Data Science& Machine Learning

Project Report

On

**Exploring Classification Techniques for Bank Marketing: A Comparative Analysis of Machine Learning Models**

Submitted by

Kanika Sagar

Group 8, Project V

**IIT Guwahati Data Science and Machine Learning**

**-- Prof. Babji Srinivasan--**

**Contribution made by each member:**

|  |  |  |  |
| --- | --- | --- | --- |
| **S.no** | **Member Name** | **Contribution** | **Email ID** |
| 1 | Kanika Sagar | 100% | sagar.kanika1424@gmail.com |

**Objective**

The primary objective of this project is to develop a predictive model that accurately determines whether a client will make a deposit or not bases on various demographic and marketing related features. Specifically, we aim to utilize the provided dataset containing attributes such as age, job, marital status, education, housing situation, loan etc. Using this information, we intend to build a robust classification model that can effectively classify clients into two categories: those likely to make a deposit and those not likely to make a deposit.

**Introduction**

Deposit is the major source of income for the bank. A deposit refers to placement of funds into an account held at a financial situations, such as bank or credit union. Deposit serves as means for individuals or business to securely store their money with the expectation of earning interest or facilitating financial transactions. Additionally, deposit are a crucial source of funding of banks, enabling them to lend money to borrowers and conduct other financial activities.

In today’s banking landscape, predicting customer behaviour, particularly deposit subscriptions, holds immense importance for banks in terms of loan management and marketing strategies. This dataset comprising various socio-economic attributes and marketing campaign details, serves as a vital resource for exploring these dynamics.

By leveraging machine learning techniques on this dataset, our main goal is to predict the term deposit (Variable Y), indicating column subscription to term deposits. This prediction holds significant importance for the banks, guiding loan management decisions and optimizing marketing strategies to enhance customer engagement and overall business performance.

**Dataset**

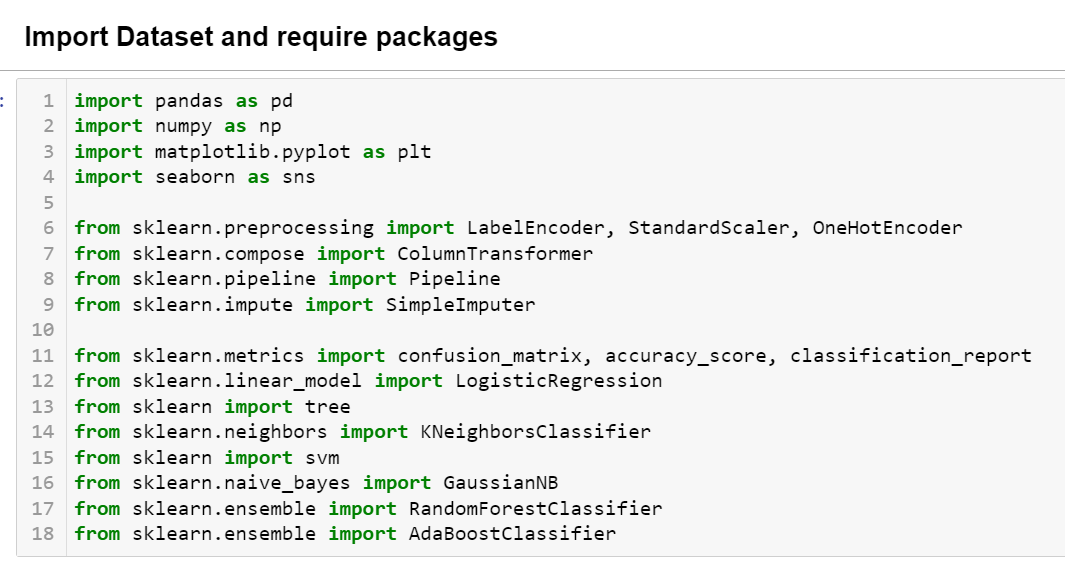
Train full data.csv :- 41188 rows , 21 column

Test full data.csv:- 4119 rows, 21 column

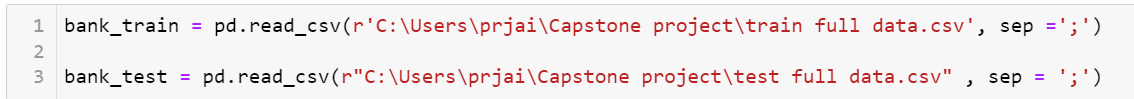
**Code Implementation**

**Importing needed library**

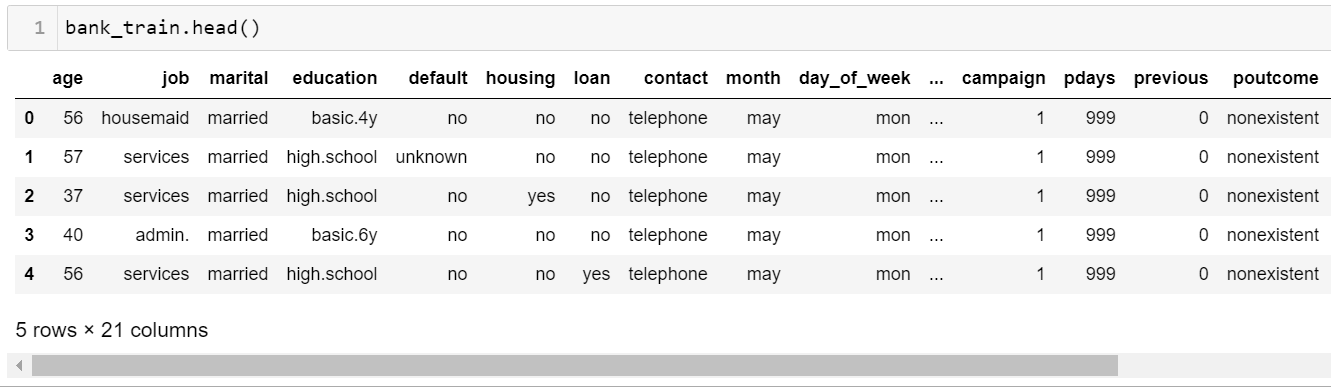
* Data analysis and visualization
* Necessary encoder, pipeline, standard scaler for feature scaling
* Data splitting, cross-validation, classification performance, accuracy score
* Necessary classifier for building machine learning models.



**Reading training and testing dataset as dataframes.**



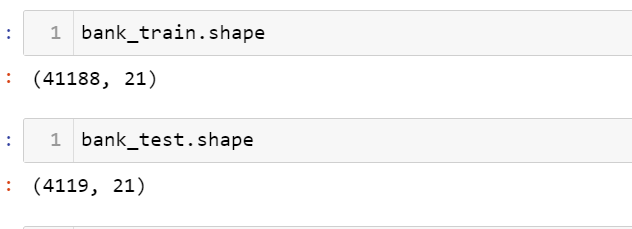
**Displaying first five entries.**



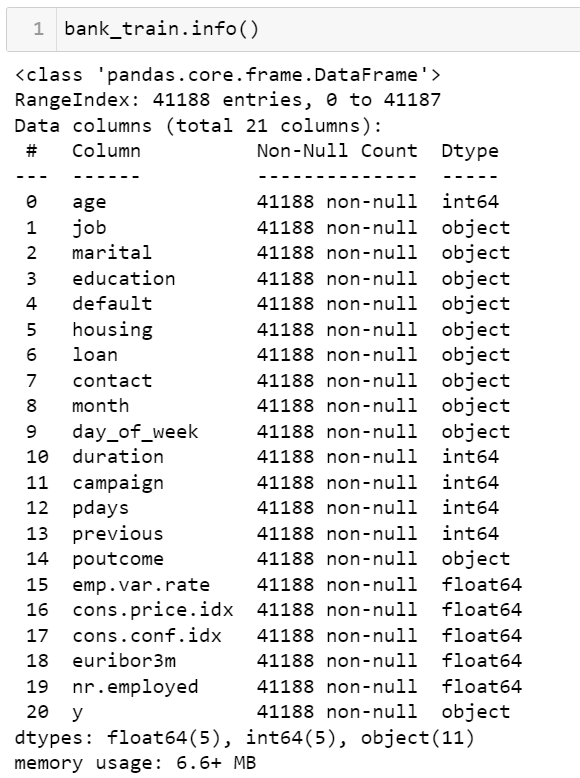
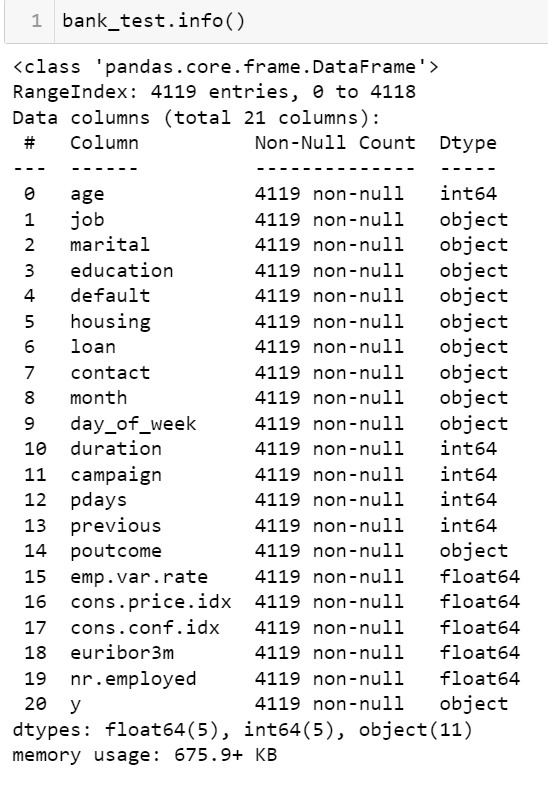


**Displaying Information about the dataset.**

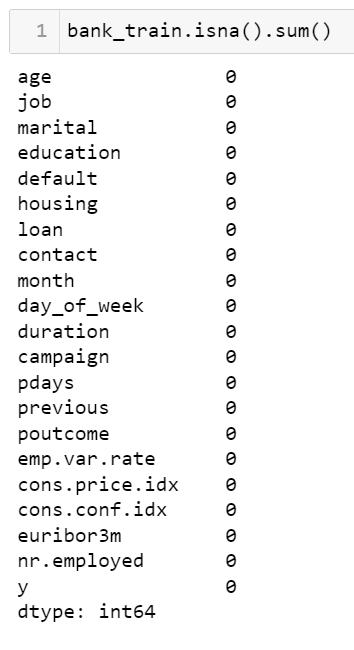
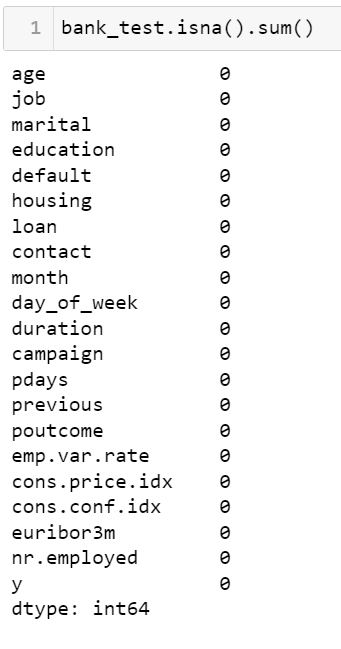
* Dataset shape.



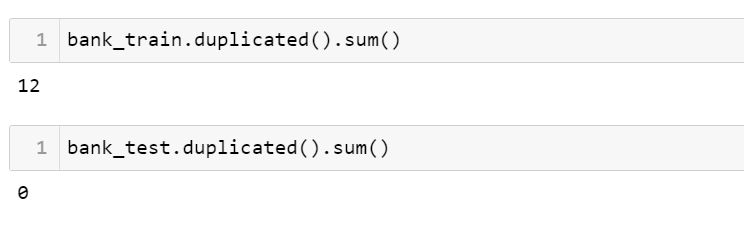
* Training Dataset consist of 41188 rows and 21 column.
* Testing Dataset consist of 4119 rows and 21 column.
* Dataset Information.

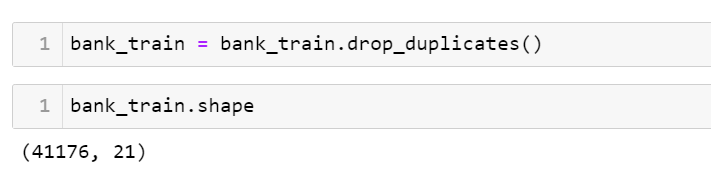
* Checking of missing/Null value.

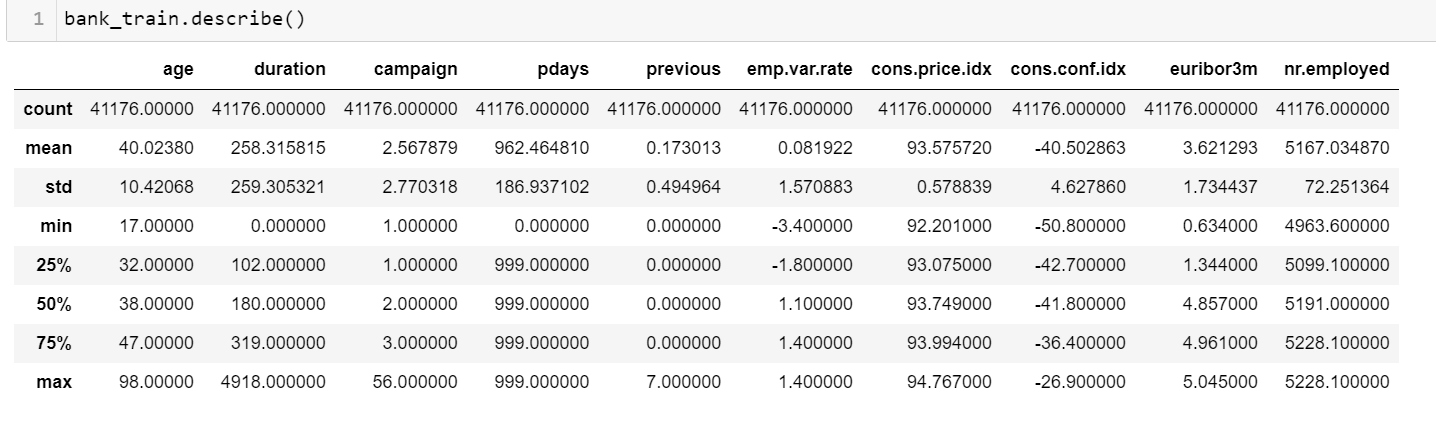
* There is no missing value in both dataset.
* Check for Duplicate values & Remove if any.

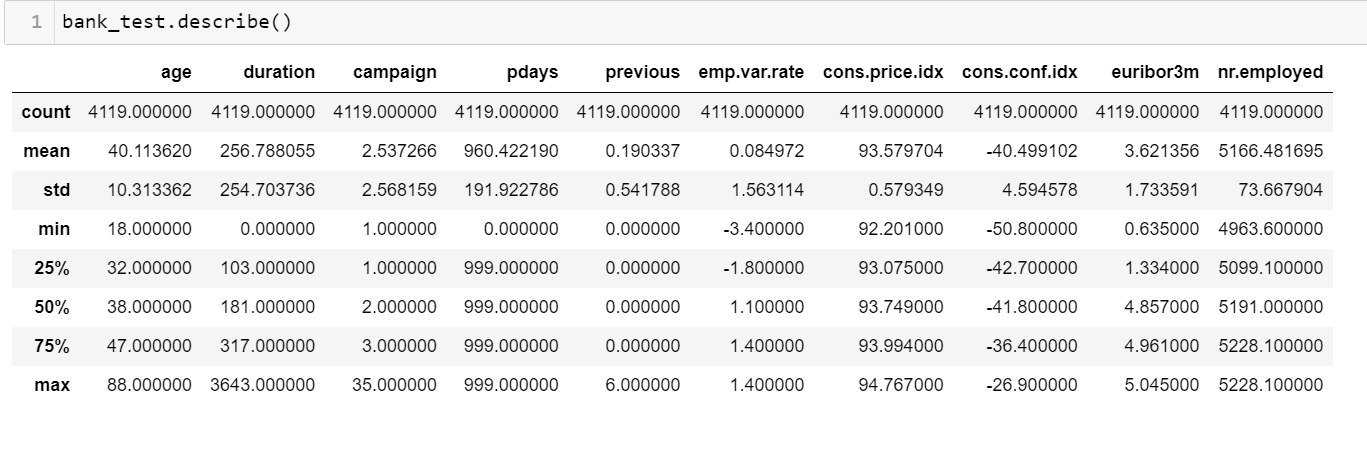


* 12 values are duplicate in Training dataset.

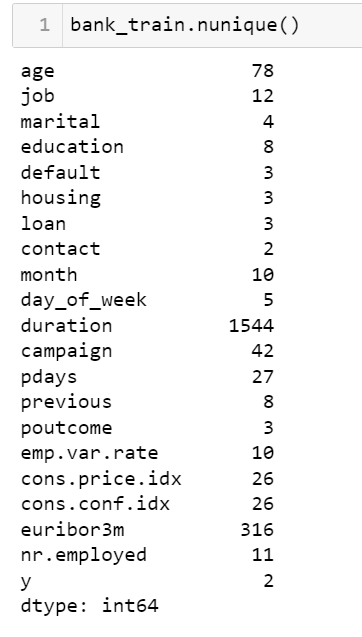


* Generating Descriptive statistics for dataset.

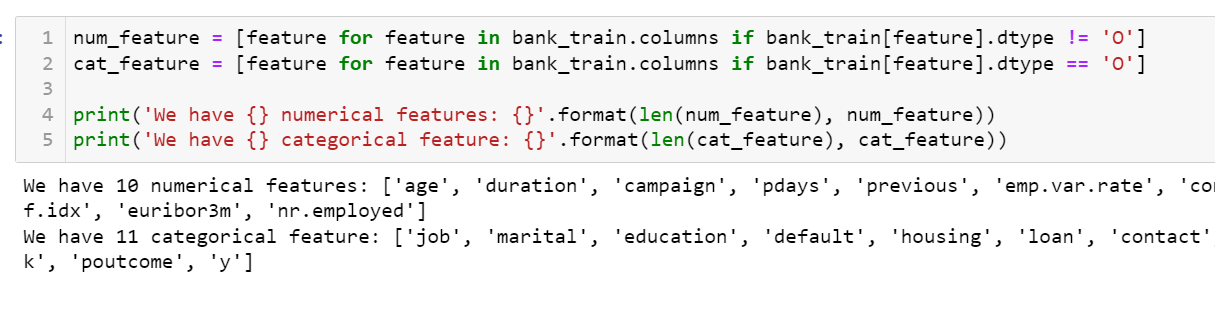




* Checking for unique values of each column.

