Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

Developing A Web based System for Breast Cancer Prediction using XGboost Classifier

Nayan Kumar Sinha, Menuka Khulal, Manzil Gurung, Arvind Lal
Department of Computer Science and Technology
Centre for Computers and Communication Technology, Chisopani, Sikkim, India

***Abstract*- In today's world cancer is the most common diseases which lead to greatest number of death. Cancer is not one disease; it is a group of more than 100 different and distinctive diseases. Cancer can involve in any tissue of the body and have many different forms and in each body part. Breast Cancer is a grim disease and it is the only type of cancer that is widespread among women worldwide. As the diagnosis of this disease manually takes long hours and the lesser availability of systems, there is a need to develop the automatic diagnosis system for early detection of cancer. So in this project we are developing a web based diagnosis system for which we have done the comparative study of the supervised machine learning classifiers to get to know which classifier is giving the best accuracy. For that we have taken dataset from the Wisconsin breast cancer database (WBCD) which is the benchmark database for comparing the results through different algorithms. In which we will use following classification techniques of machine learning like Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest (RF), Adaboost Classifier and XGboost Classifier for the classification of benign and malignant tumor in which the machine is learned from the past data and can predict the category of new input.**

Keywords- WBCD, Support Vector Machine, K-Nearest Neighbor, Random Forest, Adaboost Classifier and XGboost Classifier.

1. INTRODUCTION

Breast cancer has become one of the most common diseases among women that lead to death. Breast cancer can be diagnosed by classifying tumors. There are two different types of tumors i.e. malignant and benign tumors. Doctors need a reliable diagnosis procedure to distinguish between these tumors. But generally it is very difficult to distinguish the tumors even by the experts. So automation of diagnostic system is needed for diagnosing. As the most prevalent cancer in women, breast cancer has always had a high incidence rate and mortality rate. According to the latest cancer statistics, breast cancer alone is expected to account for 25% of all new cancer diagnoses and 15% of all cancer deaths among women worldwide. In case of any sign or symptom, usually people visit doctor immediately, who may refer to an oncologist, if required. The oncologist can diagnose breast cancer by: Undertaking thorough the patient's medical history, examination of both the breasts and also check for swelling or hardening of any lymph nodes in the armpit. Here in this project, we have used the Wisconsin Breast Cancer Dataset (WBCD) of fine needle aspiration biopsy method and with that of the dataset we have invoked the machine learning algorithms to predict whether the patient is having breast cancer or not. This paper compares performance of five classification algorithms and their combination using ensemble approach that are suitable for direct interpretability of their results. We are using an XGboost classifier approach to compare other four classification algorithms and done the analysis of each classifiers accuracy of the best fit for the prediction of breast cancer.

2. PROBLEM STATEMENT

To identify which machine learning classifier gives the best accuracy. To count the number of patients having benign and malignant and also identify the type of tumor.

3. PROPOSED METHODOLOGY

We acquire the breast cancer dataset of Wisconsin Breast Cancer diagnosis dataset and used jupyter notebook and Anaconda Spyder as the platform for the purpose of coding and get the Prediction UI (user interface) output from the flask as in local server. Our methodology involves use of supervised learning algorithms and classification technique like Support Vector Classifier, KNN, Random Forest, Adaboost and Xgboost Classifier, with Dimensionality Reduction technique.

3.1 Data Manipulation

The data that we have it is in dictionary format and in sklearn we call it 'Bunch'. We have the keys of the dataset i.e. ('data', 'target', 'target_names', 'DESCR', 'feature_names', 'filename') and the values of this are in numeric format i.e. in 2d array format. Now the 'Target' means the patient who are having the breast cancer, the tumor is benign or malignant. Here malignant means the patient is having cancer and benign means the patient doesn't have the cancer.

0 means malignant tumor

1 mean benign tumor

In this dataset we have 569 numbers of instances with 30 features or attributes. As we know the features are in numeric format, so our 30 features are with the numeric values of each of the instances.

3.2 DataFrame

So the keys and values that we have, we combine the 'data' and 'target' to make the dataframe, it is because without dataframe we cannot apply the machine learning algorithm and by using the 'feature_name' and 'target' we have given the column name and then we store that into the file, so that it can help us in future purpose. Now we have checked our dataset's information and there are no null values, all the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

features are having float64 format. Now we have taken the numerical distribution of our dataset and describe it.

3.3 Data Visualization

We have to visualize our data as because it is in numerical format so we have to take the pair plot of our dataset and it is already distributed in two categories i.e. in benign 1 and malignant 0 and we can easily distributed it in blue and orange.

