**Breast Cancer Prediction using Machine Learning Techniques**

**Japneet Singh 1, Mayank Gupta 2,**

Faculty of Engg. & Technology1,2

# Department of Computer Science and Engineering1,2,

# Manav Rachna International Institute of Research & Studies1,2

[japneet.singh064@gmail.com](mailto:japneet.singh064@gmail.com)1, [mayankgupta648@gmail.com](mailto:mayankgupta648@gmail.com)2,

***Abstract:*** Breast cancer is identified mainly among women and is a main reason for increasing mortality rate among women. Diagnosis of breast cancer is time consuming and due to less system availability, it is necessary to develop a system that can automatically diagnose breast cancer in its early stages. Different Machine Learning Algorithms have been utilized for the classification of benign and malignant tumors. The paper underlines on different models that is executed, for example Logistic Regression, forward and backward propagation for iteration, Support Vector Machine (SVM), Random Forest ,XGBoost and Modelling using confusion matrix etc. on the dataset taken from the vault of Private Hospital. Every one of these calculations has been estimated

and contrasted with deference with exactness and accuracy acquired. All the algorithms are coded in Python The experiments have indicated that Random Forest Classifier are the best for predictive analysis with train accuracy 1 and test accuracy 97. For the Prediction other algorithms are also implemented such as SVM and XGBoost.

The accuracy Obtained in the case of SVM is train accuracy 98 and test accuracy 97. The accuracy Obtained in the case of XGBoost is 96.

***Keywords****-* Classification, Machine Learning, Logistic Regression, Random Forest, Support Vector Machine, and XGBoost.

1. **INTRODUCTION**

A tumour is an abnormal lump or growth of cells. When the cells in the tumor are normal, it is benign. Something just went wrong, and they overgrew and produced a lump. When the cells are abnormal and can grow uncontrollably, they are cancerous cells, and the tumor is malignant. Breast cancer is a malignant cell growth in the breast. If left untreated, the cancer spreads to other areas of the body., breast cancer is the most common type of cancer in women in the United States, accounting for one of every three cancer diagnoses. The incidence of breast cancer rises after age 40. The highest incidence (approximately 80% of invasive cases) occurs in women over age 50. Cancer begins when healthy cells in the breast change and grow out of control, forming a mass or sheet of cells called a tumour. A tumour can be cancerous or benign. A cancerous tumour is malignant, meaning it can grow and spread to other parts of the body. A benign tumour means the tumour can grow but will not spread. Data Pre-processing it is one of the most important steps data mining. The way you collect data is often controlled incorrectly. The tumour classification process can be performed using machine learning techniques. Machine learning techniques may offer better accuracy when the data is complex and large compared to the machine. Research is ongoing in the area to various data sets on breast cancer.

**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 Random Forest ,XGBoost , SVM to find classification accuracy in breast cancer dataset. Their research shows Random Forest had produced higher accuracy in classification.

[2] Their research is about finding classification accuracy using machine learning algorithm known as forward and backward propagation with different values of i. For each value of i they have received a different result.

[3] Their paper is about using powerful machine learning classification algorithm Random Forest ,XGBoost , SVM which is usually used in data mining. Their work shows Random Forest did a better job in classification.

[4] Random forest classifier was implemented in their project to find mean accuracy.

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

[6] The model namely Logistic Regression was implemented. It was used to compare prediction accuracy breast cancer in mammography. Their study says logistic regression performed well in prediction.

[7] The performance and efficiency of the algorithms such as SVM, Random Forest, and Logistic Regression 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, SVM and Random Forest, the experiment conducted was focused on the evaluation of the algorithms.

