Breast Cancer Prediction using Machine Learning

Sivapriya J, Aravind Kumar V, Siddarth Sai S, Sriram S

Abstract: One of the most dreadful disease is breast cancer and it has a potential cause for death in women. Every year, death rate increases drastically due to breast cancer. An effective way to classify data is through classification or data mining. This becomes very handy, especially in the medical field where diagnosis and analysis are done through these techniques. Wisconsin Breast cancer dataset is used to perform a comparison between SVM, Logistic Regression, Naïve Bayes and Random Forest. Evaluating the correctness in classifying data based on accuracy and time consumption is used to determine the efficiency of the algorithms, which is the main objective. Based on the result of performed experiments, the Random Forest algorithm shows the highest accuracy (99.76%) with the least error rate. ANACONDA Data Science Platform is used to execute all the experiments in a simulated environment.

Keywords: Accuracy, Algorithm, Breast cancer

I. INTRODUCTION

Being the most frequently occurring cancer in women, breast cancer affects around 10% of women at some point in their life. It is the second leading contributor to women's death after lung cancer. 25% of all cancers in women including 12% of all new cases are caused by breast cancer [9]. Big Data has seen a rise in value due to it being used in derivation of business intelligence, business analytics and data mining to obtain reports and result predictions. Topics like medial science rise rapidly when certain approaches like data mining is applied due to better possibility of prediction of diseases, reducing medicine costs, improving health of patient by revamping the quality of healthcare along with value by saving people's lives through real time decisions. The paper provides you with a analysis of performance and comparison of accuracy in classification between the algorithms such as: Logistic Regression, SVM, Random Forest and Naïve Bayes, being the major influential algorithms of data mining used in the research community. Logistic Regression is used to perform regression analysis when the dependent variable is binary. It is a predictive analysis similar to all other regression analyses. Naïve Bayes is a powerful classification algorithm from machine learning, it is not a algorithm but a group of algorithm in which all of them have same fundamental principle. In that group of algorithms, they all classify the data independently such that no algorithm provides same classification result or analysis. Naïve Bayes can be seen using in places such as document classification.

Document classification is nothing but it is used to classify document based on fields Random forest algorithm basically creates multiple trees and select the best among them by voting methodology. Random Forest algorithm creates multiple tree, each tree having different decision independent to each other. In common words no two tree is similar in Random Forest such that every tree provides implies classification methodology. This can be useful in certain instances like these where we have to classify medical dataset. Random forest is also used in places such as ETM, detection and prediction of an object and in games where it replicates the actions produced by humans in games. SVM is classification and regression model. It can be used to both circumstances. SVM creates a hyperplane which is basically called as threshold limit to classify data. This limit is created by the dataset while training. What if there is an deviation in data, SVM creates soft threshold limit which is close near to the main threshold limit and near to the deviation. The more dataset given to the algorithm the proper the classification will be of hyperplane. The kernel used in this project is linear. The reason to use linear kernel is it is faster and it is preferred when the data can be linearly separable. Places where you can see SVM used is text classification, face detection etc. Face detection is one of the main reasons SVM is preferred widely. Algorithm's efficiency evaluation is the primary objective of this project.

II. RELATED WORK

In machine learning and data mining, classification should be a crucial task. Researchers have already done lot of researches by applying machine learning algorithm on medical dataset for classification and data mining algorithm to find a pattern in dataset for faster calculation and prediction. Many of the approaches provide good accuracy and result.

[1] In their paper, they have implemented algorithms like C4.5, ANN, SVM to find classification accuracy in breast cancer dataset. Their research shows SVM had produced higher accuracy in classification.

[2] Their research is about finding classification accuracy using machine learning algorithm known as k-Nearest Neighbor with different values of k. For each value of k they have received a different result.

[3] Their paper is about using powerful machine learning classification algorithm Naïve Bayes, C4.5 which is usually used in data mining and ANN a neural network algorithm for the tumour classification of breast cancer in dataset.

Manuscript published on November 30, 2019.

* Correspondence Author
Sivapriya J M.E*- Assistant Professor in CSE at SRMIST, Ramapuram, Chennai.
Sriram S- Student at SRMIST, Ramapuram, Chennai
Aravind Kumar V- Student at SRMIST, Ramapuram, Chennai
Siddarth Sai S- Student at SRMIST, Ramapuram, Chennai

Retrieval Number: D8292118419/2019©BEIESP
DOI:10.35940/ijrte.D8292.118419
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Their work shows C4.5 did a better job in classification.

[4] Random forest classifier was implemented in their project to find sensitivity, time consumed and mean accuracy of two data set WBCPD and WBCDD.