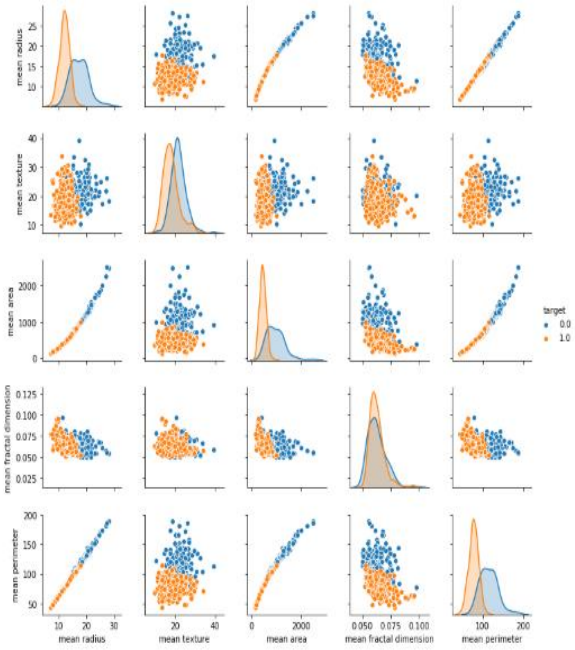

Fig 1: Pairplot of all the Features

Now we have took the counter plot of our dataset to count total how many patients are having benign and malignant tumor.

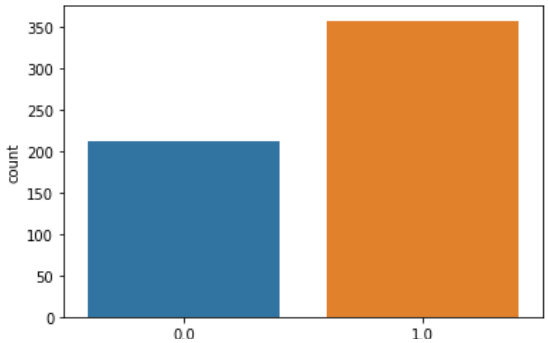

Fig 2: Total count of malignant and benign tumor patients in counterplot

So here the count of malignant tumor instances are of 220-230 and the benign tumor instances is high rather than malignant.

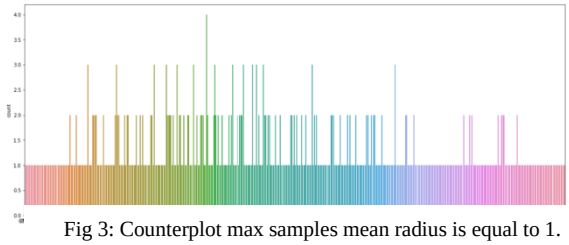

Fig 3: Counterplot max samples mean radius is equal to 1.

We have also counter plot the feature mean radius of the dataset, where we find those patients who doesn't have cancer their mean radius is near about 1 whereas the patients who are having cancer their mean radius is more than 1. We also took the correlation barplot, over here we have took the correlation with the target features.

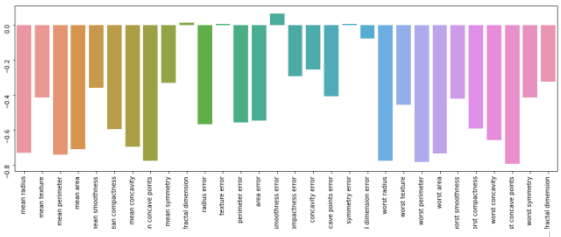

Fig 4: Correlation Barplot of all the Features

In the above correlation barplot only feature 'smoothness error' is strongly positively correlated with the target than others. The features 'mean factor dimension', 'texture error', and 'symmetry error' are very less positive correlated and others remaining are strongly negatively correlated.

3.4 Data Preprocessing

It is a technique that is used to convert the raw data into a clean data set and also refers to the transformations applied to our data before feeding it to the algorithm. For getting better results from the Machine Learning applied model, the format of the data has to be in a proper manner and in a specified format, for example, Random Forest algorithm does not support null values, 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 (mean of distances from center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness (perimeter^2 / area - 1.0), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, fractal dimension ("coastline approximation" - 1).

3.4.1 Split DataFrame in Train and Test

In our project 75% data is trained data and 25% data is test data.

3.4.2 Feature Scaling

Generally, dataset contains features which highly vary in magnitudes, units and range. So there is a need to bring all

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

features to the same level of magnitudes. This can be achieved by scaling.

