Predicting Breast Cancer using Modern Data Science Methodology

Vinoothna Manohar Botcha, Bhanu Prakash Kolla

Abstract*: Breast Cancer is the mass occurring cancer in women according to the World Health Organization(WHO), But the early prediction of breast cancer helps in the recovery for the effected one's. Reasons for breast cancer were Hormone replacement therapy or getting explore to harmful radioactive rays and due to late childbearing. The aim is to diagnose cancer by using a machine learning technique, Random Forest, for accurate solutions. The dataset we used is the Wisconsin Breast Cancer dataset. The output which the error rate was only about "0.0177".

Keywords* : Breast Cancer Prediction. Machine Learning, Data Science, Random Forest.

I. INTRODUCTION

In women, Breast Cancer is the most frequent cause of death. Breast Cancer occurred as cells began to grow out of control, which leads to the emergence of a tumor that can be discerned on an X-ray or felt as a lump. Nearly 21% of women had breast cancer, in which 16% were women for more than 50 years. This cancer develops from the breast tissue. There are mainly two types of classifications in breast cancer that are Benign tumor and Malignant Tumor. Benign tumors are not part of cancer it can be seen anywhere in the body and removes by proper medication and treatment, but Malignant tumors are cancerous, they grow abnormally out of control and can spread to other organs. The typical symptom was the lump on the breast with the chance in breast skin into a reddish color. Machine Learning helps in analyzing the data and helps to extract the information and characteristics from the given data. In cancer, machine learning techniques help in early diagnostics and prognosis of cancer [1]. There are many ML techniques to predict breast cancer like Decision trees, and Naive based Bagging Trees, Random Forest, and many more.

In this paper, we'll be using Random forest for the prediction since in decision tree have the majority of variance when utilization of different training and test sets of same data which leads to overfitting of data and the performance reduced. Whereas in bagging tree it occupies entire feature space by creating splits in the tree, which leads to decreasing variance and increasing bias. However, in the random forest, the correlation issue that occurred in bagging trees reduced along with that it prunes the tree by and de-correlated the tree by setting stopping criteria.

Revised Manuscript Received on August 20, 2019.

Vinoothna Manohar Botcha, (Correspondence Author) Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram (A.P), India.

Bhanu Prakash Kolla, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram (A.P), India.

II. REVIEW CRITERIA

Numerous ML algorithms used for the prediction of breast cancer. In those here are some research reviews. The author Moh'd Rasoul achieved 93.7% accuracy by using DWT tool [2]. The author Ashwaq Qasem achieved 95% accuracy by using marker Controller marker shed [3], and the author Junaid Ahmed achieved 84.21% accuracy by using Adaptive Reasoning Theory [4]. In this, the Wisconsin Breast Cancer data set was used, which acquired from the Machine Learning Repository of UCI, which contains 569 rows of data, and contains 32 attributes [5].

Random Forest R.F extended from Breiman's Bagging [6]. Random Forest based on the trees which group each tree based on a group of random variables. R.F is also the collection of multiple decision trees.

Fernandez-Delgado [7] used 179 ML algorithms on around 121 UCI datasets, and random forest ranked first in all.

III. RESULT AND DISCUSSION

In this paper, we used the Random forest technique for prediction of breast cancer and found that the test error rate is "0.0177". Here we use both GINI Impurity and Entropy, which gives insight to essential variables used for training.

$$\text{GINI Impurity} = 1 - \sum_{r} Ar$$

$$\text{Entropy} = \sum_{r} -Ar * \log Ar$$

Random Forest is the widely used ML technique used for predicting the diagnosis of breast cancer. The training data and the testing data is divided in 7:3 fashion, where the data can train so that the prediction can be made accessible. The confusion matrix for the given data can be seen below in table1.

