Mini project Report On

“A Breast Cancer Prediction “ Project Report submitted to

# JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY ANANTAPUR.

In Partial fulfillment of the requirements for the Award of the degree of

# BACHELOR OF TECHNOLOGY IN

**COMPUTER SCIENCE AND ENGINEERING**

BY

**S SAMEER 21125A0520**

Under the Guidance of Mrs. K Pujitha Assistant Professor



Department of Computer Science and Engineering

SREE VIDYANIKETHAN ENGINEERING COLLEGE

(Affiliated to JNTUA, Anantapuramu) Sree Sainath Nagar, Tirupati – 517 102

2022-2023

# TABLE OF CONTENTS

|  |  |
| --- | --- |
| **NAME OF THE CONTENT** | **PageNo** |
| **PROBLEM STATEMENT & ABSTRACT** | **1** |
| **INTRODUCTION** | **1-2** |
| **METHODOLOGY** | **3** |
| **MACHINE LEARNING ALGORITHMS** | **4-5** |
| **RESULTS** | **6-7** |
| **CONCLUSION** | **8** |

v

**PROBLEM STATEMENT**

The Prediciton of Breast Cancer using Machine Learning algorithms(ML) are Logistic Regression(LR), Decision Tree and Random Forest.

## ABSTRACT

Each year number of deaths is increasing extremely because of breast cancer. It is the most frequent type of all cancers and the major cause of death in women worldwide. The lack of robust prognosis models results in difficulty for doctors to prepare a treatment plan that many prolong patient survival time. Hence with requirement of time is to develop the technique which gives minimum error to increase accuracy. Three algorithms LOGISTIC REGRESSION(LR), RANDOM FOREST(RF) and DECISION TREE algorithm which

predict the breast cancer outcome have been compared in the paper using different datasets. All experiments are executed within a simulation environment and conducted in JUPYTER NOTEBOOK platform. Aim of research categorises in three domains. First domain is rediction of cancer before diagnosis, Second domain is predciton of diagnosis and treatement and Third domain focuses on outcome of different technique and suitable technique can be used depending upon requirment. This research is carried out to predict the accuracy. The future research can be carried out to predict the other different parameters and breast cancer research can be categorises on basis of other parameters.

## INTRODUCTION

Breast cancer has now overtaken lung cancer as the most commonly diagnosed cancer in women worldwide, according to statistics released by the International Agency for Research on Cancer(IARC) in December 2020. In the past two decades, the overall number of people diagnosed with cancer nearly doubled, from an estimated 10 million in 2000 to 19.3 million in 2020. Today, one in 5 people worldwide will develop cancer during teir lifetime. Projections suggest that the number of people being diagnosed with cancer will increase still further in the coming years, and will be nearly 50% higher in 2040 than in 2020. The number of cancer deaths has also inceased, from 6.2 million in 2000 to 10 million in 2020. More than one in six deaths is due to cancer. This reinforces the need to invest in both the fight against cancer and cancer prevention. The successful introduction of information and communication technologies in medical practice is an important stake in the renovation of the health system in Big Data has made a big change in BI by analyzing the large amount of unstructured, hetrogeneous, non-standard and incomplete healthcare data. It does not only forecast but also helps in decision making and is increasingly noticed as breakthrough in ongoing advancement with the goal is to improve the quality of patient care and reduces the healthcare cost. Data mining algorithms applied in healthcare

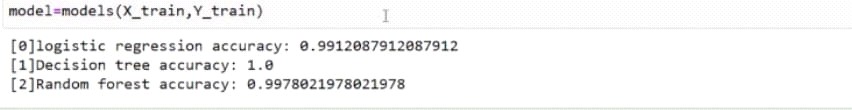
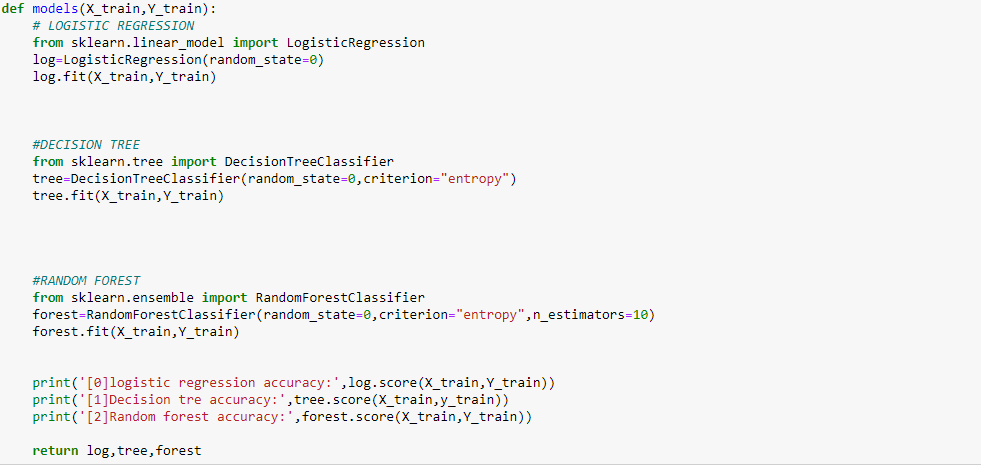
1

