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

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 08, August-2020

Breast Cancer Classification using Python Programming in Machine Learning

Shruthi S[1], Binu Xavier F[2], Ravi Kumar A[3], Yeshwanth S[4], Dr. Mahalinga V Mandi[5]
[1, 2, 3, 4] Student, Dept. of Electronics and Communication Engineering, Dr. Ambedkar Institute ofTechnology, Karnataka, India
[5] Prof. Dept. of Electronics and Communication Engineering, Dr. Ambedkar Institute of Technology, Karnataka, India.

Abstract **- Breast cancer is a disease in which cells in the breast grow out of control in a rapidly. Breast cancer occurs when a malignant (cancerous) tumor originates in the breast cells. It is the most commonly occurring cancer in women and the second most common cancer overall. Around 2 million cases were observed in 2018. The early diagnosis of breast cancer can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients affected. Further accurate classification from the data of benign tumors can prevent patients from undergoing unnecessary treatments. Thus, the correct diagnosis of breast cancer and the classification of patients into malignant or benign groups is the subject of all research done and observed. Because of its unique advantages in critical features detection from complex breast cancer datasets, machine learning (ML) is widely recognized as the methodology of choice in Breast Cancer pattern classification. This project is a relative study of the implementation of models using Logistic Regression, SVM, KNN, Random Forest, and Decision tree, which is done on the data set taken from the UCI repository. We have obtained the highest accuracy for the random forest that is 97%. We have also obtained the accuracy of 95%, 93%, 95%, 94% for logistic regression, SVM, KNN and Decision tree respectively**

Key Words: (Size 10 & Bold) Key word1, Key word2, Key word3, etc (Minimum 5 to 8 key words)…

1. INTRODUCTION

Breast cancer (BC) is the most common cancer in women, affecting about 10 percent of all women at some stages of their life. In modern times, the rate keeps increasing and data show that the survival rate is 88 percent after five years from diagnosis and 80 percent after 10 years from diagnosis. Early prediction of breast cancer so far have made heaps of improvement, death rate of breast cancer by 39 percent, starting from 1989. Due to varying nature of breast cancers symptoms, patients are often subjected to a barrage of tests, including but not limited to mammography, ultrasound and biopsy, to check their likelihoods of being diagnosed with breast cancer. Biopsy, is the most indicative among these procedures, which involves extraction of sample cells or tissues for examination. The sample of cells is obtained from a breast fine needle aspiration (FNA) procedure and then sent to a pathology laboratory to be examined under a microscope. Numerical features, such as radius, texture, perimeter and area, can be measured from microscopic images. Data, later on, obtained from FNA are analyzed in combination with various imaging data to predict probability of the patient having malignant breast cancer tumor. An automated system here would be hugely beneficial in this scenario. It will likely expedite the process and enhance the accuracy of the doctor's predictions. In addition, if supported by abundance dataset and the automated system consistently performs well, it will potentially eliminate the needs for patients to go through various other tests, such as mammography, ultrasound, and MRI, which subject patients to significant amount of pain and radiation. In all, early prediction remainsis one of the vital aspectsin the follow-up process. Data mining methods or classification can help to reduce the number of false positive and false negative decisions. Consequently, new ways like data discovery in databases has become a preferred tool for medical researcher. In this paper, using six classification models; Decision Tree, K-Neighbors, Logistic Regression, Random Forest and Support Vector Machine (SVM) have been run on the Wisconsin Breast Cancer (original) Datasets, both before and after applying Principal Component Analysis. The results obtained are than measured using various performance metrics to compare among the algorithms in order to find out the best suited model for cancer prediction.

