**SHAH AND ANCHOR KUTTCHI ENGINEERING COLLEGE, MUMBAI**

**2019-2020**



Department of Information Technology

Guide

Prof. Ashwini Deshmukh

**Breast Cancer Prediction with various ML Models**

By

|  |  |  |
| --- | --- | --- |
| Name of the Student | Class | Roll No |
| Naivedh Shah | TE-5 | 57 |
| Shaikh Sahil | TE-5 | 62 |
| Zeeshan Shaikh | TE-5 | 63 |
| Harshit Shiroiya | TE-5 | 65 |
| Mansi Thakkar | TE-5 | 68 |
| Jay Visaria | TE-5 | 72 |

**Problem Statement**

A cancer is a disease caused by uncontrolled division of abnormal cells in a part of the body. Breast Cancer is the most leading malignancy affecting 2.1 million women each year which leads to greatest number of deaths among women. The main objective is to achieve early diagnosis of the breast cancer patients. Early treatment not only helps to cure cancer but also helps in prevention of its recurrence. Breast cancer is fatal in under half of all cases and is the leading cause of death from cancer in women, accounting for 16% of all cancer deaths worldwide. Hence, this analysis becomes very helpful to doctor for diagnosis of breast cancer and also helpful to patients for early treatment.

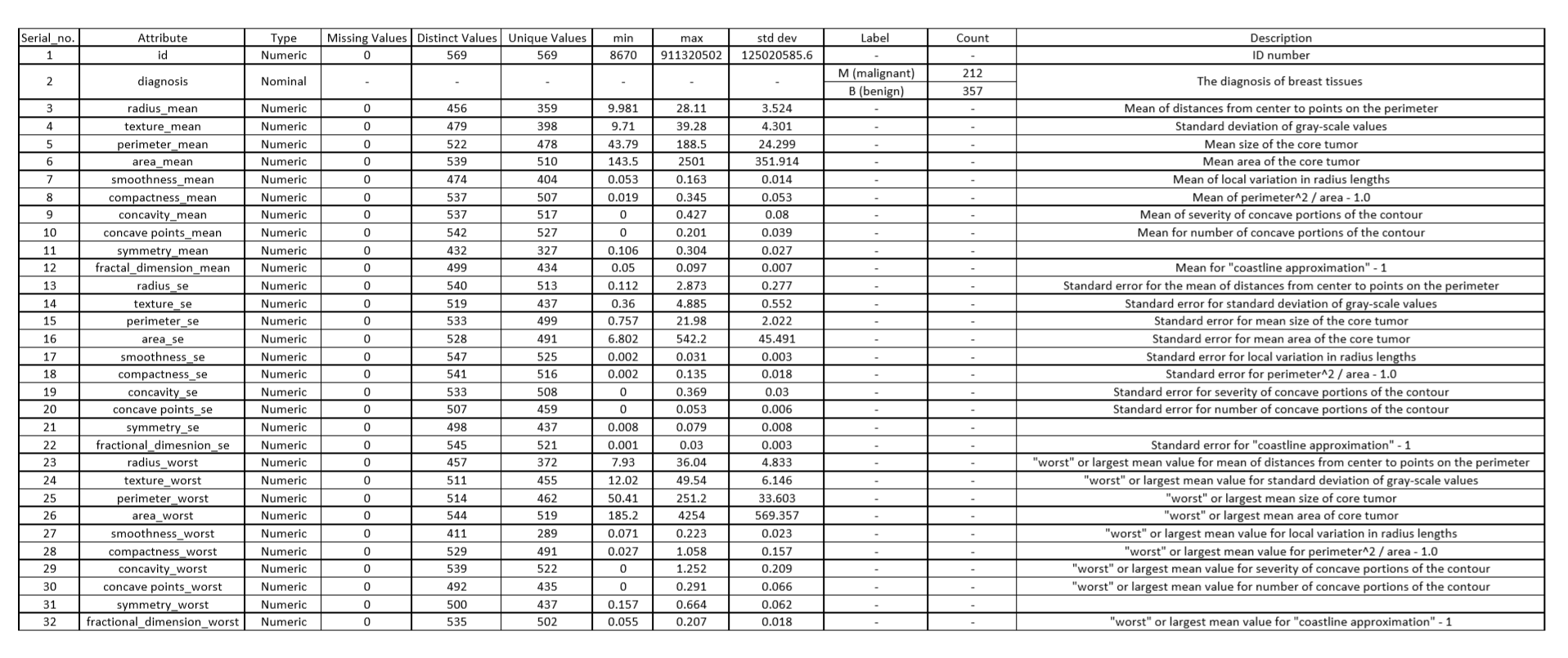
**Data Exploration**

Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a data set, including its size, accuracy, initial patterns in the data and other attributes. It is commonly conducted by data analysts using visual analytics tools, but it can also be done in more advanced statistical software

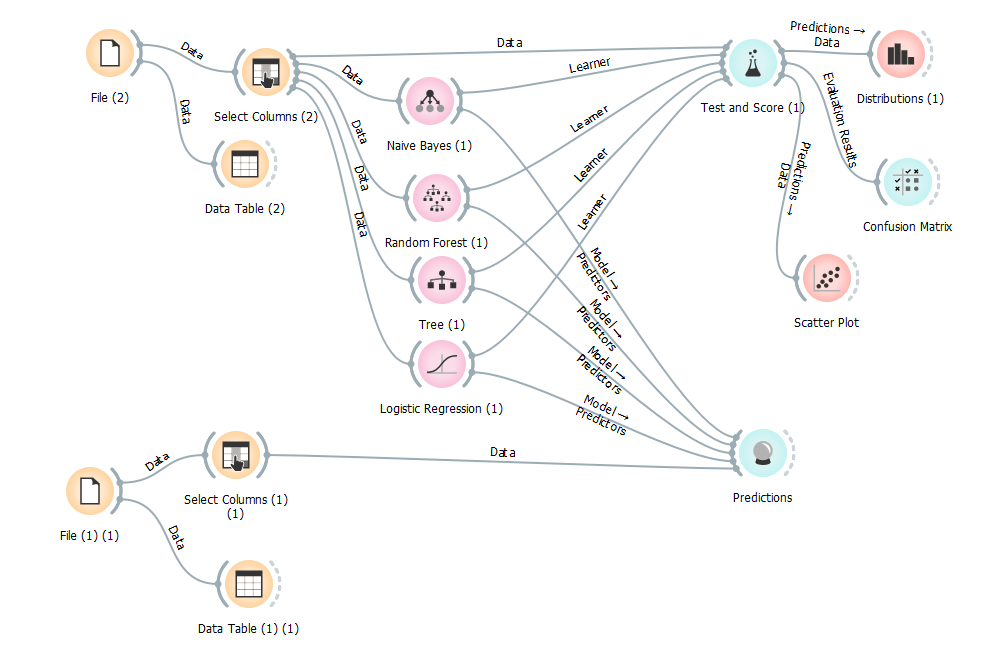
**The Role of Data Exploration**

Before it can conduct analysis on data collected by multiple data sources and stored in data warehouses, an organization must know how many cases are in a data set, what variables are included, how many missing values there are and what general hypotheses the data is likely to support. An initial exploration of the data set can help answer these questions by familiarizing analysts with the data with which they are working.

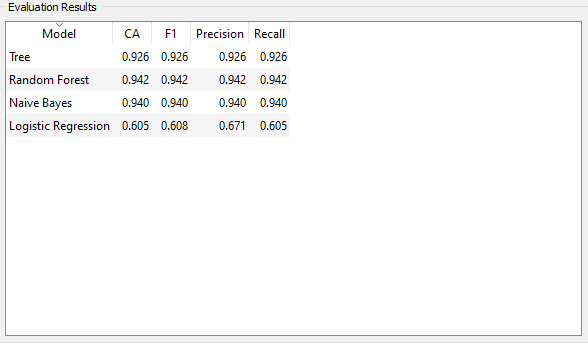
Once data exploration has uncovered the relationships between the different variables, organizations can continue the data mining process by creating and deploying data models to take action on the insights gained.