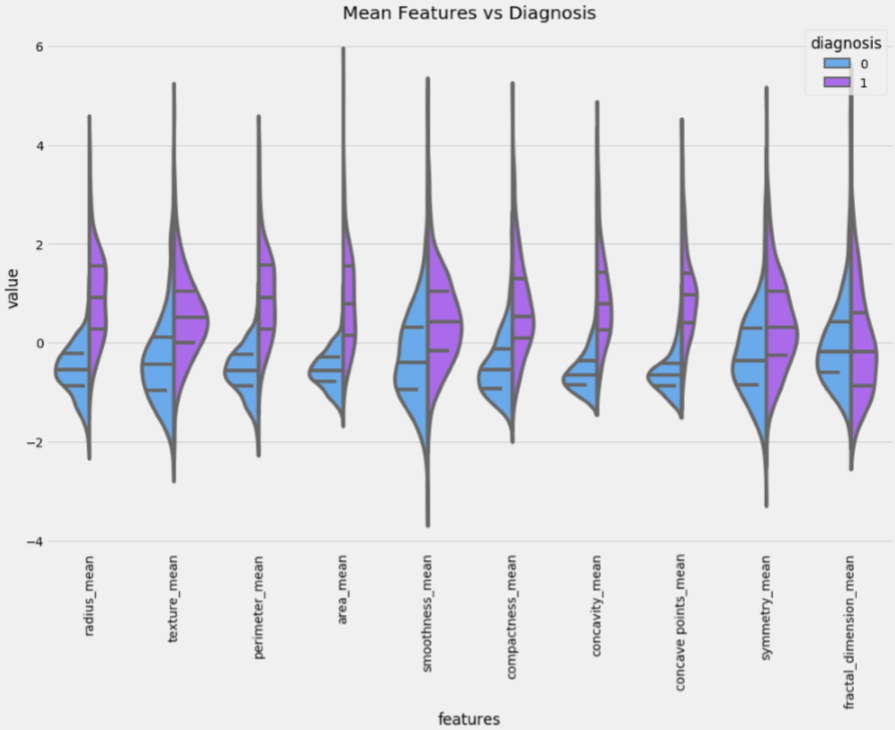
1. **Experiment Environment**

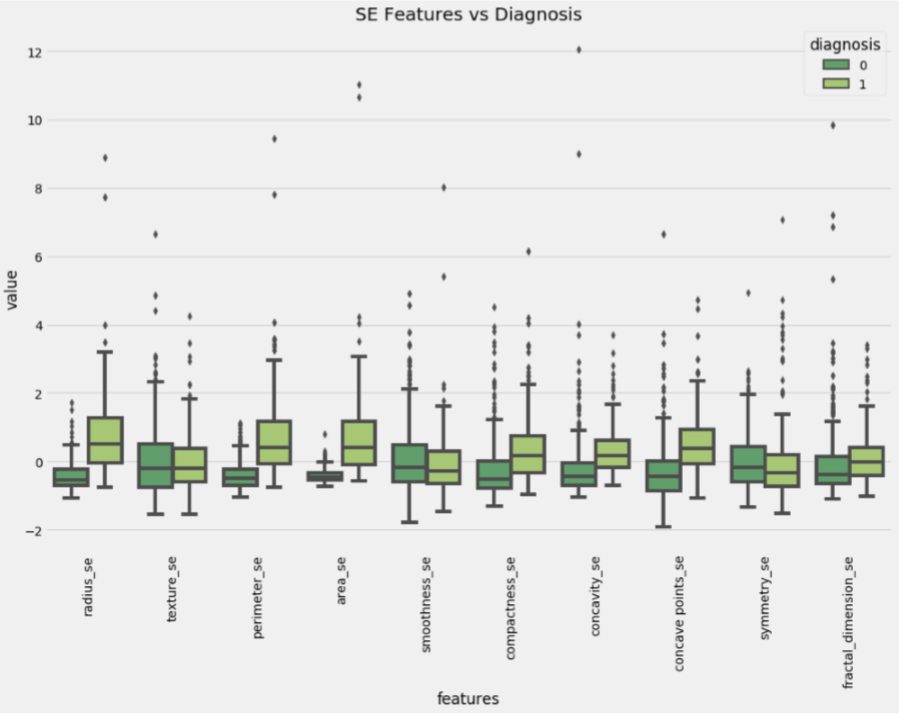
The sickit learn python libraries were used to conduct all experiments on classifiers explained in this paper. Sklearn is a collection of data mining, machine learning and deep learning algorithms used for classification, regression, data pre-processing and clustering. The sklearn libraries were used 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**