[5] In their paper, Naïve Bayes algorithm was implemented to test the classification accuracy of breast cancer dataset with specificity, sensitivity and mean accuracy.

[6] Two models namely Logistic Regression and ANN was implemented. They were used to compare prediction accuracy breast cancer in mammography. Their study says logistic regression performed well in prediction.

[7] Adaboost algorithm was used to predict the cause and effect of breast cancer and the reason for death. Modest Adaboost algorithm was used.

[8]Performance criterion of classifiers is compared by Vikas Chaurasia and Saurabh Pal for SVM with the RBF kernel, naïve bayes, rbf kernel in neural networks, simple cart and algorithm in decision trees in breast cancer dataset to find the best classifier. Their experimental results say, SVM-RBF kernel produces an accuracy of 96.84% which is higher than other classifiers.

The performance and efficiency of the algorithms such as SVM, Random Forest, Logistic Regression and Naïve bayes were compared to the similar works mentioned above. The goal is to achieve the lowest error rate and best accuracy in analysing data. The performance and efficiency of these approaches are compared using: accuracy and time to build model. Random Forest scores highest classification accuracy (99.76%) and least error rate. Unlike the other classifiers which we have chosen for this research has classification accuracy in the range of 94% and 99%.

III. EXPERIMENT

To compare the behaviors of LR, NB, SVM and Random Forest, the experiment conducted was focused on the evaluation of the algorithms. Questions raised from researchers were: Which algorithm is more effective? Which algorithm executes more efficiently? Which algorithm is more accurate in classifying?

A. Experiment Environment

The sickit learn python libraries were used to conduct all experiments on classifiers explained in this paper. Sklearn is a collection of data mining, machine learning and deep learning algorithms used for classification, regression, data pre-processing and clustering. The sklearn libraries were used to implement machine learning algorithms for various real-world problems. Developers and practitioners can build and evaluate suitable models with this framework. The experiment conducted in which environment is conducted is ANACONDA. It contains various applications in that we have preferred Spyder, which is a development environment that supports python. It is a powerful IDE for python compared to others. It also has introspection features. Since our problem might require those features and also debugging is easier in this platform it is preferred.

B. Breast cancer dataset

The UCI machine learning repository consists of The Wisconsin Breast Cancer datasets [10] which is used in this study. There are 699 instances in which 458 are benign and 241 are malignant. In addition, there are two classes malignant which contributes to 65.5% of dataset and benign 34.5%. The breast cancer dataset is obtained in a csv format from their database.

C. Data Visualization

Data visualization is a key aspect of data science. It helps one to comprehend and also convey the data to another person in a meaningful manner. Matplotlib and Seaborn are some of the several python data visualization libraries. It is essential in analysing large amounts of information and to make decisions. It employs the use of pictorial elements such as maps, plots, patterns, graph trends, etc. to provide the user with an easy method of comprehending the data.

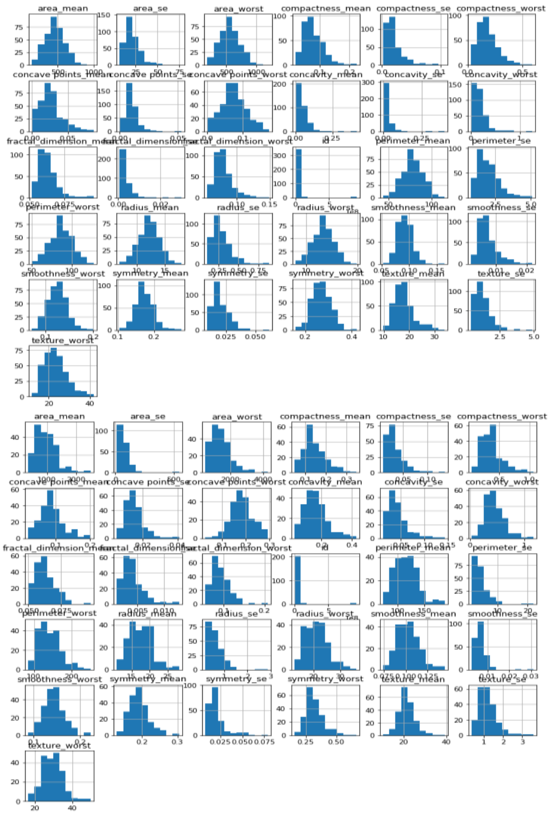

Fig 1: Data Visualization

IV. EXPERIMENTAL RESULTS

After creating predictive model, efficiency can be checked. For this, the models can be compared based on accuracy and time consumed. It was really hard to choose the algorithm which has higher performance, greater accuracy and efficiency, since all of them ended very close in accuracy. The time consumed and accuracy value of the algorithms from machine learning is shown in Table 1.