Table 1: Confusion matrix

Class	B	M	class.error
B	279	7	0.02447552
M	14	156	0.08235294

He bringing out for diagnosis is essential as it brings the class imbalance within ML[8-14]. This occurs when the data is outnumbered by other classes, which leads to deceptive of accuracy. Hence the target class should not be imbalances. Many techniques in Machine Learning requires the preprocessing of the data like neural networks, and Neural networks require the preprocessing of the data for better performance. However,

the random forest doesn't require any preprocessing. We are using Hyperparameter Optimization in this, which helps in creating to create models allows us to do a grid search. Here we used mtry, ntree, nodesize for best accuracy[15-18].

In R.F, not all the attributes are equally used in amount for classification and prediction[19-21]. Here's the visual representation of essential attributes the dataset which can be seen in the Figure1. This graph contains all the essential attributes that were used for predicting, And the scores of each attribute that used are also mentioned in table 2

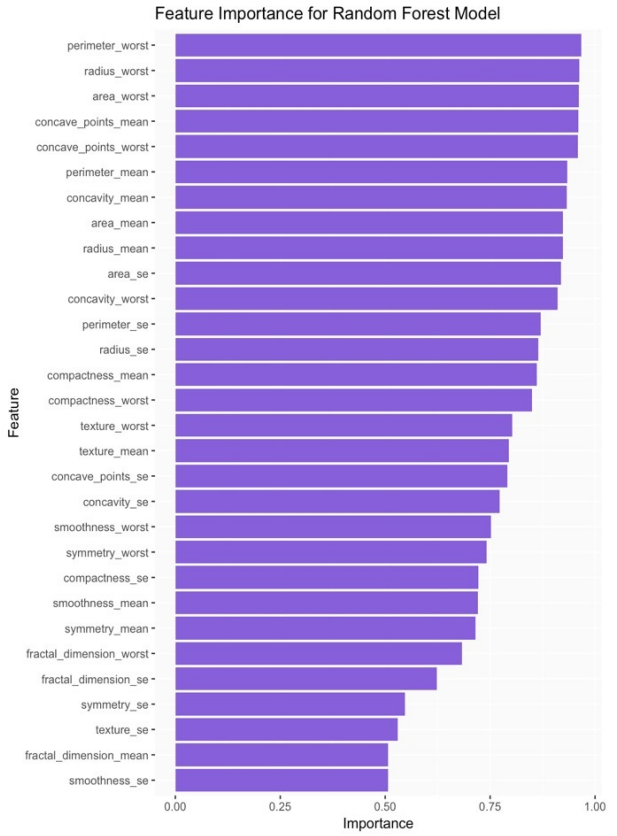

Figure 1: Feature importance of random forest model

Table 2: Attributes Scores

S.No	Names	Var_imp_scores
1	radius_mean	0.9229638
2	texture_mean	0.7948272
3	perimeter_mean	0.9336487
4	area_mean	0.9239305
5	smoothness_mean	0.7208865
6	compactness_mean	0.8613534
7	concavity_mean	0.9328877
8	concave_points_mean	0.9599548
9	symmetry_mean	0.7148807
10	fractal_dimension_mean	0.5073118
11	radius_se	0.8651584
12	texture_se	0.5300288
13	perimeter_se	0.8709070
14	area_se	0.9187269
15	smoothness_se	0.5070444
16	compactness_se	0.7222851
17	concavity_se	0.7726347
18	concave_points_se	0.7910839
19	symmetry_se	0.5474496
20	fractal_dimension_se	0.6225422
21	radius_worst	0.9622378
22	texture_worst	0.8026018
23	perimeter_worst	0.9675339
24	area_worst	0.9615693
25	smoothness_worst	0.7520053
26	compactness_worst	0.8500103
27	concavity_worst	0.9108700
28	concave_points_worst	0.9595640
29	symmetry_worst	0.7410325
30	fractal_dimension_worst	0.6824558

There is another useful method in Random Forest that is Out of Bag Error Rate(OBB error rate) [22-24], Usually, in the random forest only 2/3 part of data is used for training, and the rest can utilize for this purpose. The Predicted values of the testing dataset can be seen in table3. If the category of column and row are same, then it says that those are the correctly predicted values, other than those, they were the wrongly predicted values[25-27].