3.5 Model Selection

This is the most important phase where machine learning algorithm selection is done for the developing a system where Data Scientists use various types of Machine Learning algorithms which can be classified as: supervised learning and unsupervised learning. For this breast cancer Prediction System, we only need Supervised Learning.

3.5.1 Supervised Learning

The supervised learning algorithm learns from the training data, which helps you to predict the outcomes for unpredicted data. It helps you to optimize performance criteria using experience also helps you to solve various types of real-world computation problems and such classifiers that are used mostly briefly explained below.

3.5.1. (I) Support Vector Machine (SVM)

It is one of the most popularized Supervised Learning algorithm, which is used for Classification as well as Regression problems. However, basically, it is used for Classification problems in Machine Learning scenario. The intent of the SVM algorithm is to create the best decision boundary that can segregate n-dimensional space into classes so that it can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane of SVM.

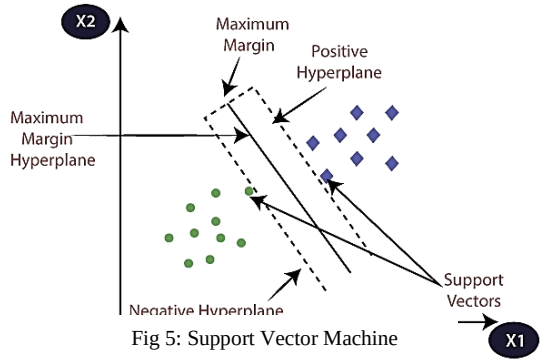

Fig 5: Support Vector Machine

3.5.1. (II) K - Nearest Neighbor (K-NN)

It is one of the simplest Machine Learning algorithms based on Supervised Learning technique. And assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. It stores all the available data and classifies a new data point based on the similarity and easily classified into a well suite category by using K- NN algorithm.

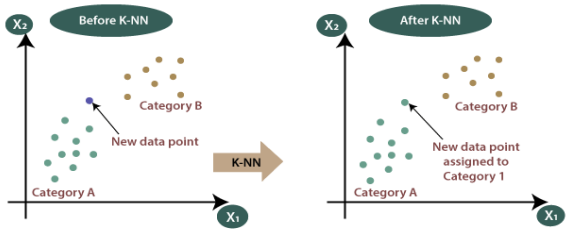

Fig 6: K - Nearest Neighbor

3.5.1. (III) Random Forest Classifier

Random Forest classifier is a learning method that operates by constructing multiple decision trees and the final decision is made based on the majority of the trees and is chosen by the random forest. It is a tree-shaped diagram used to determine a course of action. Each branch of the tree represents a possible decision, instance, or reaction. Using of Random Forest Algorithm is one of the main advantages is that it reduces the risk of over fitting and the required training time. Additionally, it also offers a high level of accuracy.

It runs efficiently in large databases and produces almost accurate predictions by approximating missing data.

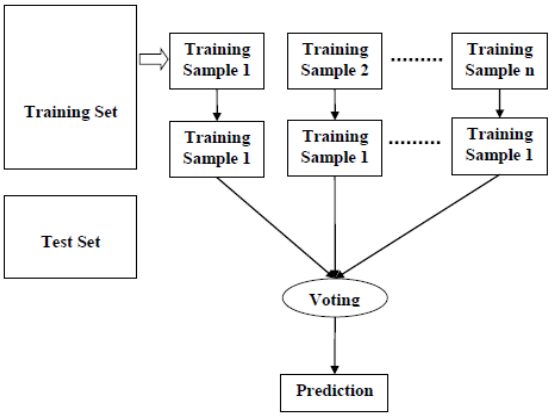

Fig 7: Random Forest Classifier

Using of Random Forest Algorithm is one of the main advantages is that it reduces the risk of over-fitting and the required training time. Additionally, it also offers a high level of accuracy and produces highly accurate predictions by estimating missing data.

3.5.1. (IV) Adaboost Classifier

Ada-boost or Adaptive Boosting is an iterative ensemble boosting classifier. It builds a robust classifier by combining all poor performing classifiers to get the high accuracy, the concept behind Adaboost is to set the multiple weighs of classifiers and train the data in each iteration, hence it ensures the exact prediction of unusual observation. AdaBoost refers to a particular method of training a boosted classifier. Adaboost classifier is a classifier in the form of

$$F_T(x) = \sum_{t=1}^{T} f_t(x)$$

Where each f_t is a weak learner that takes an object X as input and returns a value indicating the class of the object.