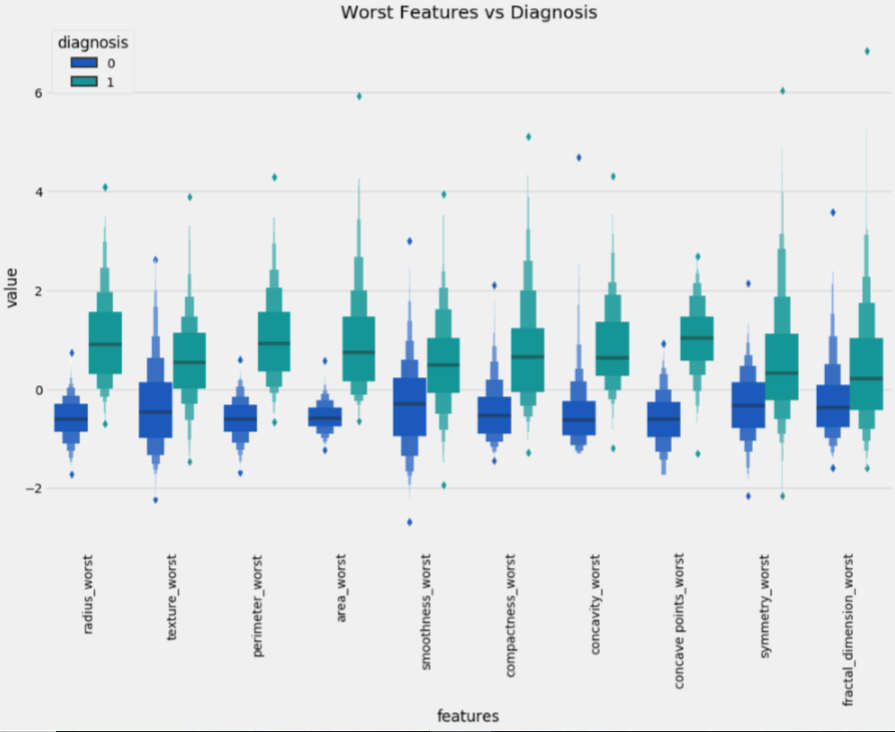
The UCI machine learning repository consists of The Breast Cancer datasets [10] which is used in this study. There are 699 instances in which 458 are benign and 241 are malignant. In addition, there are two classes malignant which contributes to 65.5% of dataset and benign 34.5%. The breast cancer dataset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







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

|  |  |  |
| --- | --- | --- |
| Algorithm | Test accuracy | Train accuracy |
| Support Vector Machine (SVM), , | 97 | 98 |
| Random Forest | 97 | 1 |
| XGBoost | 96 | 96 |
| Logistic Regression | 97 | 96 |

**V. RESULTS AND DISCUSSION**

Different machine learning algorithms such as Logistic Regression, forward and backward propagation for iteration, Support Vector Machine (SVM), Random Forest ,XGBoost and Modelling using confusion matrix for predicting breast cancer on the dataset. The maximum accuracy for the experiments have indicated that Random Forest Classifier are the best for predictive analysis with train accuracy 1 and test accuracy 97. For the Prediction other algorithms are also implemented such as SVM and XGBoost. For the Prediction other algorithms are also implemented such as SVM and XGBoost. The accuracy Obtained in the case of SVM is train accuracy 98 and test accuracy 97. The accuracy Obtained in the case of XGBoost is 96. Logistic Regression train accuracy 96 and test accuracy 97.

**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