AN APPLICATION FOR CLASSIFICATION AND PREDICTION OF BREAST CANCER

Mary Mathalin A[#1], Manthra R[*2], Madhuri Devi L[#3]

[1,2,3]*Department of Computer Science and Engineering, St. Joseph's Institute of Technology, OMR Chennai-600119, India*

Abstract - *The most frequently occurring cancer among Indian women is Breast cancer (BC). There is an opportunity of one-half for fatality during a case together of two women diagnosed with BC die within the cases of Indian women. This paper aims to present comparison of 6 Machine Learning (ML) algorithms and techniques commonly used for BC prediction which gives result whether the person has cancer or not in the form of Predicted or Not Predicted. Here the comparison between the accuracies of six algorithms are take place and gives the result with highest accuracy. The Wisconsin Diagnosis BC data set was used as a training set to match the performance of the varied machine learning techniques. Based on the result of performed experiments, the support vector machine shows the highest accuracy. Additionally, we can run this project to any other CPUs using LAN connection.*

Key Words: Pre-Processing, ML, BC, Prediction, Accuracy.

1. INTRODUCTION

Breast cancer is that the most typical reason behind the death of females in large amount around worldwide over the previous couple of decades within the developed, underdeveloped and developing countries. Early detection of carcinoma diseases and continuous supervising of clinicians can scale back the death rate. However, correct detection of carcinoma diseases all told cases and consultation of a patient for twenty-four hours by a doctor isn't obtainable since it needs additional patience, time and experience. For the correct detection of the carcinoma, an economical machine learning technique ought to be used that had been derived from a particular analysis among many Machine Learning algorithms. The diagnosing of carcinoma is completed by classifying the cancer.

Cancers will be either benign or malignant. Malignant cancers are additional harmful than benign. Sadly, not all physicians are professional in characteristic between the benign and tumour and also the classification of tumour cells could take up to a pair of days. Machine Learning will play a necessary role in predicting the tumour connected diseases. Such data, if foretold well ahead, will offer necessary insights to doctors who will then adapt their diagnosing and treatment per patient basis. the various algorithms used are: Support Vector Machine (SVM), Decision tree, Random Forest, Linear Regression (LR), Naive Bayes (NB), Logistic Regression etc.,

2. SYSTEM ARCHITECTURE

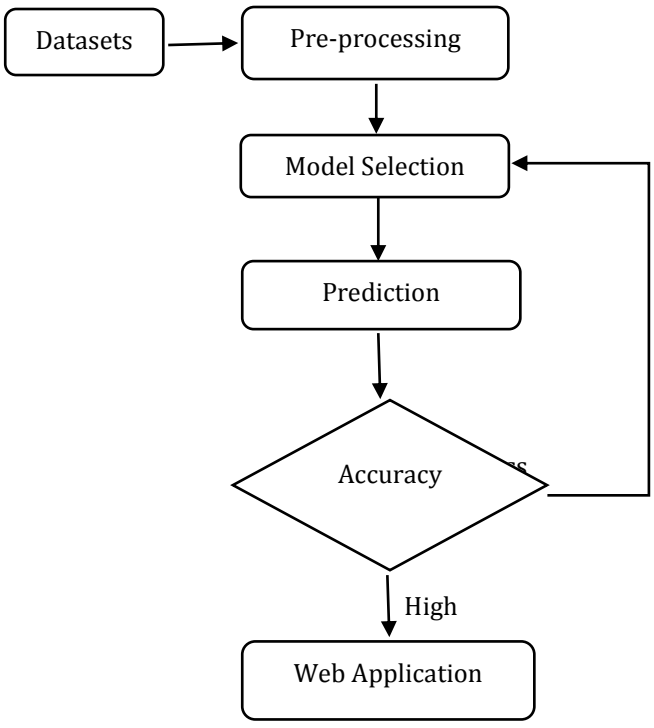

Fig-1: Flow diagram for BC Prediction

2.1 Collection of datasets

We received BC dataset of Wisconsin Breast Cancer diagnosis dataset and used Jupiter notebook as the platform for the reason of coding. Our methodology includes use of supervised getting to know algorithms and classification approach like Decision Tree, Random Forest and Logistic Regression, with Dimensionality Reduction method.

2.2 Pre-Processing

Our dataset may be Incomplete or have a few lacking characteristic values, or having best aggregate information. So, there may be a need to pre-technique our clinical dataset which has main characteristic as id, analysis and other actual valued features that are computed for each cellular nucleus like radius, texture, parameter, smoothness, area, etc. The diagnosis values are variables, it should be converted into numeric values. So, here malignant cells (Positive) are represented as 1 and benign cells (Negative) are represented as 0.