3.5.1. (V) XGboost Classifier

eXtreme Gradient Boosting or XGBoost is a library of gradient boosting algorithms optimized for modern data science problems and tools. Some of the major benefits of XGBoost are that it's highly scalable/parallelizable, quick to execute, and typically outperforms other algorithms and

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

used a more regularized model formalization, to control over-fitting, which gives it better performance.

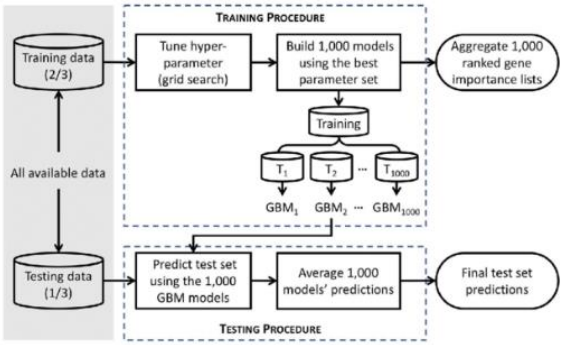

Fig 8: XGboost Classifier

Above diagram is the schematic of the XGBoost workflow. The shaded area indicates the training data and testing data. The boxes inside the dashed lines indicates training and testing procedures where T stands for tree and GBM stands for gradient boosting machine. Out of the dashed box the two oval boxes on the right depict the outputs from XGBoost.

Table 1: Comparison between SVM, KNN, RF, Adaboost, XGboost Classifiers.

Techniques	Accuracy without Standard scale	Accuracy with Standard Scale
SVM	57 %	96%
KNN	93%	57%
RF	97%	75%
Adaboost	94%	94%
XGboost	98%	98%

4. CONFUSION MATRIX

It is a summary of prediction results on a classification problem with the number of correct and incorrect predictions that are summarized with count values and broken down by each class. This is the key to the confusion matrix. It shows the ways in which your classification model get confused when it make predictions. It gives intuition not only into the errors being made by a classifier but more importantly the types of errors that are being made.

	Predicted **No**	Predicted **Yes**
Actual **No**	**TP** True Positive	**FN** False Negative
Actual **Yes**	**FP** False Positive	**TN** True Negative

Classification Rate/Accuracy:

Classification Rate or Accuracy is given by the relation:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \text{ *100}$$

Accuracy = (TP + TN) / (TP + TN + FP + FN) = (46 + 66) / (46 + 66 + 0 + 2) *100 = 98.24

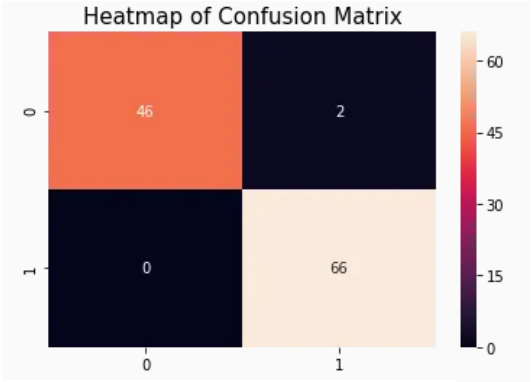

Fig 10: Heatmap of Confusion Matrix Model

The model is giving 0% type II error and it is best.

5. PROPOSED SYSTEM ARCHITECTURE

As shown in below diagram, we first collected the Dataset from Wisconsin Breast Cancer Dataset (WBCD). To applying a machine learning models, collecting appropriate data is very essential. After Collection of data, Cleaning needs to be done for removal of unwanted observations and for deleting duplicate or irrelevant values from dataset. Above mentioned Models have been comparatively studied which is used in this project and predicts the chances of breast cancer.

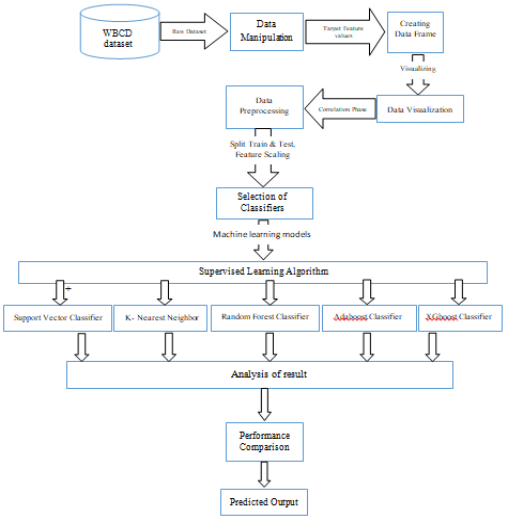

Fig 11: Work Flow

6. CONCLUSION AND FUTURE SCOPE

To analyse medical data, various data mining and machine learning methods are available. It's an important challenge in data mining and a machine learning area is to build accurate and computationally efficient classifiers for Medical applications. So in this project, we employed the machine learning classifier algorithms on the Wisconsin Breast Cancer (original) datasets and try to compare efficiency and effectiveness of those algorithms to find the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

best classification accuracy, where XGBOOST classifier is giving us the maximum accuracy.