industry play a significant role due to their high performance in predicting, diagnosis of the diseases, reducing costs of medicine, making real time decision to save people’s lives. The Most common Data mining 4odelling goals are classification and prediction which uses several algorithms for the prediction of breast cancer. Mainly gives a comparison between the performance of Three classifiers: Random Forest, Logistic Regression, Decision tree which according to research community are among the most influential datamining algorithms and among the top 10 data mining algorithms. Our objective is to predict and diagnosis breastcancer, using machine-learning algorithms, and find out the most effective based on the performance of each classifier in terms of confusion matrix, accuracy, precision and sensitivity.

## 1. JUPYTER NOTEBOOK OVERVIEW

* + 1. **General Requirements**
       - JUPYTER NOTEBOOK
       - Import necessary libraries for the project
       - Datasets from the Sources

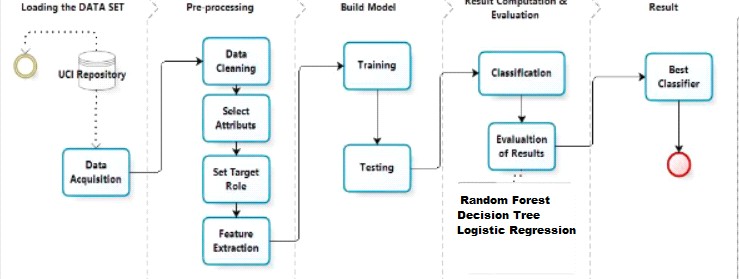
## Expected Values from the JUPYTER NOTEBOOK



2

## METHODOLOGY

The main objective of our experiment is to identify the effective and predictive algorithm for the detection of breast cancer, therefore we applied machine learning classifiers RANDOM FORESTS, LOGISTIC REGRESSION and DECISION TREE on Breast Cancer Wisconsin Diagnostic dataset and evaluate the results obtained to define which model provides a higher accuracy.



Our methodology begins with data acquisition followed by pre-processing, which contains four steps viz: datacleaning, select attributes, set target Role and features extraction. The prepared data is used to build machine learning algorithms that can predict the breast cancer for a new set of measurements. To evaluate the algorithms performances, we show the model new data for which we have labels. This is usually done by splitting the labeled data we have collected into two parts whit Train\_test\_split method. 75% of the data is used to build our machine learning model, and is called the training data or training set. 25% of the data will be used to access how well the model works and is called test data, test set. After testing the models we compare the obtained results to select the algorithm that provides the high accuracy and identify the most predictive algorithm for the detection of breast cancer.

3

## MACHINE LEARNING ALGORITHMS

In our project, the predictive analysis of the machine learning algorithms is achieved. The machine learning algorithms applied in our project are:

## RANDOM FORESTS

Random forests or random decision forests are an ensemble method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set.

## LOGISTIC REGRESSION

Logistic regression is a very powerful modeling tool, is a generalization of linear regression.Logistic Regression is used to assess the likelihood of a disease or health condition as a function of a risk factor (and covariates). Both simple and multiple logistic regression , assess the association between independent variable(s) (Xi) -- sometimes called exposure or predictor variables — and a dichotomous dependent variable(Y) -- sometimes called the outcome or response variable. It is used primarily for predicting binary or multiclass dependent variables.

## DECISION TREE

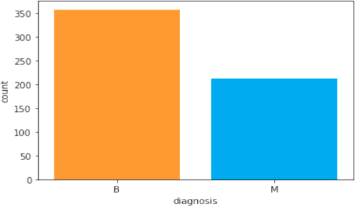
Decision Tree is a predictive modeling tool that can be applied across many areas. It can be constructed by an algorithmic approach that can split the dataset in different ways based on different conditions.

