|  |
| --- |
| **Breast Câncer Classification** |
| **Progress** |
| **Rabi Kumar Singh** |

**Analysis Plan**

**Step-1**: Reading the CSV files, dropping the unnecessary columns, getting the information of features of the dataset,

**Step-2**: Identifying the explanatory and response variable. Separate the categorical and numerical features of the dataset. Find Nominal or ordinal features among the categorical dataset.

**Step-3**: Check for balance and imbalance of the dataset, EDA for response variable, perform the Up or down sampling in order to make the dataset balanced.

**Step-4**: Finding the missing values, impute the missing values based on the respective distribution of each of the features or remove the row based on the percentage of missing values.

**Step-5**: Counts the different features of the dataset, Find the correlation among the numerical features among themselves.

**Step-6**: If the correlation is very high, find the multicollinearity among the features. To find the multicollinearity Calculate VIF (Variance Inflation Factor). If VIF>10, remove those features, the only way to remove those features

**Step-7**: Label encoding the response variables,

**Step-8**: Splitting the dataset into train and test.

**Step-9**: Feeding the training dataset into the Logistic Regression model, perform cross validation , find the respective metrics like ROC, confusion matrix, F1 score, recall etc.

**Step-10**: Perform the same model on test dataset and get the corresponding metrics.

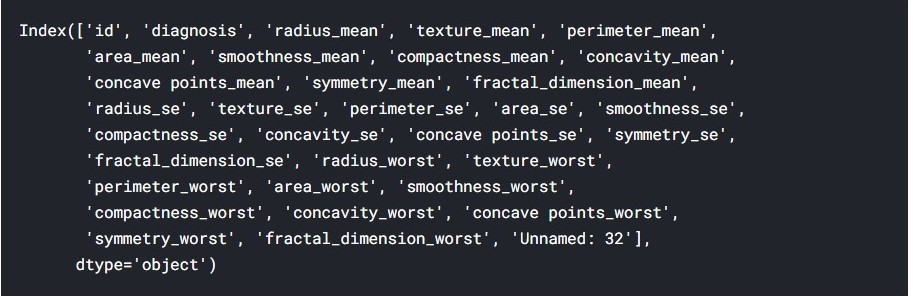
**Step-11**: Use the entire feature to get test error and other metrics.

**Step-12**: Perform the backward subset selection in order to find the important features, and then those features feed into the model and repeat the step-9

**My work**:- <https://www.kaggle.com/jurk06/breat-cancer-1>

**Dataset:**

Dataset has been collected from open source and is available at through the UW CS ftp server : *ftp ftp.cs.wisc.educd math-prog/cpo-dataset/machine-learn/WDBC/.* Besides that it is also available at Kaggle. Since this dataset is used for the classification work. The features that are available in this dataset are listed below:



***Fig-1***

**Details of Scales:-**

Since we can see that “id” is one the features of the dataset which is serial number of each patients who underwent into the breast cancer diagnosis. Since “id” is serial number hence it is interval scale. It contains 32 features and 532 rows.

|  |  |
| --- | --- |
| ***Features*** | ***Scales*** |
| id | Interval |
| diagnosis | Nominal |
| rest 30 features | Ratio |

The above is the details of the scales of the each feature which are available in the dataset.

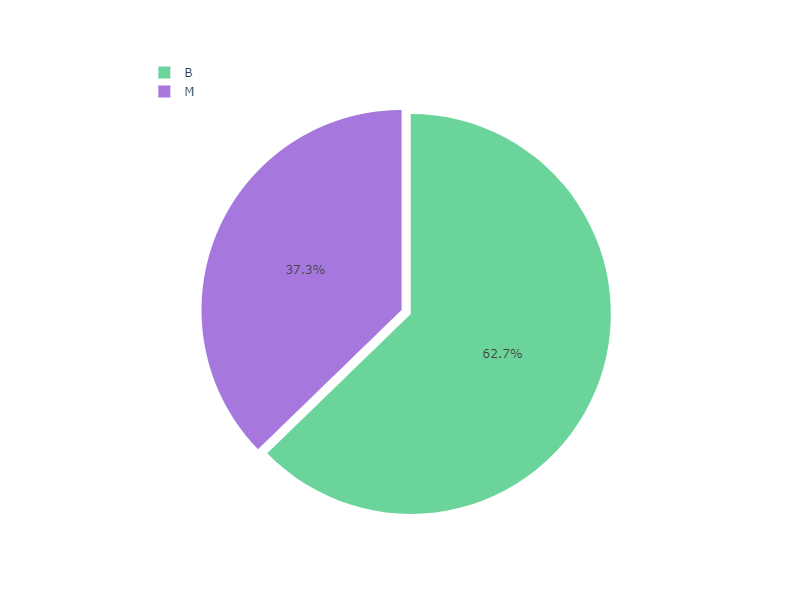
**Explanatory and Response variable:**

As we can see that there

|  |  |
| --- | --- |
| Features | variables Type |
| diagnosis | Response |
| rest 30 features | Expalanatory |