Well in Future Scope, various new deep learning algorithms are required to be implemented for the detection of different stages and categories of breast cancer simultaneously.

REFERENCES

[1] Ch. Shravya, K. Pravalika, Shaik Subhani, "Prediction of Breast Cancer Using Supervised Machine Learning Techniques International", Journal of Innovative Technology and Exploring Engineering (IJITEE) Volume-8 Issue-6, April 2019.

[2] Mamta Jadhav[1], Zeel Thakkar[2], Prof. Pramila M. Chawan[3], "Breast Cancer Prediction using Supervised Machine Learning Algorithms", International Research Journal of Engineering and Technology (IRJET)Volume: 06 Issue: 10 Oct 2019.

[3] R. Chtihrakkannan, P. Kavitha, T. Mangayarkarasi, R. Karthikeyan, "Breast Cancer Detection using Machine Learning", International Journal of Innovative Technology and Exploring Engineering (IJITEE) Volume-8 Issue-11, September 2019.

[4] Mandeep Rana[1], Pooja Chandorkar[2], Alishiba Dsouza[3], Nikahat Kazi[4], "Breast Cancer Diagnosis and Recurrence Prediction using Machine Learning techniques", IJRET: International Journal of Research in Engineering and Technology Volume: 04 Issue: 04 Apr-2015.

[5] Varsha J. Gaikwad, "Detection of Breast Cancer in Mammogram using Support Vector Machine", International Journal of Scientific Engineering and Research (IJSER) Volume 3 Issue 2, February 2015.

[6] Susmitha Uddaraju[1], M. R. Narasingarao[2], "A Survey of Machine Learning Techniques Applied for Breast Cancer Prediction", International Journal of Pure and Applied Mathematics (IJPAM) Volume 117 No. 19 2017.

[7] Rajkamal kaur Grewal Babita Pandey, "Two Level Diagnosis of Breast Cancer Using Data Mining", International Journal of Computer Applications (IJCA) Volume 89 – No 18, March 2014.

[8] Priyanka Gupta, Prof. Shalini L, "Analysis of Machine Learning Techniques for Breast Cancer Prediction", International Journal Of Engineering And Computer Science (IJECS) Volume 7 Issue 5 May 2018.

[9] Ravi Aavula, R. Bhramaramba, "An Extensible Breast Cancer Prognosis Framework for Predicting Susceptibility, Recurrence and Survivability", International Journal of Engineering and Advanced Technology (IJEAT) Volume-8 Issue-5, June 2019.

[10] Dania Abed Aljawad1, Ebtesam Alqahtani2, Ghaidaa AL-Kuhaili3, Nada Qamhan4, Noof Alghamdi5, Saleh Alrashed6, Jamal Alhiyafi7, Sunday O. Olatunji8, "Breast Cancer Surgery Survivability Prediction Using Bayesian Network and Support Vector Machines", 978-1-4673-8765-1/17/$31.00 ©2017 IEEE

[11] Mehrdad J. Gangeh, Senior Member, IEEE, Simon Liu, Hadi Tadayyon, and Gregory J. Czarnota, "Computer Aided Theragnosis Based on Tumour Volumetric Information in Breast Cancer", DOI 10.1109/TUFFC.2018.2839714, IEEE

[12] Madhuri Gupta1, Bharat Gupta2, "A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques", 978-1-5386-3452-3/18/$31.00 ©2018 IEEE

[13] Afsaneh Jalalian, Babak Karasfi, "Machine Learning Techniques for Challenging Tumor Detection and Classification in Breast Cancer", 978-1-7281-2842-9/18/$31.00 ©2018 IEEE

[14] U. Karthik Kumar1, M.B. Sai Nikhil2 and K. Sumangali3, "Prediction of Breast Cancer using Voting Classifier Technique", 978-1-5090-5905-8/17/$31.00 ©2017 IEEE

[15] Xingyui Li1 (Member, IEEE), Marko Radulovic2, Ksenija Kanjer2, and Konstantinos N. Plataniotis1, "Discriminative Pattern Mining for Breast Cancer Histopathology Image Classification via Fully Convolutional Auto-encoder ", (Fellow, IEEE) DOI 10.1109/ACCESS.2019.2904245, IEEE Access