**Data Exploration Sheet**

**Implementation in Orange Tool**

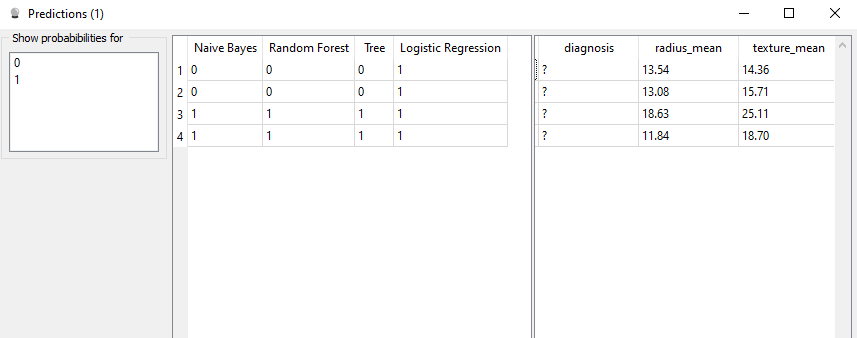
****

**Evaluation Result**



Visualization techniques with different models were used to see the output. We are getting the accuracy of ***94.2%*** with the **“Random Forest”** model.

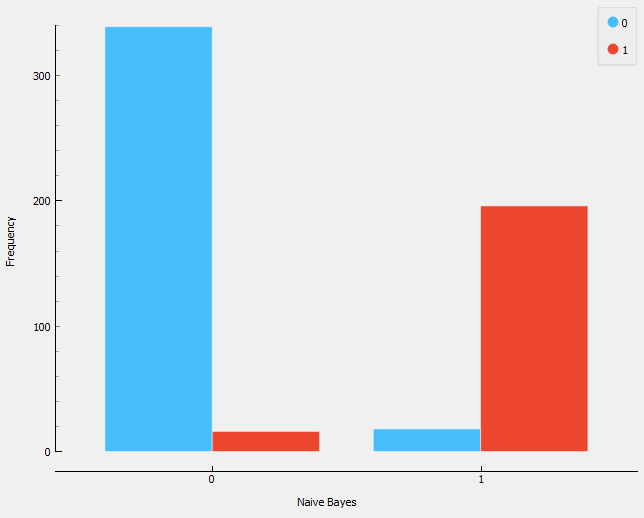
**Predicting on Test Data**



**Visualization**

0 Benign

1 Malignant

****

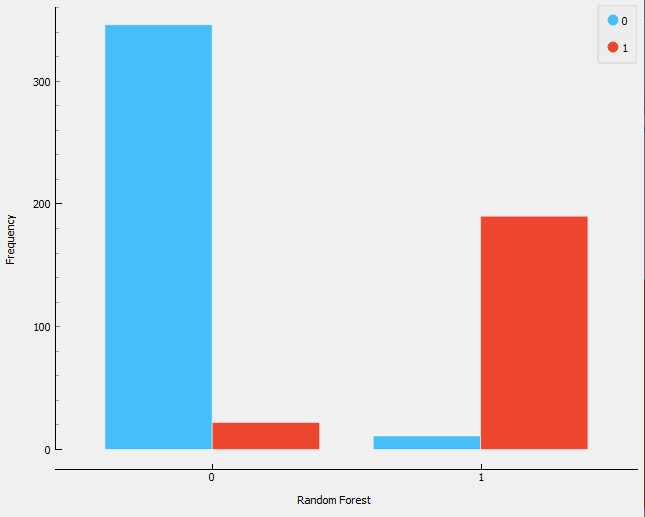
Naive Bayes

Benign Malignant

(62.39% overall) (37.61% overall)

Benign(59.58%) Malignant(2.81%) Benign(3.16%) Malignant(34.45%)