**Note**: There are total of 32 variables. Since “***diagnosis***” is **response variable** while rest of excluding “***id***” are the **explanatory** variables.

**Response Variables:-** It can bee seen that



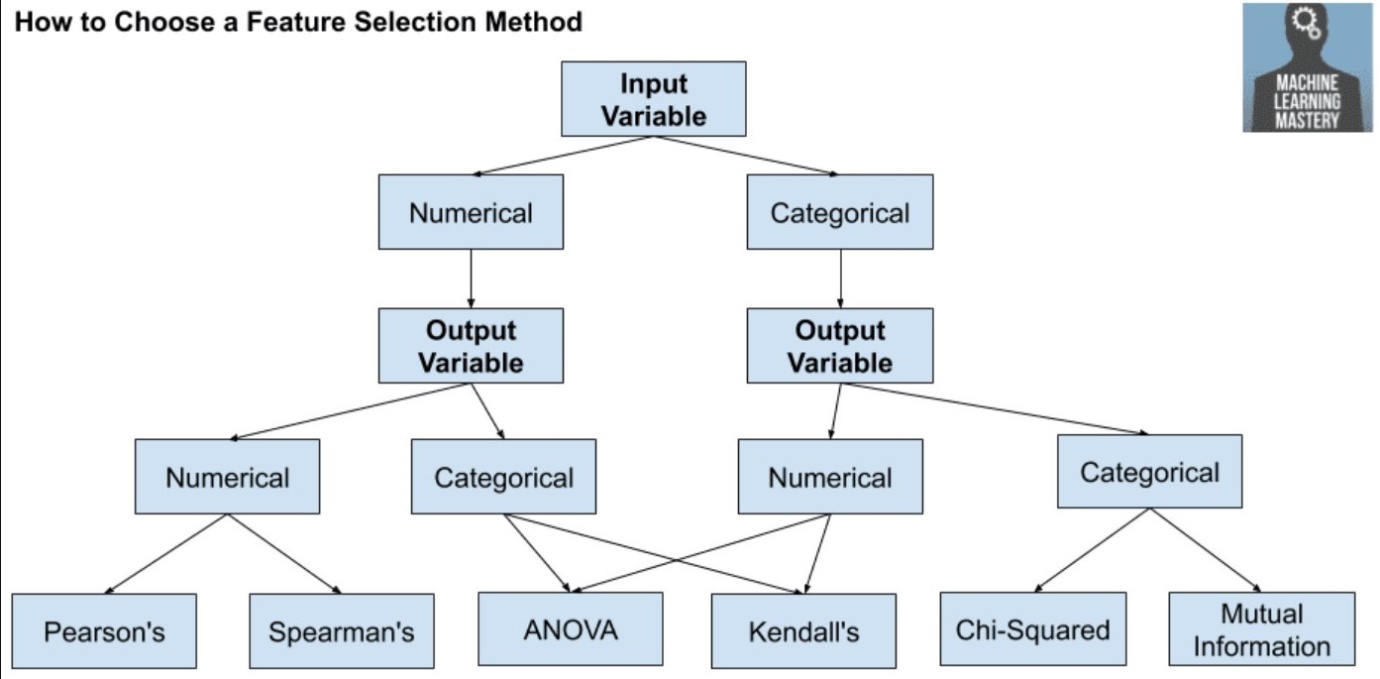
***Fig-2***

From the figure it can be observed that there “B” is 62% while “M” is almost 38%. We can see a kind of data imbalanced. Though in our analysis I have considered two cases:

**Case-I**: Where we have ignored the data imbalanced

**Case-II**: Where we have considered the data imbalanced and sampling is performed.

**Missing Values:-** After carefully analysing the dataset we have seen that it is perfect dataset which contains no null values. Hence our work became easy as no imputation has been performed here.