Table 1: Experimental Results

Algorithm	Accuracy	Time taken
LR	99.06%	0.02s
SVM	98.59%	0.03s
NB	94.83%	0.01s
Random Forest	99.76%	0.02s

V. RESULTS AND DISCUSSION

From Table 1 we can clearly notice that Random Forest requires 0.02s to build model unlike Naïve Bayes which requires just 0.01s. The accuracy of SVM is 97.41% and time taken by it is 0.03s which is higher than other models.

We can say that the performance of Random Forest (99.76%) is better when it comes to classification with comparison to the other algorithm's accuracy obtained. Other algorithms have accuracy that varies between 94% and 99%.

This means that the Random Forest algorithm portrays the highest correctly classified instance value and the lowest incorrectly classified instance value in comparison to the other classifiers. In summary, Random Forest was able to show its efficiency on the basis of time and accuracy. Random forest performs better when it comes to classification because it creates multiple trees with each tree providing different result. These trees are analysed and trained by itself with dataset and the best tree among the trees is been selected by voting methodology which is the tree with highest vote or tree which classifies most number of instances and chosen for the type of problem or situation.

VI. CONCLUSION

Medical dataset can not only be classified with the previously mentioned algorithms from machine learning, there are many algorithms and techniques which may perform better than these. Production of accurate classifier which perform efficiently for medicinal application is the main challenge we face in machine learning. Four main algorithms were implemented in this study were NB, SVM, Random Forest and LR on Breast Cancer dataset. Our main aim for the research is to discover the algorithm which performs faster, accurate and efficiently. Random Forest surpasses all the other algorithms with an accuracy of 99.76%. In conclusion, the Random Forest algorithm achieves the lowest error rate along with highest precision which might be the best choice of algorithm for this problem and prediction of disease.

REFERENCES

1. Ahmad LG, Eshlaghy AT, Poorebrahimi A, Ebrahimi M and Razavi AR. Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. Journal of Health & Medical Informatics
2. Seyyid Ahmed Medjahed, Tamazouzt Ait Saadi, Abdelkader Benyettou. Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules. International Journal of Computer Applications (0975 - 8887)
3. Abdelghani Bellaachia, Erhan Guven Predicting Breast Cancer Survivability Using Data Mining Techniques. 2006 SIAM Conference on Data Mining
4. 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
5. Diana Dumitru. Prediction of recurrent events in breast cancer using the Naive Bayesian classification. 2000 Mathematics Subject Classification.
6. Turgay Ayer, MS; JagpreetChhatwal, PhD; OguzhanAlagoz, PhD; Charles E. Kahn, Jr, MD, MS; Ryan W. Woods, MD, MPH; Elizabeth S. Burnside, MD, MPH, MS. Comparison of Logistic Regression and Artificial Neural Network Models in Breast Cancer Risk Estimation. RadioGraphics 2010
7. JarceThongkam, Guandong Xu, Yanchun Zhang and Fuchun Huang. Breast Cancer Survivability via Adaboost Algorithm. HDKM '08 Proceedings of the second Australasian workshop on Health data and knowledge management
8. Rasool Fakoor, Faisal Ladhak, Azade Nazi, Manfred Huber. Using deep learning to enhance cancer diagnosis and classification. 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013
9. Cancer Statistics, 2016. CA: A Cancer Journal for Clinicians
10. UCI Machine Learning Repository: Breast Cancer Wisconsin Dataset. link: https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/

AUTHOR'S PROFILE

Sivapriya J M.E- Assistant Professor in CSE at SRMIST, Ramapuram, Chennai. M.E in SSN College of Engineering
Mail: sivapriya@srmist.edu.in

Sriram S- Student at SRMIST, Ramapuram, Chennai pursuing B. Tech in CSE Department, III Year 2017-2021. Completed High School at Velammal Matric. Hr. Sec School.
Mail: sriram_sridhar17@srmuniv.edu.in

Aravind Kumar V- Student at SRMIST, Ramapuram, Chennai pursuing B. Tech in CSE Department, III Year 2017-2021. Completed High School at Jaigopal Garodia Matric. Hr. Sec. School.
Mail: aravindkumar_v17@srmuniv.edu.in

Siddarth Sai S- Student at SRMIST, Ramapuram, Chennai pursuing B. Tech in CSE Department, III Year 2017-2021. Completed High School at Chinmaya Vidyalaya Hr. Sec. School.

Mail: siddarthsai_s17@srmuniv.edu.in

Retrieval Number: D8292118419/2019©BEIESP
DOI:10.35940/ijrte.D8292.118419
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication