Breast Cancer Prediction using Supervised Machine Learning Algorithms – Part II

Zeel Thakkar[1], Mamta Jadhav[2], Prof. Pramila M. Chawan[3]

[1]B.Tech Student, Dept. of Computer Engineering, VJTI College, Mumbai, Maharashtra, India
[2]B.Tech Student, Dept. of Computer Engineering, VJTI College, Mumbai, Maharashtra, India
[3]Associate Professor, Dept. of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India

Abstract - *Breast Cancer is the most leading malignancy affecting 2.1 million women each year which leads to greatest number of deaths among women. Early treatment not only helps to cure cancer but also helps in prevention of its recurrence. And hence this system mainly focuses on prediction of breast cancer where it uses different machine learning algorithms for creating models like decision tree, logistic regression, random forest which are applied on pre-processed data which suspects greater accuracy for prediction. Amongst all the models, Random Forest Classification leads to best accuracy with 98.6%. These techniques are coded in python and uses numpy, pandas, seaborn libraries.*

Index Terms: Decision Tree, Logistic Regression, Random Forest Classification, Numpy, Pandas, Seaborn.

1. INTRODUCTION

According to World health organization, Breast cancer is the most frequent cancer among women and it is the second dangerous cancer after lung cancer. In 2018, from the research it is estimated that total 627,000 women lost their life due to breast cancer that is 15% of all cancer deaths among women. In case of any symptom, people visit to oncologist. Doctors can easily identify breast cancer by using Breast ultrasound, Diagnostic mammogram, Magnetic resonance imaging (MRI), Biopsy. Based on these test results, doctor may recommend further tests or therapy. Early detection is very crucial in breast cancer. If chances of cancer are predicted at early stage then survivability chances of patient may increase. An alternate way to identify breast cancer is using machine learning algorithms for prediction of abnormal tumor. Thus the research is carried out for the proper diagnosis and categorization of patients into malignant and benign groups.

Three types of tumors are as follows:-

- Benign tumors are not cancerous they cannot spread or they can grow very slowly. And if doctors remove them, then they cannot cause any harm to the human body.
- In Premalignant tumors the cells are not cancerous but they have potential to become malignant.
- Malignant cells are cancerous and they can spread rapidly in body.

In machine learning, cancer classification can be done using benign or malignant cells could significantly improve our ability for prognosis in cancer patients, poor progress has been made for their application in the clinics. However, before gene expression profiling can be used in clinical practice, studies with larger data samples and more adequate validation are needed. And thus, our aim is to develop a prediction system that can predict chances of breast cancer on a huge data.

2. LITERATURE REVIEW

Data mining is been applied on medical data of the past and current research papers. Thorough study is done on various base reports. Jacob et al. [4] have compared various classifier algorithms on Wisconsin Breast Cancer diagnosis dataset. They came across that Random Tree and SVM classification algorithm produce best result i.e. 100% accuracy. However they mainly worked on 'Time' feature along with other parameters to predict the outcome of non-recurrence or recurrence of breast cancer among patients. In this paper, "Time" feature has not been relied upon for prediction of recurrence of the disease. Here, prediction is based on "Diagnosis" feature of WBCD dataset.

[5] used the SEER dataset of breast cancer to predict the survivability of a patient using 10-fold cross validation method. The result indicated that the decision tree is the best predictor with 93.6% accuracy on the dataset as compared to ANN and logistic regression model.

3. PROPOSED SYSTEM

3.1 Problem Statement

"To identify and compare which model is better for prediction of breast cancer symptoms at early stage to save someone's life by using data mining techniques and machine learning models on WBCD dataset."