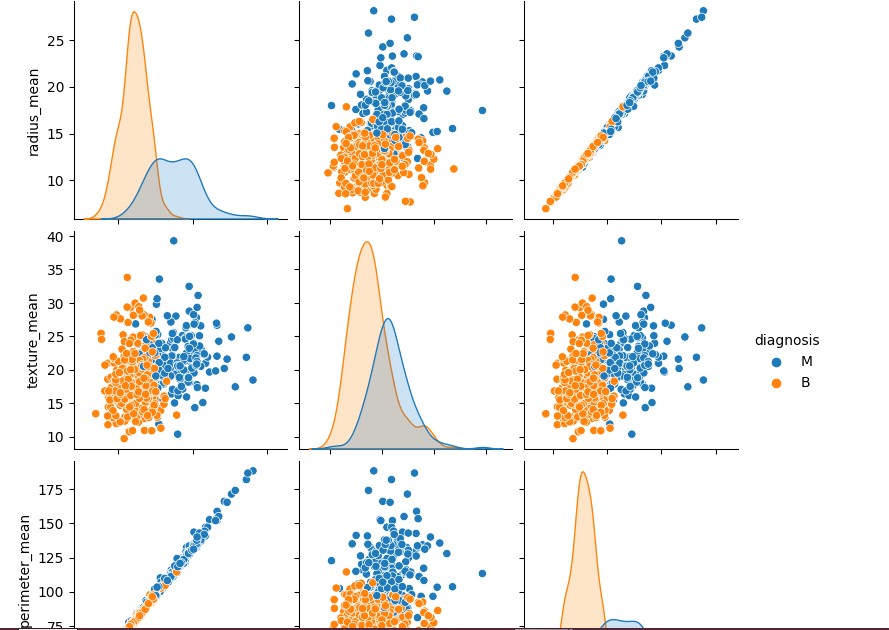
## DATASET ACQUISITION

In our study, we use Breast Cancer Wisconsin Diagnostic dataset from University of Wisconsin Hospitals Madison Breast Cancer Database . The features of dataset are computed from a digitized image of a breast cancer sample obtained from fine-needle aspirate (FNA). The characteristics of the cell nuclei present in the image are determined from these features. Breast Cancer Wisconsin Diagnostic has 569 instances (Benign: 357 Malignant:212), 2 classes (62.74% benign and 37.26% malignant), and 11 integer-valued attributes (-Id -Diagnosis -Radius -Texture -Area -Perimeter -Smoothness -Compactness

-Concavity -Concave points -Symmetry -Fractal dimension).

4

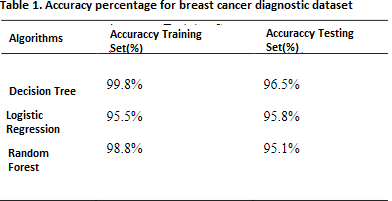




5

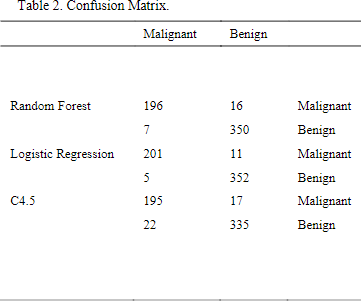
## RESULTS

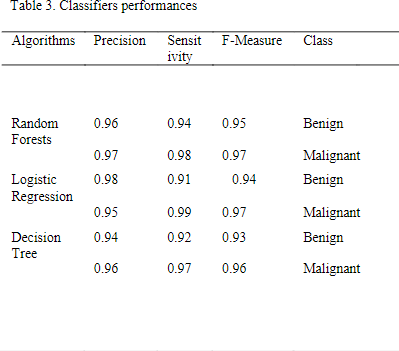
After applying Machine Learning Algorithms on Breast Cancer Wisconsin Diagnostic dataset.We used Confusion Matrix, Accuracy, Precision, Sensitivity, F1 Score, AUC as performance metrics to evaluate and compare the models and identify the best algortihm for the brest cancer Prediction. Confusion Matrix is the way to measure the performance of a classification problem where the output can be of two or more type of classes. A confusion matrix is a table with two dimensions viz. “Actual” and “Predicted” and furthermore, both the dimensions have “True Positives (TP)”, “True Negatives (TN)”, “False Positives (FP)”, and “False Negatives (FN)”. Accuracy is mostcommon performance metric for classification algorithms. It defined as the number of correct predictions made as a ratio of all predictions made. Precision, used in document retrievals, may be defined as the number of correct documents returned by our ML model. Sensitivity may be defined as the number of positives returned by your ML model. F1 score gives us the harmonic mean of precision and Sensitivity. Mathematically, F1 score is the weighted average of the precision and Sensitivity.



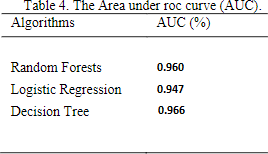
Since confusion matrices are a useful way to assess the classifier, each row in Table 2 represents the rates in anactual class while each column displays the predictions.

6



Table 3 present the calculated performance measures ofclassification models based on confusion matrix results, precision sensitivity f1 score for benign and malignant.

The Decision Tree has the highest AUC score 0.96% while the AUC score of Logistic Regression 0.94% is the lowest as shown in table 4.



7

## CONCLUSION

The report concluded an extensive analysis of the performance of these algorithms using a specific dataset. The evaluation metrics such as accuracy, precision, F1-score were used to compare the performance of these algorithms.It was found that the Decision Tree Algorithm performed the best in terms of overall performances, achieving the highest accuracy, precsion and F1-score. The Random Forests Algorithm also performed well, but slightly behind the Decision Tree. The Logistic Regression algorithm performed reasonably well, but not as well as Decision Tree and Random Forests Algorithm.This report has shown that the use of Machine Learning Algorithms, specifically Decision Tree Algorithm, can be a powerful tool for the predicition of Breast Cancer dieseases. The Decision Tree Algorithm perfomed the best among all two algorithms considered in this report, achieving high levels of accuracy, precision, and F1 score. It is also relatively easy to interpret and computationally efficient. Based on this research, it is recommended to use the Decision Tree Algorithm for predicating Breast Cancer Diseases in the future.

8