Table 3: Predicted Values

Class	B	M
B	70	1
M	1	41

In this paper, the OOB error rate is checked for across 100 trees using the random forest. These 100 trees contain a different set of training data so that the output prediction will be more accurate; here are the 100 trees OOB error rate, which seen in Figure 2.

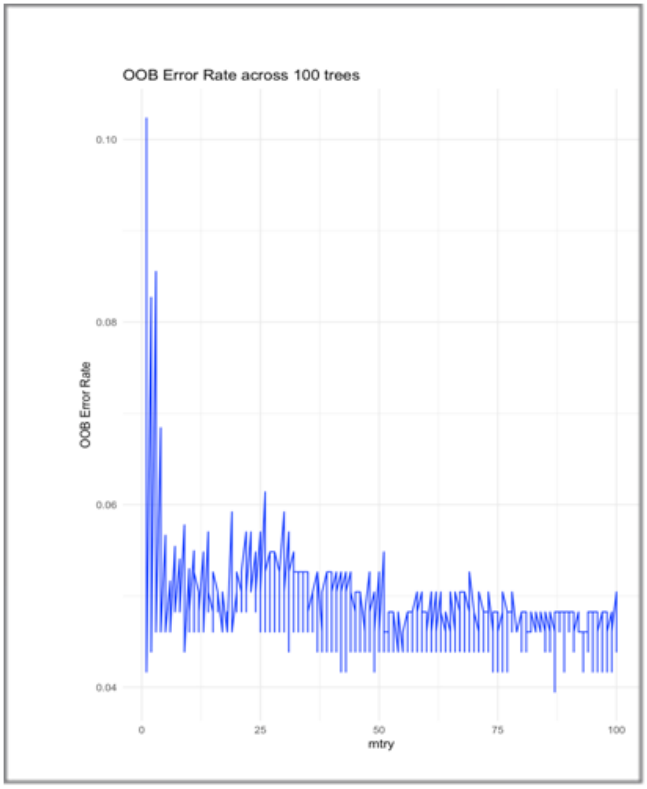

Figure 2: OBB Rate across 100 trees

The importance of OOB is, it gives us the evidence to show that OBB estimate is provided equal accuracy as using a test set of equal size as training data set.

IV. CONCLUSION

Breast Cancer is the occurring in the mass of women, and early prediction and diagnosis helps in a long and healthy life. In this paper, we discussed the random forest, a machine learning technique for the Wisconsin Breast Cancer.