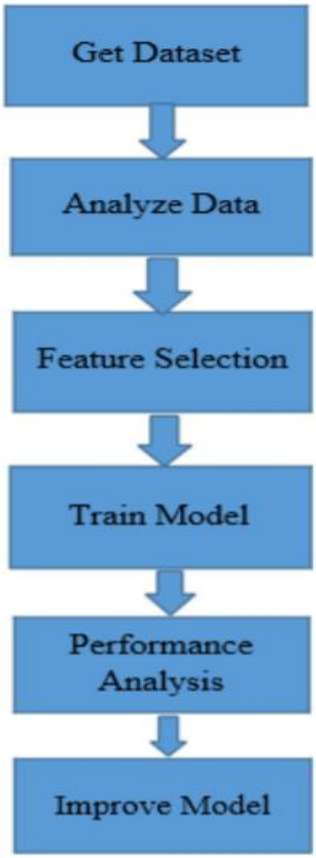
2. PROBLEM FORMULATION

Many people are affected from breast cancer at the present time. Causing of this disease depends on man factors and cannot be simply determined. In addition, the identification method that determines whether or not the cancer is benign or malignant additionally needs an excellent deal of effort from a doctors and physicians. Once many tests are concerned within the identification of breast cancer, like clump thickness, uniformity of cell size, uniformity of cell form, etc., the ultimate result could also be troublesome to get, even for doctors. This has given an increase within the previous few years to the utilization of machine learning and computing generally as diagnostic tools. The diseases that take numerous lives, diagnostic computer-based applications are used wide. Robotics are taking part in an awfully necessary role in operational rooms. Also, the skilled systems are conferred within the intensive treatment rooms. In turn, using another side of Artificial intelligence for breast cancer designation isn't unworthy. It's reported that breast cancer illness is that the second commonest cancer that affects girls, and was the rife cancer within the world by the year of 2002. This cancer may be a quite common sort of cancer among girls and therefore the second highest reason behind cancer death. Within the United State, regarding one in eight girls over their time period includes a risk of developing breast cancer. With the uncontrolled division of one cell inside the breast leads to beginning to the breast cancer which results in a visible mass, called a tumor. The tumor can be either benign or malignant. The correct designation in determinant whether or not the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 08, August-2020

tumor is benign or malignant may result in saving lives. Therefore, the necessity for precise classification within the clinic may be an explanation for nice concern for specialists and doctors. This importance of Artificial intelligence has been actuated for the last twenty five years, once scientists began to understand the quality of taking bound selections to treat specific diseases. The employment of machine learning and data processing as tools in diagnosing becomes terribly effective and one amongst the crucial diseases in medicines wherever the classification task plays a really essential role is that the diagnosis of breast cancer. Therefore, machine learning techniques will facilitate doctors to create an correct identification for breast cancer and make the proper classification of being benign or malignant tumor. There is little question that analysis of information taken from the patient and selections of doctors and specialists are the foremost necessary factors within the identification, however knowledgeable systems and artificial Intelligence techniques like machine learning for classification tasks, conjointly facilitate doctors and specialists in a great deal. We aim in this paper from to compare different classification learning algorithms significantly to predict a benign from malignant cancer in breast cancer dataset. We aim to investigate different machine learning techniques and we will use several algorithms and apply on breast cancer dataset. We will focus on machine learning algorithms: Naïve Bayes, K-nearest neighbor, logistic regression, reinforcement algorithm, support vector machine algorithm. We will primarily study these various algorithms and analyze their result.

3. METHODOLOGY



3.1 WORKING PRINCIPLE

The main plan of principal component analysis (PCA) is to cut back the dimensionality of a data set consisting of the many variables related with one another, either heavily or gently, whereas holding the variation present within the data set, up to the utmost extent. The identical is finished by remodeling the variables to a replacement set of variables, that are referred to as the principal elements (or merely, the PCs) and are orthogonal, ordered specified the retention of variation present within the original variables decreases as we tend to move down within the order. So, during this method, the first principal element retains most variation that was gift within the original elements. The principal elements are the Manfred Eigen vectors of a co variance matrix, and therefore they're orthogonal. Importantly, the dataset on that PCA technique is to be used should be scaled. The results are sensitive to the relative scaling. As a layman, it's a technique of summarizing information. Imagine some wine bottles on a board. every wine is delineate by its attributes like color, strength, age, etc. however redundancy can arise as a result of several of them can live connected properties. Thus what PCA can neutralize this case is summarize every wine within the stock with less characteristics. Intuitively, Principal part Analysis will provide the user with a lower-dimensional image, a projection or "shadow" of this object once viewed from its most in-formative viewpoint.

4. RESULT ANALYSIS

The next step after applying implementing machine learning models is to seek out how effective is that the model, i.e. how the models performed on the datasets. This is carried out by running the models on the test dataset which was set earlier. The test dataset comprised of 30% of the dataset for Breast Cancer prediction.10-fold cross-validation was also done for Breast cancer pre-diction. In order to determine and compare the performances of the different algorithms, several metrics have been used.

4.1 Performance metrics

Several performance metrics have been used to figure out the performance of the Machine Learning algorithms in this our thesis. As the paper sincerely deals with classification problems, performance metrics relating to classifications are discussed here. For Breast Cancer prediction, if the target variable is 1(malignant), then it is a positive instance, meaning the patient has Breast cancer. And if the target variable is 0 (benign), then it is a negative instance, stating that the patient does not have the cancer.

4.2 Confusion Matrix

Summarization the performance of a classification algorithm is based on a technique which is known as confusion matrix. It is arguably the easiest way to regulate the performance of a classification model by comparing how many positive instances are correctly/incorrectly classified and how many negative instances are correctly/incorrectly classified. In a confusion matrix, as shown here, the rows represent the actual labels while the columns represent the predicted labels.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 08, August-2020

True Positives (TP)
These are the occurrences where both the predictive and actual class is true (1), i.e., when the patient has complications (breast cancer in this case) and is also classified by the model to have complications.

True negatives (TN)
True negatives are the occurrences where both the predicted class and actual class is False (0), i.e., when a patient does not have complications and is also classified by the model as not having complications.

False Negative (FN)
These are occurrences where the predicted class is False (0) but actual class is True (1), i.e., case of a patient being classified by the model as not having complications even though in reality, they do.

False Positive (FP)
False positives are the occurrences where the predicted class is True (1) while the actual class is False (0), i.e., when a patient is classified by the model as having complications even though in reality, they do not.

Normalized matrix
Normalized Confusion Matrix represents results in a more efficient way. The results are similar to that of the confusion matrix. The values are distributed within the range of 0-1. An even distribution of data makes prediction easier.

Accuracy
Evaluation of classification models is done by one of the metrics called accuracy. Accuracy is the fraction of prediction. It determines the number of correct predictions over the total number of predictions made by the model.

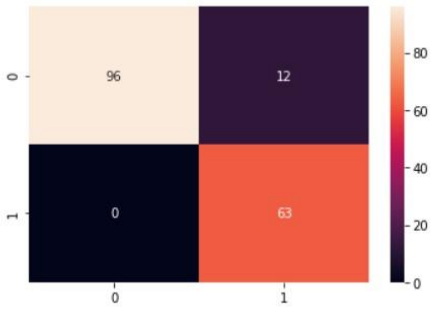
Recall
It is a measure of the proportion of patients that were predicted to have the complications among those patients that actually have the complications.

Precision
It is described as a measure of proportion of patients that actually have complications among those classified to have complications by the model.

Specificity
Classifier's performance to spot negative results is related by Specificity. It is exactly the negative of Recall. It is a measure of the number of patients who are classified as not having complications among those who actually did not have the complications.

F1 Score
Weighted average of precision and recall is known as F1 score. Therefore false positives and false negatives are taken by this score into the consideration. Intuitively it is not as simple to grasp as accuracy, however F1 is typically additional helpful than accuracy.

```
Out[23]: <matplotlib.axes._subplots.AxesSubplot at 0x7f83140ef4d0>
```



Model 0

	precision	recall	f1-score	support
0	0.97	0.96	0.96	90
1	0.93	0.94	0.93	53
accuracy			0.95	143
macro avg	0.95	0.95	0.95	143
weighted avg	0.95	0.95	0.95	143

0.951048951048951

Model 3

	precision	recall	f1-score	support
0	0.94	0.99	0.96	90
1	0.98	0.89	0.93	53
accuracy			0.95	143
macro avg	0.96	0.94	0.95	143
weighted avg	0.95	0.95	0.95	143

0.951048951048951

Model 2

	precision	recall	f1-score	support
0	0.98	0.97	0.97	90
1	0.94	0.96	0.95	53
accuracy			0.97	143
macro avg	0.96	0.96	0.96	143
weighted avg	0.97	0.97	0.97	143

