Assignment 3

We are using the dataset to predict between two categories and the data is discrete, so this is a classification problem. We have to classify whether the cancer is malignant or benign and we will use logistic regression for that.

Observations from the Data Quality Check:-

1. We have to predict the diagnosis column from the dataset.

2. We can remove the id and Unnamed: 32 columns because it doesn't contain any relevant information.

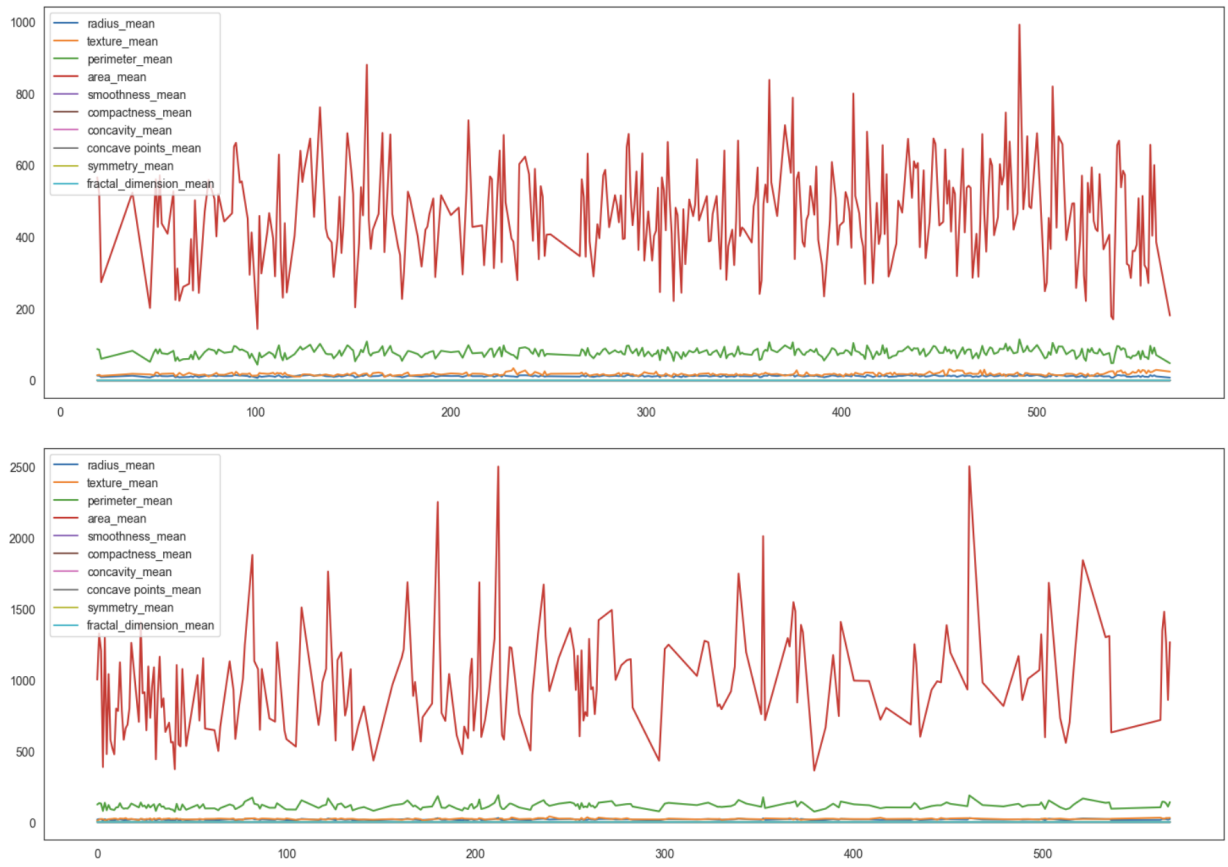
3. Here, we don't have a 0 null value, and all other features are in float64 except the diagnosis.

4. Since the majority of the components have variances that are close to zero, variable information is poor for analyzing the data.

5. Most of the features are positively skewed.

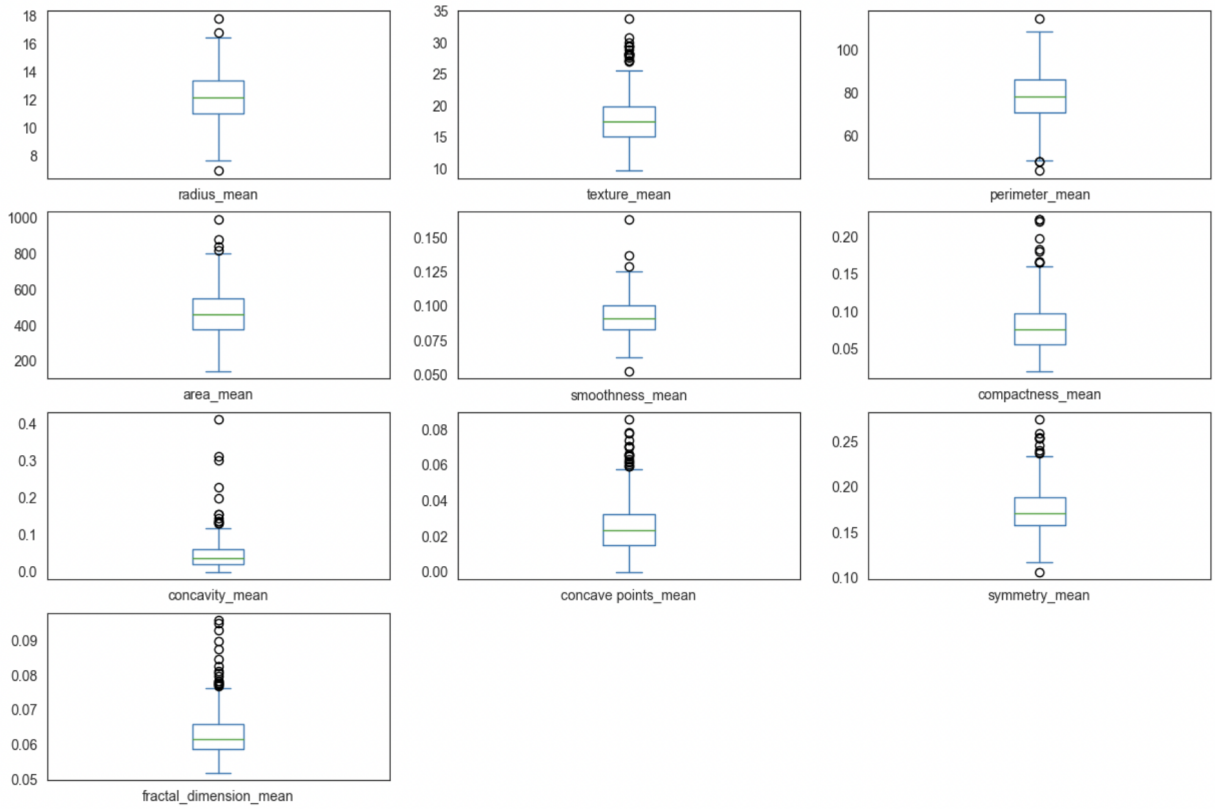
Data Visualisation:-

We are interested mainly in the mean values of these features, so we will segregate those features, in order to make work easier and the code more readable. We will group the mean value features and plot the graph.

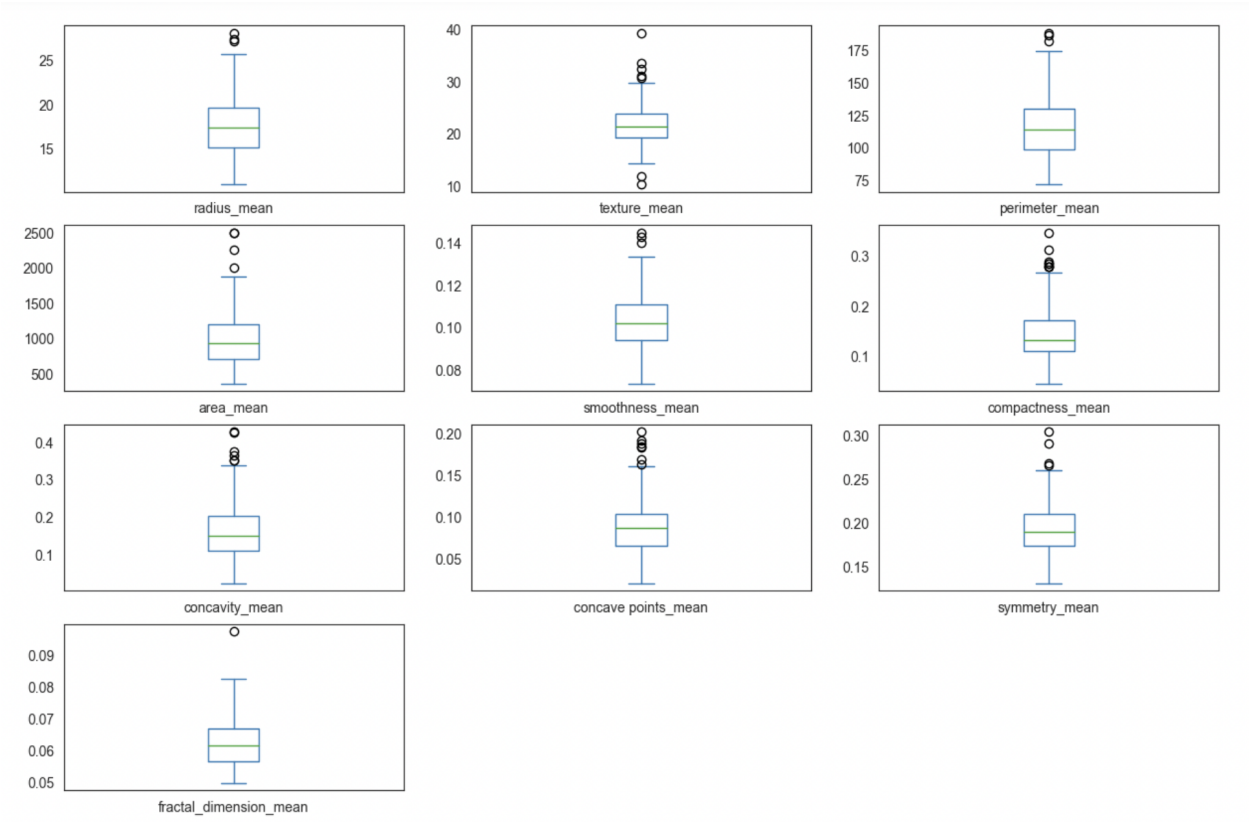


The plot of mean features for both Benign and Malignant

Box Plot:-



Box Plot for Benign



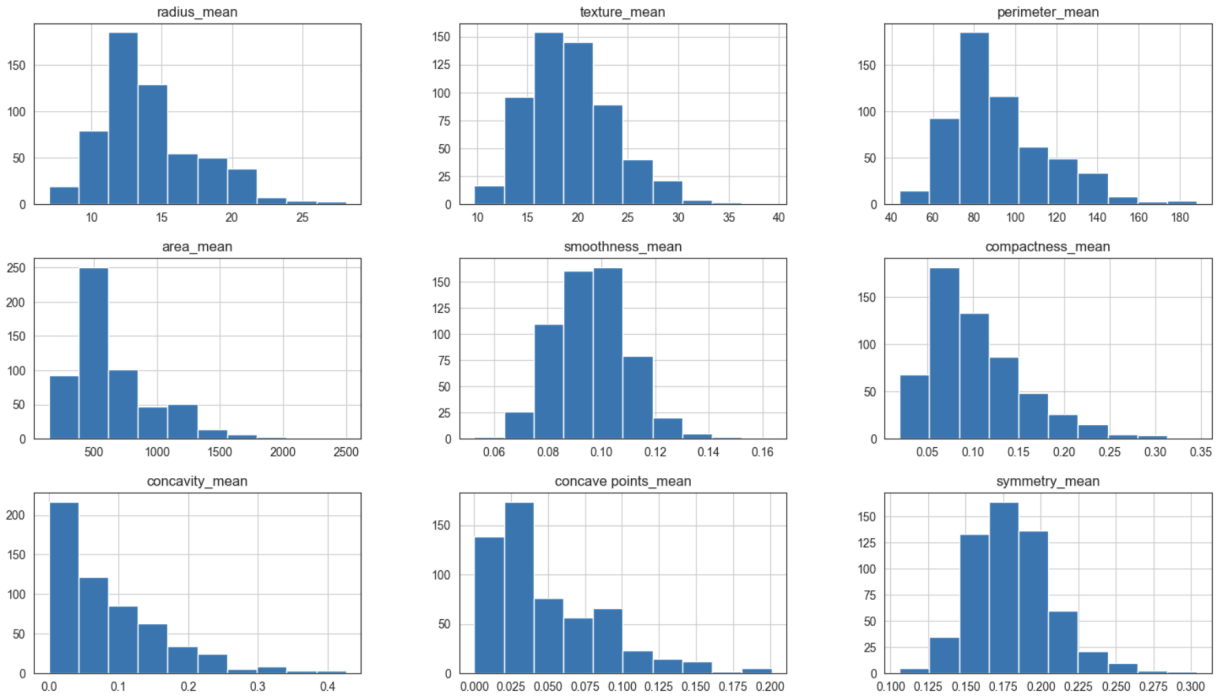
Box Plot for Malignant

Observations of Box plot:-

1. We can observe that the characteristics of perimeter, radius, area, concavity, and compactness may be distributed exponentially.

2. we can see that the symmetry, smoothness, and texture characteristics may have a Gaussian or nearly Gaussian distribution. This is intriguing because the input variables are typically assumed to have a Gaussian univariate distribution by many machine learning approaches.

Histogram:-

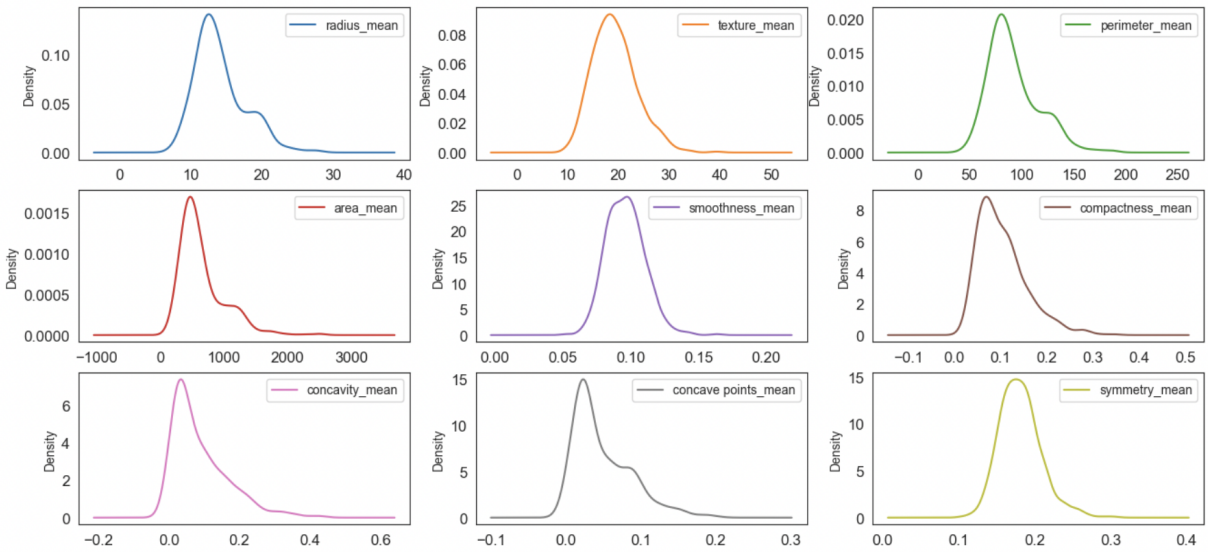


Observations of the histogram:-

1. We can see that perhaps the attributes concavity and concavity_point may have an exponential distribution.

2. We can also see that perhaps the texture, smooth, and symmetry attributes may have a Gaussian or nearly Gaussian distribution.

Density plot:-

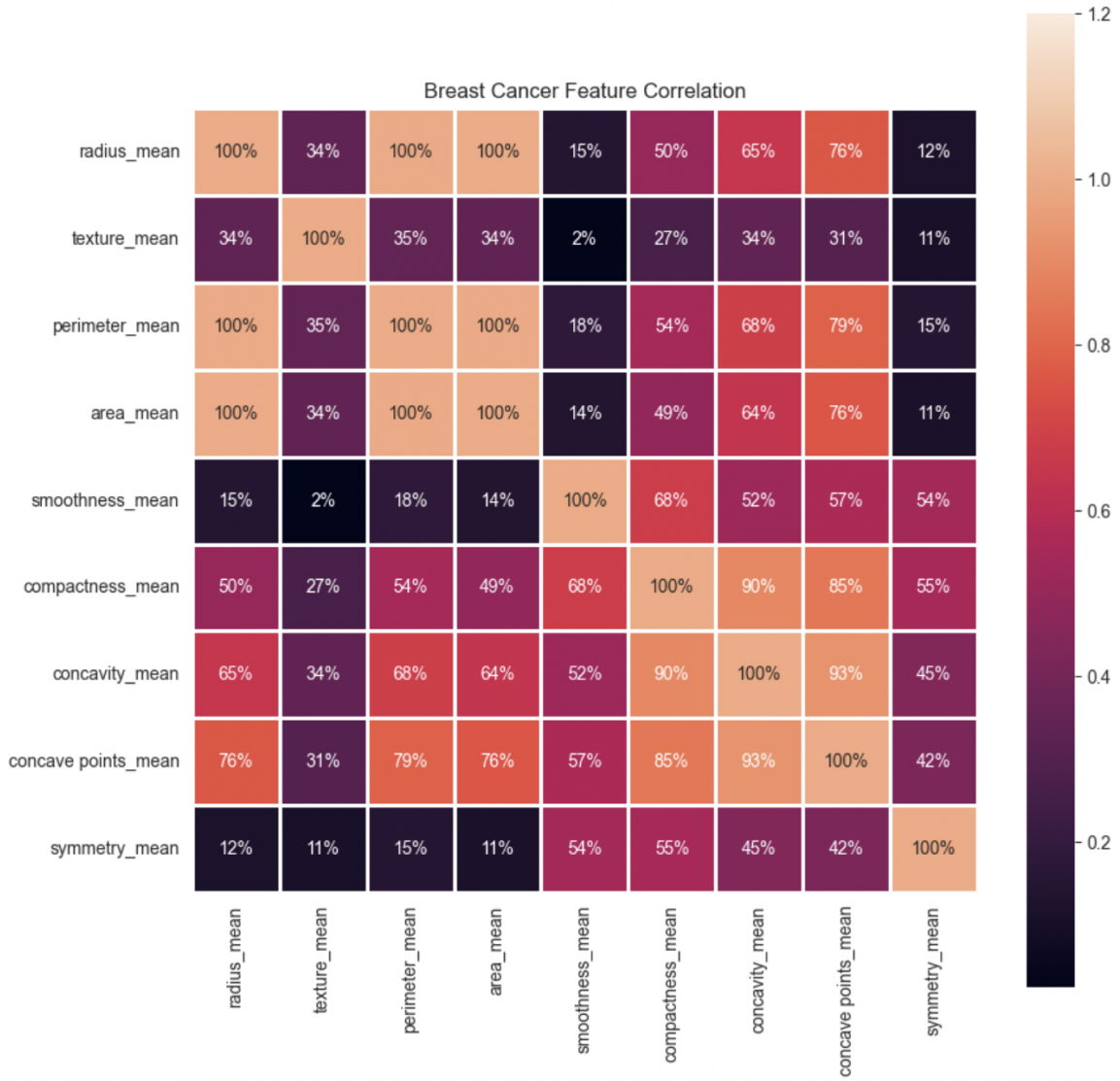


Observations of Density plot:-

1. We can see that perhaps the attributes perimeter, radius, area, concavity, and compactness may have an exponential distribution.

2. We can also see that perhaps the texture, smooth and symmetry attributes may have a Gaussian or nearly Gaussian distribution.

Heatmap:-



Heatmap of correlation matrix which calculates the correlation between each pair of features

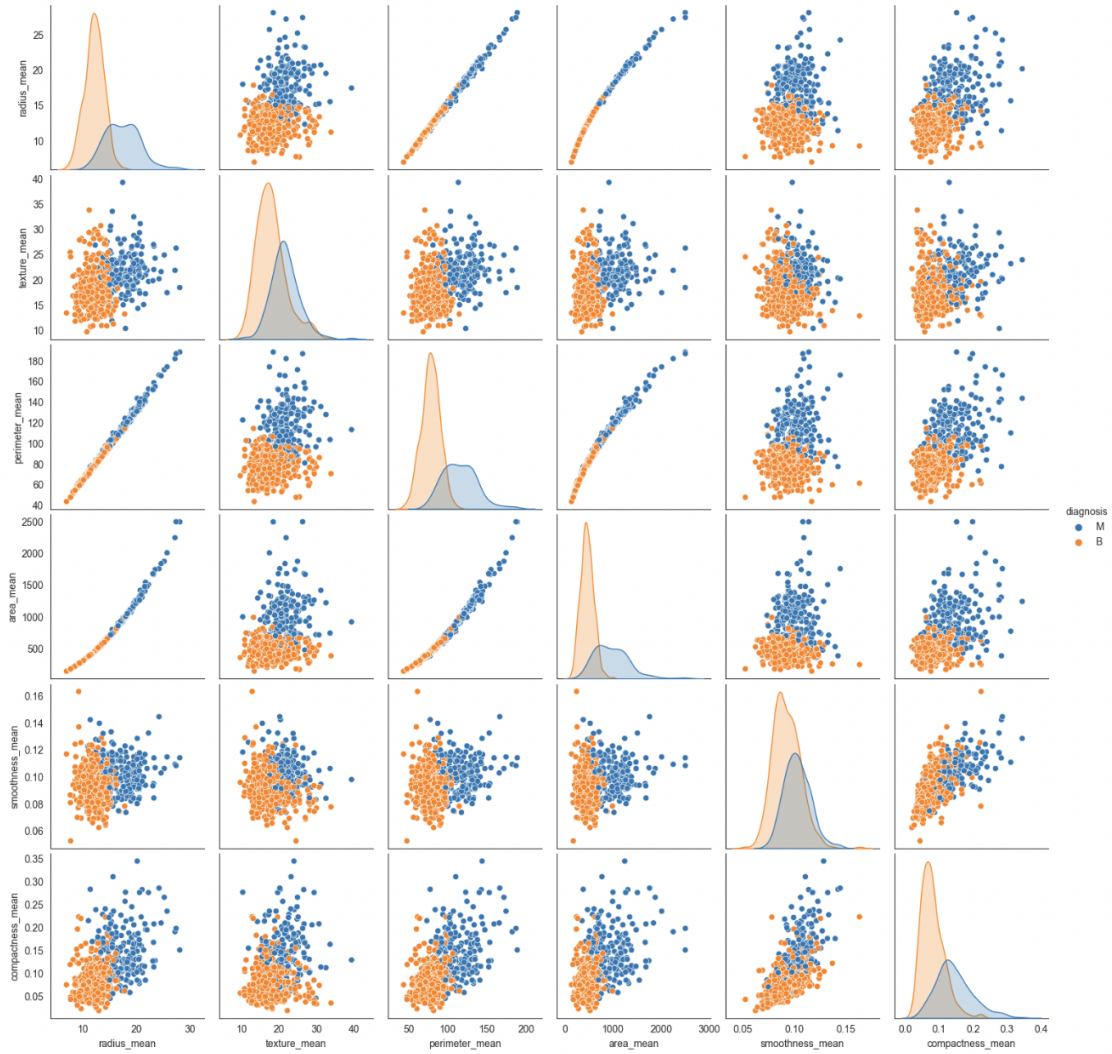
We can see that the mean parameters have a strong positive correlation with values between 0 and 0.75.

1. The mean area of the tissue nucleus has a high positive association with the mean values of radius and parameter.

2. A few parameters have fairly positive correlations (r between 0.5 and 0.75), such as concavity and area, concavity and perimeter, etc.

3. Similar to this, we observe a substantial negative association between fractal_dimension and the mean values of the radius, texture, and parameter.

Seaborn pairplot:-



Observations:-

1. Cancer can be categorized using the average values of the cell's radius, perimeter, area, and compactness spots. These metrics' larger values frequently exhibit a link with malignant tumors.

2. mean values of texture and smoothness do not show a particular preference for one diagnosis over the other.

Feature Engineering:-

1. Features "id" and "Unnamed: 32" are not useful
2. Label Encoding:- After encoding the class labels(diagnosis) in an array y, the malignant tumors are now represented as class 1(i.e presence of cancer cells) and the benign tumors are represented as class 0 (i.e no cancer cells detection), respectively

Compare the model performance using different features:-

Based on the correlation matrix I selected the following features to predict the output.

features_selection = ['radius_mean', 'perimeter_mean', 'area_mean', 'concavity_mean', 'concave points_mean']

Accuracy from the above features was 91%

After that, I train the model for all features excluding id and Unnamed: 32

Performed the Target Feature separation, standardization, and Splitting of the data.

Got below accuracy after the training:-

	precision	recall	f1-score	support
0	0.98	0.97	0.98	108
1	0.95	0.97	0.96	63
accuracy			0.97	171
macro avg	0.97	0.97	0.97	171
weighted avg	0.97	0.97	0.97	171

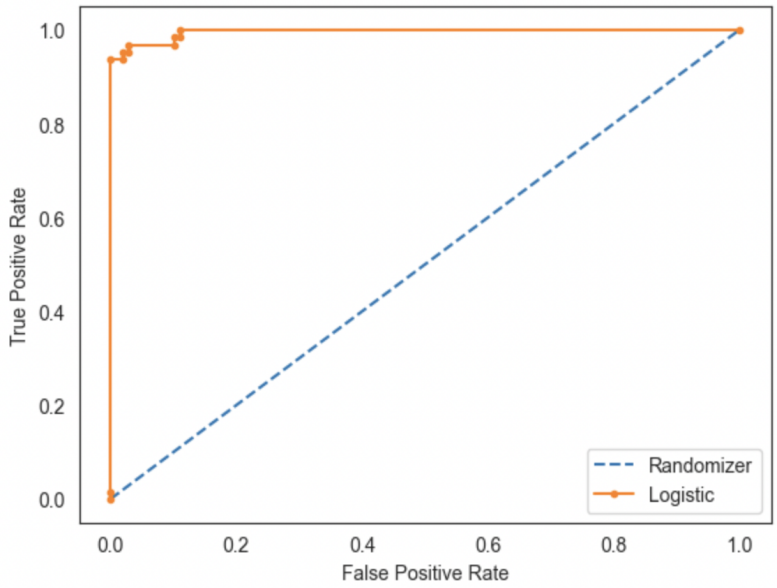
```
Accuracy: 0.9707602339181286
Precision: 0.953125
Recall: 0.9682539682539683
0.9899497487437185
0.9707602339181286
```

Confusion matrix:-

[[105 3]
 [2 61]]

Error Analysis using ROC and AUC:-

```
Randomized Predictions: ROC AUC=0.500
Logistic Regression Classifier: ROC AUC=0.996
```



Conclusion

Score 97% accuracy.