Correctly Classified Wrongly Classified Wrongly Classified Correctly Classified

****

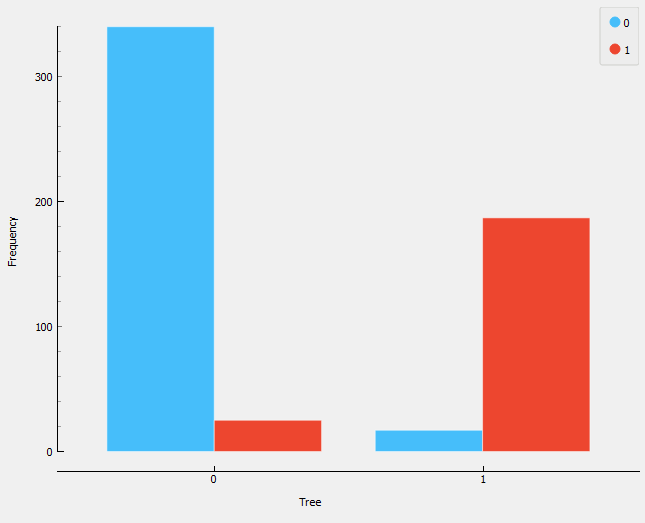
Random Forest

Benign Malignant

(67.47% overall) (35.33% overall)

Benign(60.81%) Malignant(3.87%) Benign(1.93%) Malignant(33.39%)

Correctly Classified Wrongly Classified Wrongly Classified Correctly Classified



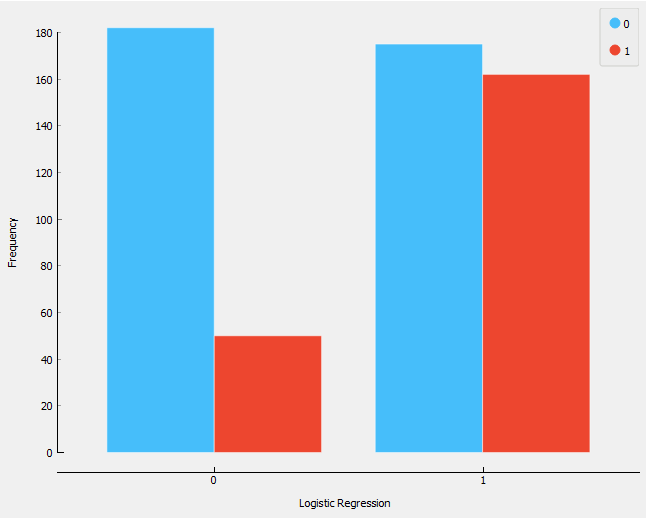
Decision Tree

Benign Malignant

(64.15% overall) (35.85% overall)

Benign(59.75%) Malignant(4.39%) Benign(2.99%) Malignant(32.86%)

Correctly Classified Wrongly Classified Wrongly Classified Correctly Classified



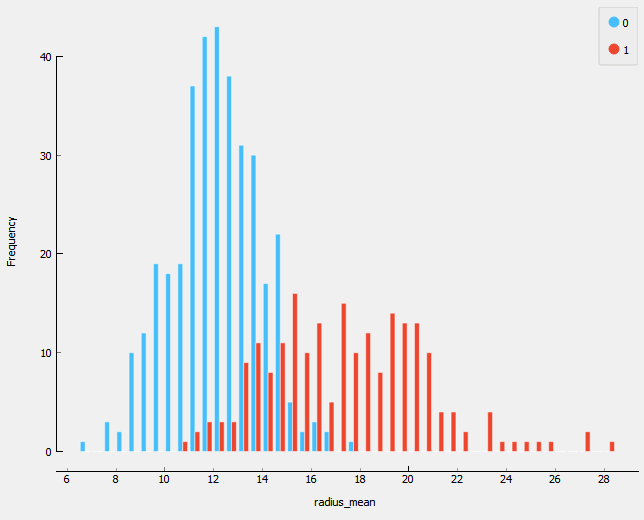
Logistic Regression

Benign Malignant

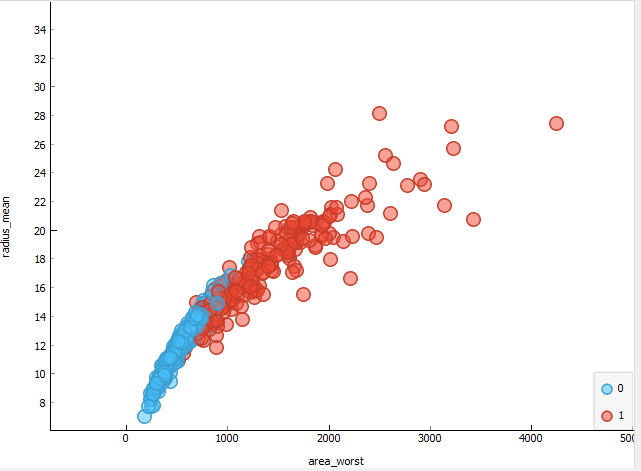
(40.77% overall) (59.23% overall)

Benign(31.99%) Malignant(8.79%) Benign(30.76%) Malignant(28.47%)

Correctly Classified Wrongly Classified Wrongly Classified Correctly Classified

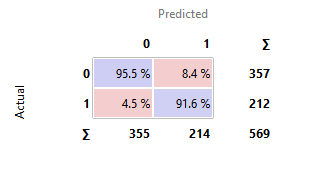
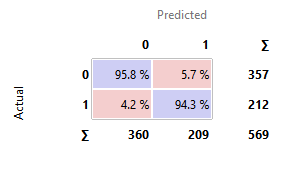
****

radius\_mean vs Frequency

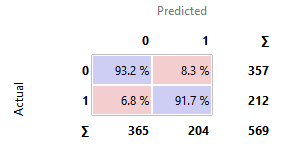
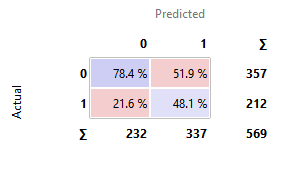


area\_worst vs radius\_mean

**Confusion Matrix**

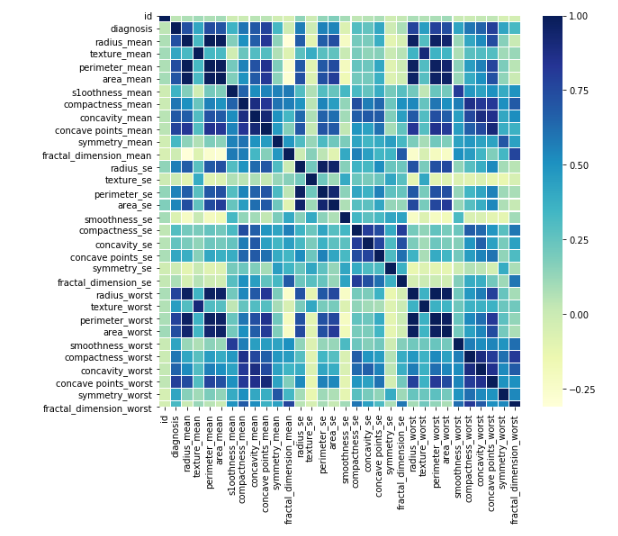


Naive Bayes Random Forest

****

Decision Tree Logistic Regression

**Correlation Matrix**

****

**BI Decision**

Dataset used by us was ***“Breast Cancer”*** and from the analysis we know that each and every factor in the dataset is very crucial to decide whether cancer detected is mild or not, also it will help the doctors for the diagnosis.

According to the visualisation that we have obtained, most of the breast cancer found in women’s is **“Benign”** that is, mild in nature.

From the above graph and histograms ,we can interpret and understand that the factors such as radius ,perimeter and area play a major role in determining whether the cancer is malignant or benign.

Once a patient is detected with breast cancer the size of the tumour can help them to prevent it from getting worse. According to the analysis it can be concluded that the detected cancer must be cured in the initial stages that is when the radius of the tumour is “**less than 12units”** because after that the condition could get worse and the state of the tumour becomes “**Malignant”**. Hence, the algorithm can help doctor for diagnosis of breast cancer and to provide the patients with early treatment.