0.965034965034965

Model 1

	precision	recall	f1-score	support
0	0.98	0.92	0.95	90
1	0.88	0.96	0.92	53
accuracy			0.94	143
macro avg	0.93	0.94	0.93	143
weighted avg	0.94	0.94	0.94	143

0.9370629370629371

5. CONCLUSIONS

In terms of accuracy, Random Forest have scored high figures of 0.97, without applying PCA. K-Neighbors (0.9349) and Logistic regression (0.923) are not far behind either. SVM scores 0.917 in accuracy. Decision Tree performs the worst among all six resulting 0.834. Application of PCA declines the accuracy of all the algorithms except Decision tree. However, the accuracy figures are still higher than that of Random Forest,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 08, August-2020

again, performs best after PCA is applied, even though there is a fall in accuracy (0.917). Considering the other performance matrix into account, a lot can be determined regarding the performance of the algorithms. Decision tree and K-Neighbors performs better without the introduction of PCA, while Logistic Regression and SVM performs better after PCA is applied to the dataset. SVM and Logistic Regression scores a perfect 1.000 when it comes to recall, which is vital in terms of disease prediction, after PCA is applied, even though there are declines in the values of all other performance metrics of both the mentioned algorithms. Keeping in mind that PCA reduces the run time exponential to huge extends in datasets (both small and large alike) and keeping the recall score into consideration, we can conclude that Logistic Regression and Support Vector Analysis with PCA performs better when it comes to Breast Cancer Prediction for this dataset used.

6. FUTURE ENHANCEMENT

Despite attaining accurate results and accuracies with the six algorithms we have used, we wish to confirm the results we obtained are not biased thanks to the scale of our dataset. We would like to search out an even bigger dataset and perform similar analysis and see if the results are the identical. Furthermore, since our dataset is kind of obsolete (collected within the 90s), more criteria for prediction and improved technology must have been available to attain more accurate numerical data. It would also put our analysis to the test, if we can identify the right parameters from our current and future datasets in order to generate ROC curves. Additionally, besides the models we have tried, we would conjointly wish to attempt other algorithms such as Ada boost in order to compare results and continue our search for the best model for prediction. The idea of applying other feature selection on the currently used models is also under consideration, such as the Recursive Feature Elimination and the Correlation Heat Map.

REFERENCES

[1] Breast cancer facts and figures 2003-2004 (2003). American Cancer Society.
[2] Stages | Mesothelioma | Cancer Research UK Breast cancer survival statistics September 26, 2017
[3] Pendharkar PC, Rodger JA, Yaverbaum GJ, Herman N, Benner M (1999) Association, statistical, mathematical and neural approaches for mining breast cancer patterns. Expert Systems with Applications 17: 223-232.
[4] Deepsense.ai what is reinforcement learning? The complete guide July 05, 2018
[5] Hacker Noon Absolute Fundamentals of Machine Learning – Hacker Noon January 15, 2018
[6] Furundzic, D.; Djordjevic, M.; Bekic, A.J. Neural networks approach to early breast cancer detection. J. Syst. Archit. 1998, 44, 617–633. [CrossRef]
[7] Floyd, C.E.; Lo, J.Y.; Yun, A.J.; Sullivan, D.C.; Kornguth, P.J. Prediction of breast cancer malignancy using an artificial neural network. Cancer 1994, 74, 2944–2948. [CrossRef]
[8] Fogel, D.B.; Wasson, E.C.; Boughton, E.M. Evolving neural networks for detecting breast cancer. Cancer Lett. (1995), 96, 49–53. [CrossRef]
[9] Fogel, D.B.; Wasson, E.C.; Boughton, E.M.; Porto, V.W.; Angeline, P.J. Linear and neural models for classifying breast masses. IEEE Trans. Med. Imaging (1998), 17, 485–488. [CrossRef] [PubMed]
[10] Setiono, R. Extracting rules from pruned neural networks for breast cancer diagnosis. Ar-tif. Intell. Med. (1996), 8, 37–51.