Proposed System Architecture

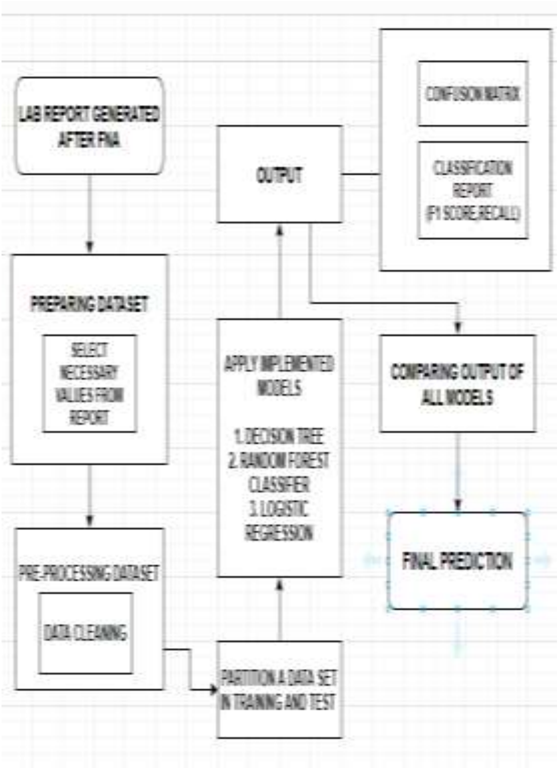

LAB REPORT GENERATED AFTER FNA
PREPARING DATASET
SELECT NECESSARY VALUES FROM REPORT
PRE-PROCESSING DATASET
DATA CLEANING
PARTITION A DATA SET IN TRAINING AND TEST
APPLY IMPLEMENTED MODELS
1. DECISION TREE
2. RANDOM FOREST CLASSIFIER
3. LOGISTIC REGRESSION
OUTPUT
CONFUSION MATRIX
CLASSIFICATION REPORT (F1 SCORE,RECALL)
COMPARING OUTPUT OF ALL MODELS
FINAL PREDICTION

Fig – 1: Work Flow

As shown in above diagram initially, from the obtained medical report a WDBC (Patients data) is prepared and will be taken as the input for the process and pre-process the input. Then, divide the data into Training and Testing data and train the models with 3 mentioned algorithms. Finally, Compare all the 3 modules created based on their accuracy and predict the cancer output for given data.

3.3 Proposed Methodology

We have used **jupyter notebook** as the platform for the purpose of coding. Our methodology involves use of supervised learning algorithms and classification technique like Decision Tree, Random Forest and Logistic Regression, with Dimensionality Reduction technique.

1. Collect the data from the medical reports or digitized image of FNA and prepare a numeric dataset.
2. Read this data using Pandas and process it as our requirements.
3. Apply All 3 algorithmic modules on this data and find the appropriate one based on their accuracy.

4. IMPLEMENTATION

4.1 Preparing Dataset

The dataset was created by Dr. William H. Wolberg, physician at the University of Wisconsin Hospital at Madison, Wisconsin, USA. To create the dataset Dr. Wolberg used fluid samples, taken from patients with solid breast masses and an easy-to-use graphical computer program called Xcyt, which is capable of perform the analysis of cytological features based on a digital scan. Wisconsin Breast Cancer Dataset have many features which are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass.

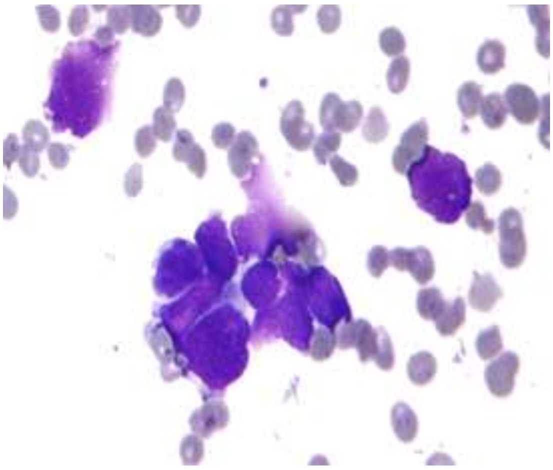
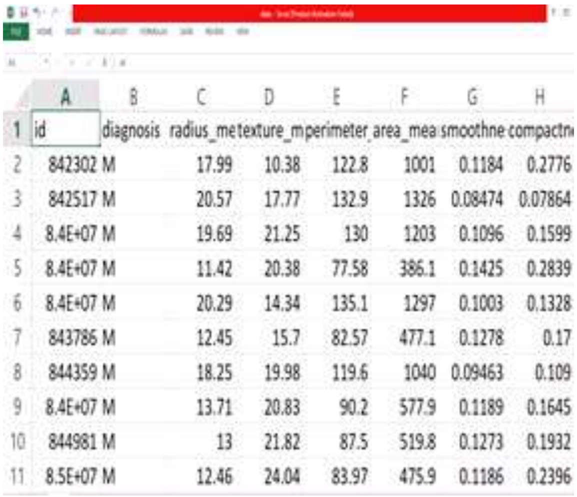


Fig – 2: FNA (initial level data)



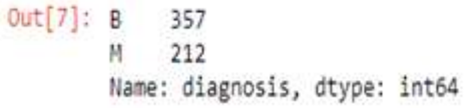
	A	B	C	D	E	F	G	H
1	id	diagnosis	radius_me	texture_m	perimeter	area_mea	smoothne	compactn
2	842302	M	17.99	10.38	122.8	1001	0.1184	0.2776
3	842517	M	20.57	17.77	132.9	1326	0.08474	0.07864
4	8.4E+07	M	19.69	21.25	130	1203	0.1096	0.1599
5	8.4E+07	M	11.42	20.38	77.58	386.1	0.1425	0.2839
6	8.4E+07	M	20.29	14.34	135.1	1297	0.1003	0.1328
7	843786	M	12.45	15.7	82.57	477.1	0.1278	0.17
8	844359	M	18.25	19.98	119.6	1040	0.09463	0.109
9	8.4E+07	M	13.71	20.83	90.2	577.9	0.1189	0.1645
10	844981	M	13	21.82	87.5	519.8	0.1273	0.1932
11	8.5E+07	M	12.46	24.04	83.97	475.9	0.1186	0.2396

Fig – 3: Dataset Prepared

4.2 Data Processing and Exploration:

Our dataset may be Incomplete or have some missing attribute values, or having only aggregate data. So, there is a need to pre-process our medical dataset which has major attribute as id, diagnosis and other real valued features which are computed for each cell nucleus like radius, texture, parameter, smoothness, area, etc.

We have to importing the necessary libraries as numpy, pandas, matplotlib, seaborn, sklearn, etc. We also examined the dataset using panda's head() function and then using shape() for printing the dimensions of dataset. After that, we found that out of the 569 persons, 357 are labeled as B (benign) and 212 as M (malignant).



```
Out[7]: B    357
        M    212
        Name: diagnosis, dtype: int64
```

Fig – 4: Count Of Diagnosis

4.3 Data Visualization

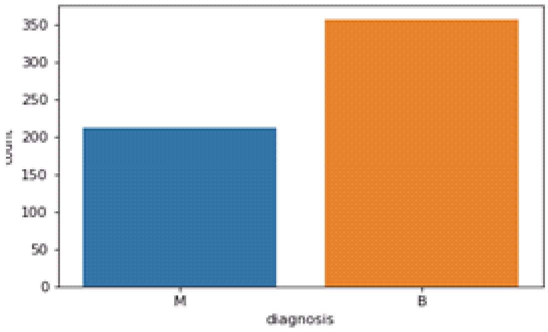


Fig – 5: Visualizing Diagnosis Feature

In our project, matplotlib is used to find the data distribution of the features. Here, we have also found the correlation between different features and then plot a chart of it using heatmap() method of seaborn library.

4.4 Data Categorization

Categorical data are variables that contain label values rather than numeric values. The number of possible values is often limited to a fixed set. We have used Label Encoder to label the categorical data. Label Encoder is the part of Sklearn library in Python and used to convert categorical data, or text data, into numbers, which our predictive models can better understand.

4.5 Data Splitting & Feature Scaling:

The data we use is usually split into training data and test data. In our project 75% data is trained data and 25% data is test data. This is done using SciKit-Learn library in Python using the train_test_split method.

