Classification Accuracy method to find the accuracy of the above mentioned models. Classification Accuracy is the ratio of number of correct predictions to the total number of input samples.

So finally we built our model for classification using the Random Forest Classification algorithm as it gives best results for the dataset we used

## IV. CONCLUSIONS

In this study we applied two different data mining classification techniques for the prediction of breast cancer risk and their performance and accuracy were compared in order to evaluate which algorithm classifies the dataset best. By looking at the comparison done we concluded that Random Forest algorithm classifies the dataset better than the Logistic Regression algorithm. Hence, an effective and efficient classifier for breast cancer detection has been identified while a number of attributes covered by the classification algorithm can be increased. This can be done by increasing the sample size of the training set used to train the classifier. Thus, resulting into more accurate model.

## REFERENCES

[1] https://www.cancer.gov/

[2] "Destroying the Cancer Threat". America's Biopharmaceutical Companies [Blog Post] Retrieved from https://innovation.org/diseases/immunologic/immunotherapy/destroying-cancer-threat

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

  

[4] Han J., Kamber M., "Data Mining Concepts and Techniques". Morgan Kaufman Publishers, 2001.

[5] Williams, Kehinde & Idowu, Peter & Balogun, Jeremiah & Oluwaranti, Adeniran. (2015). Breast cancer risk prediction using data mining classification techniques.Transactions on Networks and Communications. 3. 10.14738/tnc.32.662.

[6] Cuong Nguyen, Yong Wang, Ha Nam Nguyen "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic". J. Biomedical Science and Engineering, 2013, 6, 551-560

[7] Gupta S, Kumar D, Sharma A. "Data Mining Classification Techniques Applied For Breast Cancer Diagnosis and Prognosis". Indian Journal of Computer Science and Engineering. 2 (2011).

[8] W.N. Street, W.H. Wolberg and O.L. Mangasarian. "Nuclear feature extraction for breast tumor diagnosis". IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, volume 1905, pages 861-870, San Jose, CA, 1993.

[9] G. D. Rashmi, A. Lekha, and N. Bawane,"Analysis of efficiency of classification and Prediction algorithms (Naïve Bayes) for Breast Cancer dataset,"2015 Int.Conf.Emerg. Res. Electron. Comput. Sci. Technol., pp. 108–113, 2015.

[10] E. Aličković and A. Subasi, "Breast cancer diagnosis using GA feature selection and Rotation Forest." Neural Comput. Appl., vol. 28, no. 4, pp. 753–763, 2017.

[11] D S Jacob et al, "A Survey on Breast Cancer Prediction Using Data Mining Techniques", IEEE Conference on Emerging Devices and Smart System, pp.256-258 (2018).

[12] Liu, Y-Q, Wang, C, Zhang, L. "Decision tree based predictive models for breast cancer survivability on imbalanced data". In: 3rd international conference on bioinformatics and biomedical engineering, 11-13 June 2009, Beijing, China, 2009

  