After that the datasets are split into 2. Such as 75% for training and 25% for testing process.

2.3 Model Selection:

This is the most crucial segment in which algorithm selection is completed for the developing system. Data Scientists use various styles of Machine Learning algorithms which may be classified as: supervised getting to know and unsupervised gaining knowledge.

For this Prediction System, we handiest need Supervised Learning

2.4 Algorithm Used:

Supervised Learning is a kind of system in which each input and favoured output records are provided. Input and output facts are labelled for type to offer a learning foundation for future data processing. Supervised systems provide the studying algorithms with known quantities to support future judgments.

We have different types of classification algorithms in ML.

1.Decision Tree Algorithm

2. Random Forest Classification

3. Logistic Regression

4. SVM

5.Naive Bays

6.LR

2.5 Experimental Result

After using 6 algorithms we have to calculate the accuracy of each algorithms and the result will be displayed based on high accuracy of the algorithms. To test the right prediction, we have to check confusion matrix object and upload the expected outcomes diagonally so that you can be range of accurate prediction and then divide via the use of popular quantity of predictions.

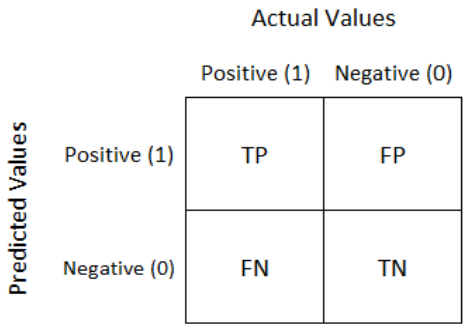

Fig-2: Confusion-matrix

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

TP-True Positive

TN-True Negative

FP-False Positive

FN-False Negative

After using the algorithms we've got accuracies with special fashions:

- Logistic Regression —85.8%
- SVM —99.78%
- Decision tree —89.8%
- Random Forest — 92.6%
- Naïve Bays — 90.6%
- Linear Regression — 91.6%

So ultimately, we've built our type model and we see that Support Vector Machines algorithm gives the highest accuracy.

3. WEB DEVELOPMENT MODULES

Flask is a web framework for Python, it provides functionality for building web applications, including managing HTTP requests and rendering templates. Flask application to create our API. Flask's frame work is more explicit and is easier to learn because it has less Application. In this project, flask framework plays a vital role in predicting the breast cancer and will suggest the hospitals.

- In this module, the user can give the necessary inputs like name, mobile number, email-id, password, confirm password in order to register.
- After that the user can give a user name password in order to logging in. Once the user login the page, OTP will send to the registered mail id.
- The prediction page will be displayed only the user enter the correct OTP. The prediction page consists of input data like area mean, radius mean etc.
- Above predictions can show the patient/person has been affected by the disease of breast cancer. The result will be displayed as BC predicted or BC not predicted.

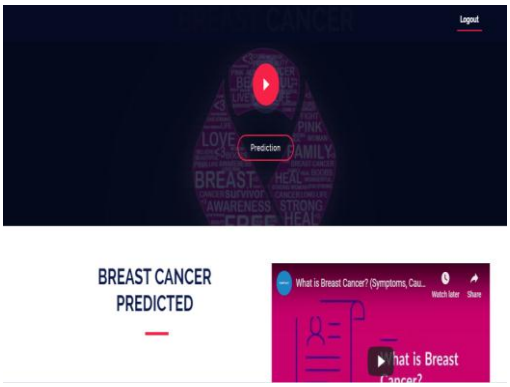

Fig-3: BC Predicted Page

4. CONCLUSION

Medical dataset can not only be classified with the previously mentioned algorithms from machine learning, there are many algorithms and techniques which may perform better than these. Production of accurate classifier which perform efficiently for medicinal application is the main challenge we face in machine learning. Six algorithms were implemented namely NB, SVM, Random Forest, Logistic Regression, Decision Tree and LR on Breast Cancer dataset. Our main aim is to discover the algorithm which performs faster, accurate and efficiently. SVM surpasses all the other algorithms with an accuracy of 99.78%. In conclusion, the SVM algorithm achieves the highest accuracy which might be the best choice of algorithm for this problem and prediction of disease.

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

[8]. Rasool Fakoor, Faisal Ladhak, Azade Nazi, Manfred Huber. Using deep learning to enhance cancer diagnosis and classification. 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013

[9]. Cancer Statistics, 2016. CA: A Cancer Journal for Clinicians

[10]. UCI Machine Learning Repository: Breast Cancer Wisconsin Dataset.