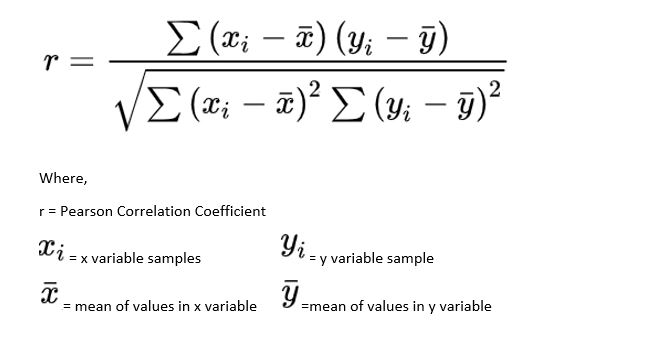


***Fig-3***

In our case we see that our input or explanatory variables are numerical and output or response variable is categorical one. Since in my case I have taken categorical as numerical and find out the

**Pearson’s Correlation Coefficient:-**

To find out the relationship between the response variable and each explanatory variable, I have calculated Pearson’s coefficient. The formula for the correlation coefficient is given by



|  |
| --- |
| Case-1: when *r=*-1, it shows that both the variables are negatively correlated. With the increase of the one variable other will decrease.  Case-2: when r=0, it shows that both the features are independent.  Case-3: When r=1, it shows that both the features are highly related. |

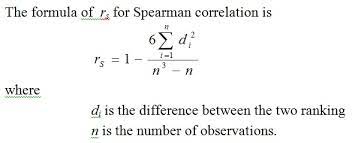
Now we have the required coefficient calculation of the all features related with response variables.

|  |  |  |
| --- | --- | --- |
| Features | Pearson Coefficient | P-value |
| id | 0.039769 | 3.44E-01 |
| diagnosis | 1 | 0.00E+00 |
| radius\_mean | 0.730029 | 8.47E-96 |
| texture\_mean | 0.415185 | 4.06E-25 |
| perimeter\_mean | 0.742636 | 8.44E-101 |
| area\_mean | 0.708984 | 4.73E-88 |
| smoothness\_mean | 0.35856 | 1.05E-18 |
| compactness\_mean | 0.596534 | 3.94E-56 |
| concavity\_mean | 0.69636 | 9.97E-84 |
| concave points\_mean | 0.776614 | 7.10E-116 |
| symmetry\_mean | 0.330499 | 5.73E-16 |
| fractal\_dimension\_mean | -0.012838 | 7.60E-01 |
| radius\_se | 0.567134 | 9.74E-50 |
| texture\_se | -0.008303 | 8.43E-01 |
| perimeter\_se | 0.556141 | 1.65E-47 |
| area\_se | 0.548236 | 5.90E-46 |
| smoothness\_se | -0.067016 | 1.10E-01 |
| compactness\_se | 0.292999 | 9.98E-13 |
| concavity\_se | 0.25373 | 8.26E-10 |
| concave points\_se | 0.408042 | 3.07E-24 |
| symmetry\_se | -0.006522 | 8.77E-01 |
| fractal\_dimension\_se | 0.077972 | 6.31E-02 |
| radius\_worst | 0.776454 | 8.48E-116 |
| texture\_worst | 0.456903 | 1.08E-30 |
| perimeter\_worst | 0.782914 | 5.77E-119 |
| area\_worst | 0.733825 | 2.83E-97 |
| smoothness\_worst | 0.421465 | 6.58E-26 |
| compactness\_worst | 0.590998 | 7.07E-55 |
| concavity\_worst | 0.65961 | 2.46E-72 |
| concave points\_worst | 0.793566 | 1.97E-124 |
| symmetry\_worst | 0.416294 | 2.95E-25 |
| fractal\_dimension\_worst | 0.323872 | 2.32E-15 |

**Observation**:- We can see that some of the features which are correlated or independent on the basis of Pearson’s correlation values.

**Spearman’s Correlation Coefficient**-

it is same as with Pearson’s correlation coefficient.



We get almost the similar values of correlation as that we get in case of Pearson’s correlation coefficient.

**Difference between Spearman and Pearson’s coefficient**

* Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not)

**Multi-collinearity**

As we can see that there many features in the dataset i.e. there are 32 features including the response variable. There is high chance that there might be multicollinearity among the features. To detect the multicolinearlity we have multiple options.

**Reason for the Multicollinearity**

It arises when there are

* One correlation between independent and dependent variables.
* One independent variable is correlated with a linear combination of two or more

**Why we need to examine Multi-collinearity**

Since we know one thing that if the coefficients of the features are much higher, that variable will contribute towards the prediction of the response variables. If we are able to distinguish among those variable or features, the prediction of the response variable will improve. All we need to find that variable which is affected by multi-collinearity.

**How to Deal with Multicollinearity**

To deal with multicollinearity we have one option which is dealt below

* Variance inflation factor (VIF) measures how much the behaviour (variance) of an independent variable is influenced, or inflated, by its interaction/correlation with the other independent variables.
* 1 = not correlated.
* Between 1 and 5 = moderately correlate.
* Greater than 5 = highly correlate.
* VIF> 10 is used and remove the feature which has that high VIF.

|  |
| --- |
| feature VIF |
| 0 id 1.163246 |
| 1 radius\_mean 63309.441555 |
| 2 texture\_mean 251.432664 |
| 3 perimeter\_mean 58123.587378 |
| 4 area\_mean 1287.411538 |
| 5 smoothness\_mean 393.514898 |
| 6 compactness\_mean 201.166758 |
| 7 concavity\_mean 157.899635 |
| 8 concave points\_mean 154.297834 |
| 9 symmetry\_mean 184.429153 |
| 10 fractal\_dimension\_mean 629.688286 |
| 11 radius\_se 237.257123 |
| 12 texture\_se 24.758627 |
| 13 perimeter\_se 211.410744 |
| 14 area\_se 73.436569 |
| 15 smoothness\_se 26.330278 |
| 16 compactness\_se 44.948568 |
| 17 concavity\_se 33.254271 |
| 18 concave points\_se 54.030035 |
| 19 symmetry\_se 37.207715 |
| 20 fractal\_dimension\_se 27.549724 |
| 21 radius\_worst 9677.820028 |
| 22 texture\_worst 343.005975 |
| 23 perimeter\_worst 4487.783848 |
| 24 area\_worst 1139.047176 |
| 25 smoothness\_worst 375.678745 |
| 26 compactness\_worst 132.994862 |
| 27 concavity\_worst 86.311570 |
| 28 concave points\_worst 148.786616 |
| 29 symmetry\_worst 219.018071 |
| 30 fractal\_dimension\_worst 423.466387 |

**Observation**

We can make some of the observations. Those are

* VIF of most of the features are more than 10.
* As it is not possible to remove the all the features.
* There is alternative way to remove the features. The higher VIF values will be removed and the remaining features will be used in modelled.

**Alternative Method to remove features**

There are two methods that I have performed in order to remove some of the features which might not be

* In our case VIF is not working well. So I have chosen an alternative option. We have to perform some statistical analysis. We have to calculate the p-value for each feature. We remove those features which have higher p-value.
* Next is normality test is performed for each of the variables. Those features which do not follow the Gaussian distribution will be not used for the modelling. To perform the normality test, we can use QQ plot, distribution plot and histogram. Beside that some KS test and chi-square goodness of fit are also performed.

# Q-Q plot

* Another popular plot for checking the distribution of a data sample is the quantile-quantile plot, Q-Q plot, or QQ plot for short.
* It is a technique to compare whether two sets of sample points are from or they follow same distributions.
* one distribution is known , we have to check for the other distribution
* If the unknown sample of dataset follow given distribution, we will have a scatter plot, where data points will be in a straight line y = x.
* The idea is to plot the quintile values of two distributions/samples and see
* If they make a straight line or not. If the quintiles of two sample sets are similar or in a better case, identical then sample set is from the same distribution. The process of QQ plot
* Arrange the dataset in increasing order.
* Calculate the percentiles for each increasing dataset
* Plot those percentiles with the help of Scatter plot
* Done

# Shapiro-Wilk Test

* Evaluates a data sample and quantifies how likely it is that the data was drawn from a Gaussian distribution
* The Shapiro-Wilk test is believed to be a reliable test of normality, although there is some suggestion that the test may be suitable for smaller samples of data, e.g. thousands of observations or fewer.
* The function returns both the W-statistic calculated by the test and the p-value.

**Dataset Spliting**:

Dataset is containing the whole features. Before feeding into the model we will split the dataset in train and test. Train dataset is fed into the machine learning model and accuracy score is calculated. Test dataset is unseen data which is used for test performance. The main reason for using splitting the dataset is to check for the ***Simpson’s paradox***. Since we have not observed such phenomenon.

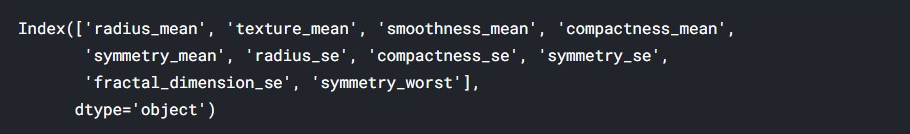
**Feature selection Method**:

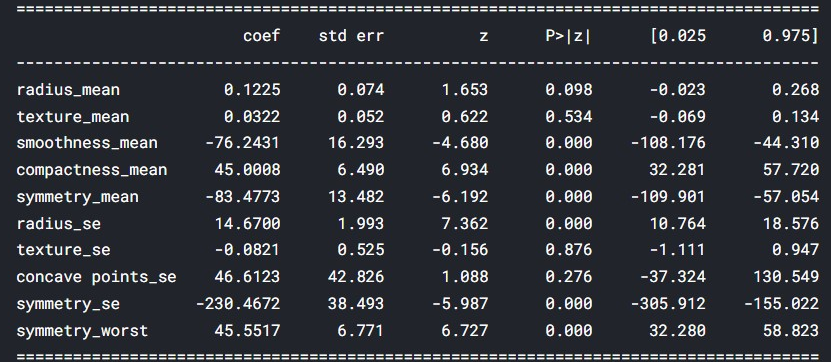
Here I have mentioned of the technique of choosing the variable to fit into the model and perform the accuracy rate. There are various advantage of feature selection process. These are as follow

* Improved accuracy
* Simple models are easier to interpret.
* Shorter training times
* Enhanced generalization by reducing overfitting
* Easier to implement by software developers
* Reduced risk of data errors by model use
* Variable redundancy
* Bad learning behaviour in high dimensional spaces

**Forward Selection:-**

* Forward selection is an iterative method in which we start with having no feature in the model. In each iteration, we keep adding the feature which best improves our model till an addition of a new variable does not improve the performance of the model.
* The procedure starts with an empty set of features [reduced set].
* At each subsequent iteration, the best of the remaining original attributes is added to the set.
* The pre-set criteria can be the roc\_auc for classification and the r squared for regression for example.
* This selection procedure is called greedy, because it evaluates all possible single, double, triple and so on feature combinations. Therefore, it is quite computationally expensive.
* When sample size < number of predictor, then we can use the forward selection method.

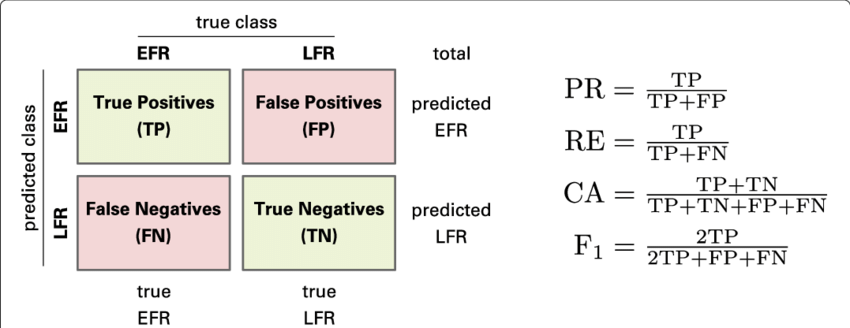


****

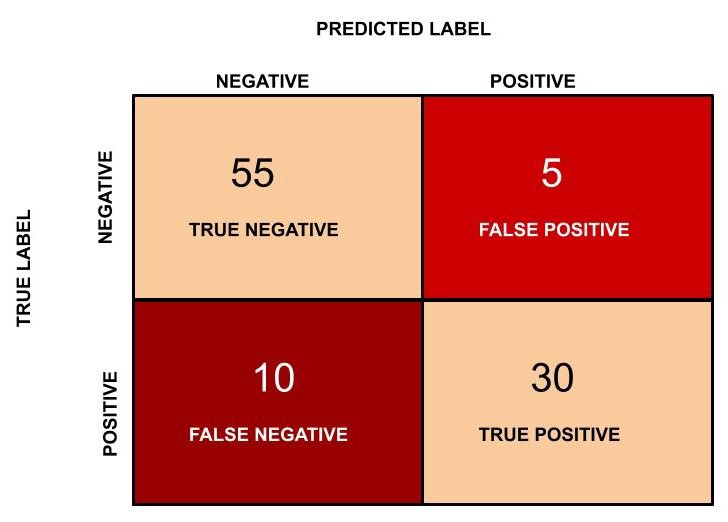
**Metrics for Classification**

The metrics are used for obtaining the accuracy of the model is

* **Accuracy Score**- In classification accuracy is for a given points how many points are correctly classified. It is good metrics for the balanced class but it will not give good values for the case of imbalanced dataset.
* **Confusion Metrics**- It is tool to find out the accuracy of the model. It is matrix of the true class and predicted class.

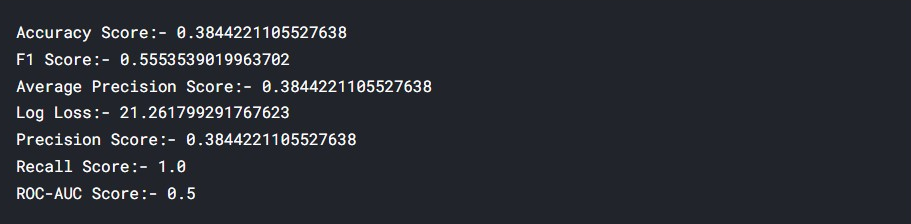


* **Precision**:- It is the ratio of True Positive/(True Positive + False Positive). It is one of the good metrics for imbalanced dataset. As we want the Precision to be 1 i.e. when FP=0.
* **Recall**:- Ratio of True Positive/(True Positive + False negative) . It is one of the good metrics for imbalanced dataset. As we want the Recall to be 1 i.e. when FN=0. Since in our case we want recall to be high. We want to reduce the Type-I error
* **F-1 score**- It is the harmonic mean of Precision and Recall.
* **ROC-AUC score:**  It is stand for Receiver Operating Characteristics curve. Drawing the curve between the True Positive Rate and False Positive rate. For each threshold value of the probabilities, we get different value of FPR and TPR; according to this we plot the curve.
* **Precision Recall curve:** It is plot between the precision and recall. It is good measure for the imbalanced dataset. A precision-recall curve is a plot of the precision (y-axis) and the recall (x-axis) for different thresholds, much like the ROC curve. Dumb model gives an area of equal to 0.5.

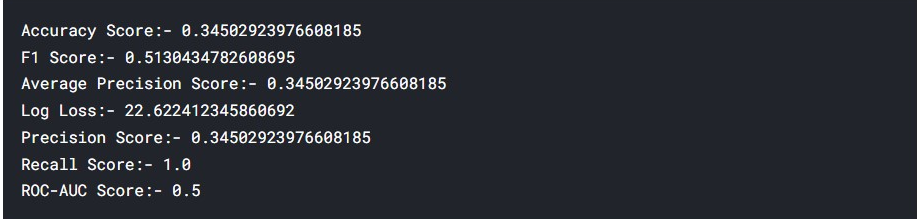


**Useful Metrics**

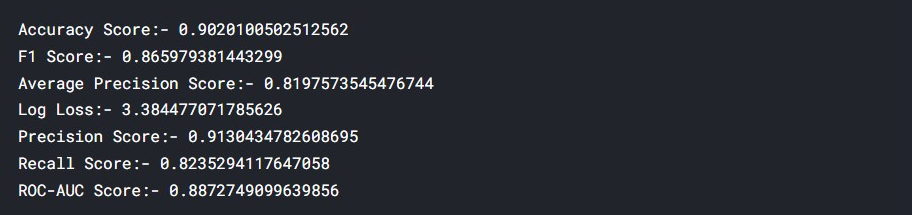
Before applying the feature selection on Train data



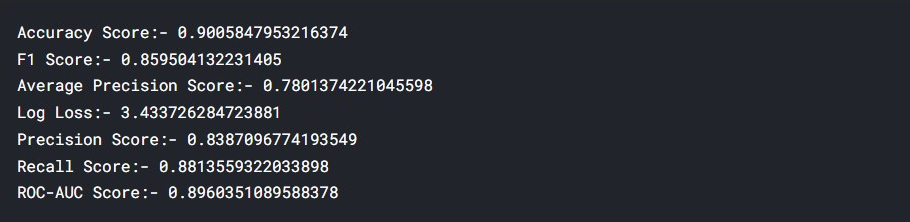
Test dataset-



After applying the feature selection on Train data



Test dataset



**Conclusion**

We can see that after using the different models like Logistic Regression, Random Forest and XGBoost. The first one is giving quite better results as it is avoiding the over fit. After the feature selection we conclude that few features are quite important as

* We see that accuracy score has increased 3 times.
* Log-loss has decreased.
* Precision is increased.
* ROC-AUC has increased.