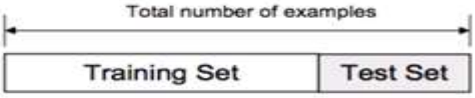


Fig - 6: Data Splitting

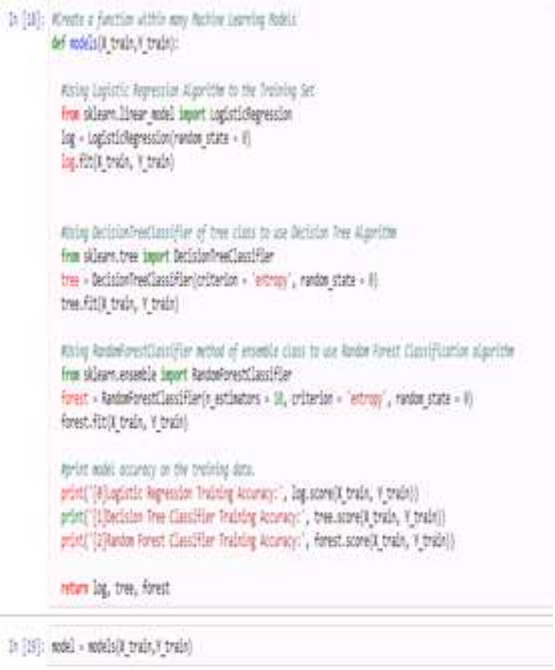
We have used StandardScaler method from SciKit-Learn library to bring all features of our dataset to the same level of magnitudes.

4.6 Model Implementation

A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. It is generally used when the dependent variable is binary and provides a constant output.

Decision tree is the most powerful and popular tool for classification and prediction in machine learning.The Decision trees algorithm consists of two parts: nodes and rules (tests).A Decision tree is like tree structure, where node denotes a test on an attribute, Branch represents an outcome of the test, and each leaf node holds a class label.

Random forest algorithm is a supervised classification algorithm. In this classifier, the **higher the number** of trees in the forest gives **the high accuracy** results.



```
In [18]: #Create a function within many Machine Learning Models
         def models(X_train,Y_train):

           #Using Logistic Regression Algorithm to the Training Set
           from sklearn.linear_model import LogisticRegression
           log = LogisticRegression(random_state = 0)
           log.fit(X_train, Y_train)


           #Using DecisionTreeClassifier of tree class to use Decision Tree Algorithm
           from sklearn.tree import DecisionTreeClassifier
           tree = DecisionTreeClassifier(criterion = 'entropy', random_state = 0)
           tree.fit(X_train, Y_train)

           #Using RandomForestClassifier method of ensemble class to use Random Forest Classification algorithm
           from sklearn.ensemble import RandomForestClassifier
           forest = RandomForestClassifier(n_estimators = 10, criterion = 'entropy', random_state = 0)
           forest.fit(X_train, Y_train)

           #print model accuracy on the training data.
           print('[0]Logistic Regression Training Accuracy:', log.score(X_train, Y_train))
           print('[1]Decision Tree Classifier Training Accuracy:', tree.score(X_train, Y_train))
           print('[2]Random Forest Classifier Training Accuracy:', forest.score(X_train, Y_train))

           return log, tree, forest

In [19]: model = models(X_train,Y_train)
```

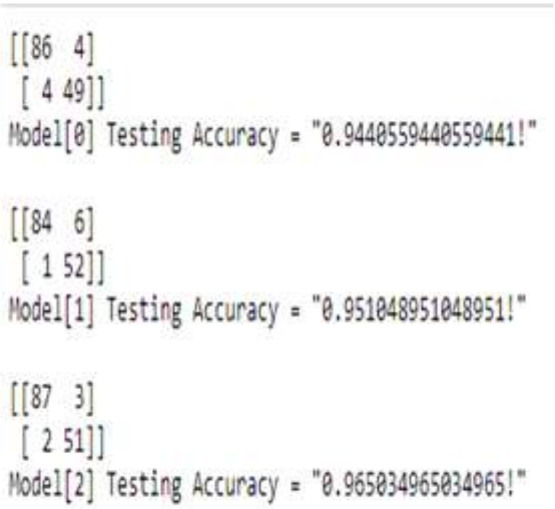
Fig – 7 : Implementing Algorithms

After implementing all these models, we got accuracy as observed above are as follows:-