REFERENCES

1. Maity, N. G. (2017). Machine learning for improved diagnosis and prognosis in healthcare. EEE Aerospace Conference, 1-9.
2. M. R. Al-Hadidi, A. A. (2016). Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm. 9th International Conference on Developments in eSystems Engineering (DeSE), 35-39.
3. A. Q. (2014). Breast cancer mass localization based on machine learning. IEEE 10th International Colloquium on Signal Processing and its Applications, Kuala Lumpur, 31-36.
4. J. A. Bhat, V. G. (2015). Cloud Computing with Machine Learning Could Help Us in the Early Diagnosis of Breast Cancer. Second International Conference on Advances in Computing and Communication Engineering, Dehradun, 644-648.
5. Repository, U. (n.d.). UCI Breast Cancer Wisconsin (Original)Dataset
6. Breiman, L. (2001). Machine Learning24(2). 123-140.
7. Ho, T. K. (1995). Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, 278–282.
8. Kolla, B.P., Dorairangaswamy, M.A. & Rajaraman, A. 2010, "A neuron model for documents containing multilingual Indian texts", 2010 International Conference on Computer and Communication Technology, ICCCT-2010, pp. 451.
9. Kolla, B.P. & Raman, A.R. 2019, Data Engineered Content Extraction Studies for Indian Web Pages.
10. Prakash, K.B. 2018, "Information extraction in current Indian web documents", International Journal of Engineering and Technology(UAE), vol. 7, no. 2, pp. 68-71.
11. Prakash, K.B. 2017, "Content extraction studies using total distance algorithm", Proceedings of the 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology, iCATccT 2016, pp. 673.
12. Prakash, K.B. 2015, "Mining issues in traditional Indian web documents", Indian Journal of Science and Technology, vol. 8, no. 32.
13. Prakash, K.B. 2015, "Mining issues in traditional indian web documents", Indian Journal of Science and Technology, vol. 8, no. 32, pp. 1-11.
14. Prakash, K.B., Ananthan, T.V. & Rajavarman, V.N. 2014, "Neural network framework for multilingual web documents", Proceedings of 2014 International Conference on Contemporary Computing and Informatics, IC3I 2014, pp. 392.
15. Prakash, K.B. & Dorai Rangaswamy, M.A. 2019, Content extraction studies for multilingual unstructured web documents.
16. Prakash, K.B. & Dorai Rangaswamy, M.A. 2016, "Content extraction studies using neural network and attribute generation", Indian Journal of Science and Technology, vol. 9, no. 22, pp. 1-10.
17. Prakash, K.B., Dorai Rangaswamy, M.A. & Ananthan, T.V. 2014, "Feature extraction studies in a heterogeneous web world", International Journal of Applied Engineering Research, vol. 9, no. 22, pp. 16571-16579.
18. Prakash, K.B., Dorai Rangaswamy, M.A., Ananthan, T.V. & Rajavarman, V.N. 2015, "Information extraction in unstructured multilingual web documents", Indian Journal of Science and Technology, vol. 8, no. 16.
19. Prakash, K.B., Dorai Rangaswamy, M.A. & Raman, A.R. 2010, "Text studies towards multi-lingual content mining for web communication", Proceedings of the 2nd International Conference on Trendz in Information Sciences and Computing, TISC-2010, pp. 28.
20. Prakash, K.B., Kumar, K.S. & Rao, S.U.M. 2017, "Content extraction issues in online web education", Proceedings of the 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology, iCATccT 2016, pp. 680.
21. Prakash, K.B. & Rajaraman, A. 2016, "Mining of Bilingual Indian Web Documents", Procedia Computer Science, pp. 514.
22. Prakash, K.B., Rajaraman, A. & Lakshmi, M. 2017, "Complexities in developing multilingual on-line courses in the Indian context", Proceedings of the 2017 International Conference On Big Data Analytics and Computational Intelligence, ICBDACI 2017, pp. 339.
23. Prakash, K.B., Rajaraman, A., Perumal, T. & Kolla, P. 2016, "Foundations to frontiers of big data analytics", Proceedings of the 2016 2nd International Conference on Contemporary Computing and Informatics, IC3I 2016, pp. 242.
24. Prakash, K.B. & Rangaswamy, M.A.D. 2016, "Content extraction of biological datasets using soft computing techniques", Journal of Medical Imaging and Health Informatics, vol. 6, no. 4, pp. 932-936.
25. Prakash, K.B., Rangaswamy, M.A.D. & Raja Raman, A. 2012, ANN for multi-lingual regional web communication.
26. Prakash, K.B., Rangaswamy, M.A.D. & Raman, A.R. 2013, "Attribute based content mining for regional web documents", IET Seminar Digest, pp. 368.
27. Prakash, K.B., Rangaswamy, M.A.D. & Raman, A.R. 2012, Statistical interpretation for mining hybrid regional web documents.

AUTHORS PROFILE

Vinoothna Manohar Botcha studying B.Tech Computer Science Engineering at Koneru Lakshmaiah Education Foundation. My specializations include Data science, Deep learning.

Dr. Kolla Bhanu Prakash working as research group head in Koneru Lakshmaiah Education Foundation. His specialisations include Data science, Deep learning, Soft Computing and Internet of Things. He is IEEE Senior Member and Member of ACM.