1. Logistic Regression — 95.8%
2. Decision Tree Algorithm — 95.8%
3. Random Forest Classification — 98.6%

4.7 Output

1. **Confusion Matrix:**



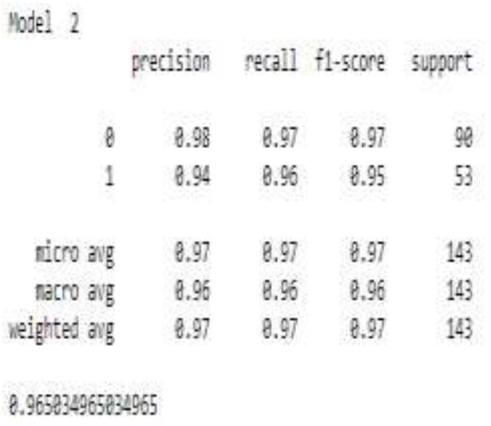
```
[[86  4]
 [ 4 49]]
Model[0] Testing Accuracy = "0.9440559440559441!"

[[84  6]
 [ 1 52]]
Model[1] Testing Accuracy = "0.951048951048951!"

[[87  3]
 [ 2 51]]
Model[2] Testing Accuracy = "0.965034965034965!"
```

Fig - 8: All models Output

We have imported confusion matrix method of metrics class. It is a way of tabulating the number of mis-classifications, i.e., the number of predicted classes which ended up in a wrong classification based on the true classes.

2. **Classification Report:**

As we observed in above output that Random Forest Classifier Provides fastest and the accurate prediction of Breast Cancer presence in a body. So, below output gives a detailed idea about this model's efficiency.



```
Model  2
              precision    recall  f1-score   support

           0       0.98      0.97      0.97        90
           1       0.94      0.96      0.95        53

   micro avg       0.97      0.97      0.97       143
   macro avg       0.96      0.96      0.96       143
weighted avg       0.97      0.97      0.97       143

0.965034965034965
```

Fig - 9 : Accuracy of Random Forest Model

5. CONCLUSION

In this paper, different types of models are reviewed and their accuracies are computed and compared with each other. Three main algorithms implemented are Decision Tree, Random Forest Classifier and Logistic Regression on Breast Cancer Dataset. Our work mainly focused in the advancement of predictive models to achieve good accuracy in predicting valid disease outcomes using supervised machine learning methods. Random Forest surpasses all the other algorithms with an accuracy of 99.53 %.

REFERENCES

[1]https://towardsdatascience.com/building-a-simple-machine-learning-model-on-breast-cancer-data-eca4b3b99fa3

[2]https://www.ijitee.org/wpcontent/uploads/papers/v8i6/F3384048619.pdf

[3]https://ieeexplore.ieee.org/document/7943207

[4] Shomona G. Jacob, R. Geetha Ramani, "Efficient Classifier for Classification of Prognosis Breast Cancer Data Through Data Mining Techniques", *Proceedings of the World Congress on Engineering and Computer Science 2012*, vol. I, October 2012.

[5] D. Delen, G. Walker, A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods", *Artificial Intelligence in Medicine*, vol. 34, no. 2, pp. 113127, 2004.

BIOGRAPHIES

Mamta Jadhav is currently pursuing B. Tech in Computer Engineering, from Veermata Jijabai Technological Institute (VJTI), Matunga, Mumbai. She has completed her Diploma in Computer engineering from K J Somaiya polytechnic Vidyavihar, Mumbai.

Zeel Thakkar is currently pursuing B. Tech in Computer Engineering, from Veermata Jijabai Technological Institute (VJTI) Matunga, Mumbai. She has completed her Diploma in Computer engineering from K J Somaiya polytechnic, Vidyavihar, Mumbai.

Pramila M. Chawan is working as an Associate Professor in the Computer Engineering Department of VJTI, Mumbai. She has done her B.E. (Computer Engg.) and M.E (Computer Engineering) from VJTI COE, Mumbai University. She has 27 years of teaching experience and has guided 75+ M. Tech. projects and 100+ B. Tech. projects. She has published 99 papers in the International Journals, 21 papers in the National/ International conferences/ symposiums. She has worked as an Organizing Committee member for 13 International Conferences, one National Conference and 4 AICTE workshops. She has worked as NBA coordinator of Computer Engineering Department of VJTI for 5 years. She had written proposal for VJTI under TEQIP-I in June 2004 for creating Central Computing Facility at VJTI. Rs. Eight Crore (Rs. 8,00,00,000/-) were sanctioned by the World Bank on this proposal.