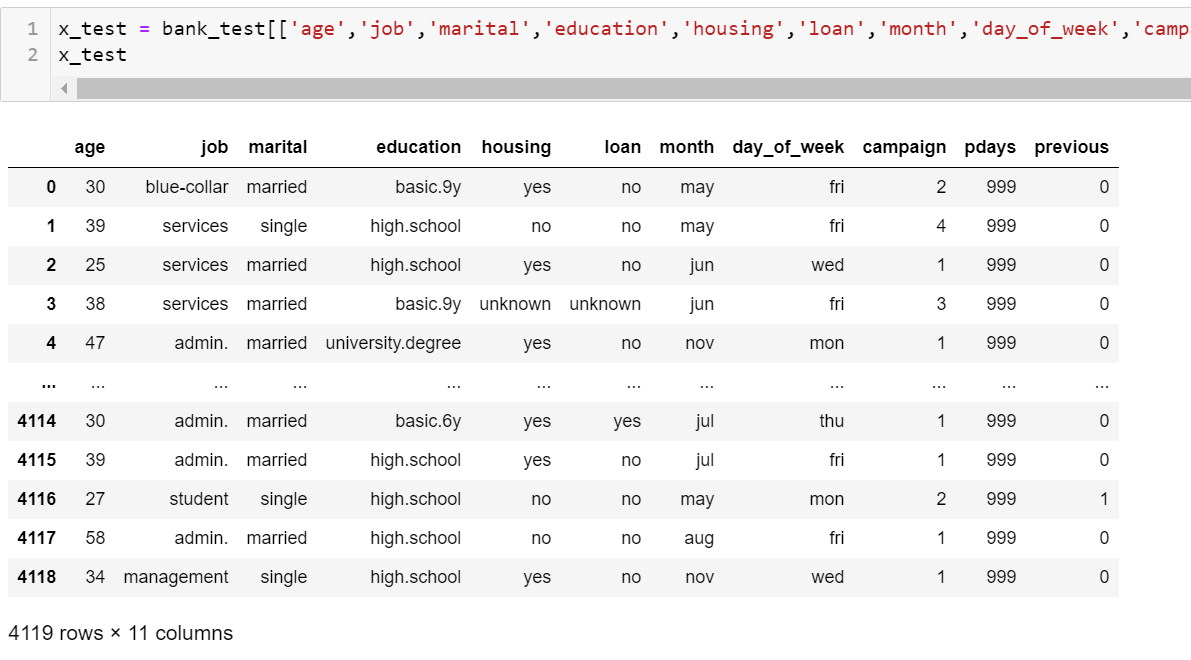
* Differentiating numerical and categorical features in the dataset and creating the copy of dataset.



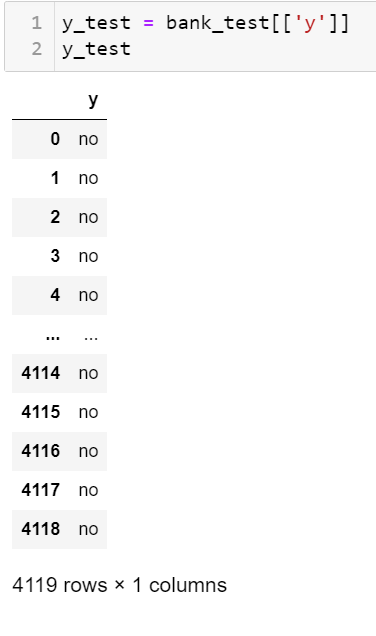
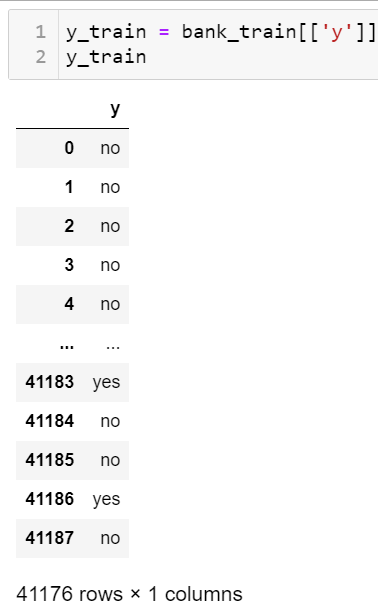
* Both dataset contain 10 numerical and 11 categorical features.
* **Defining X & Y dataset for training and testing set.**
* X\_train



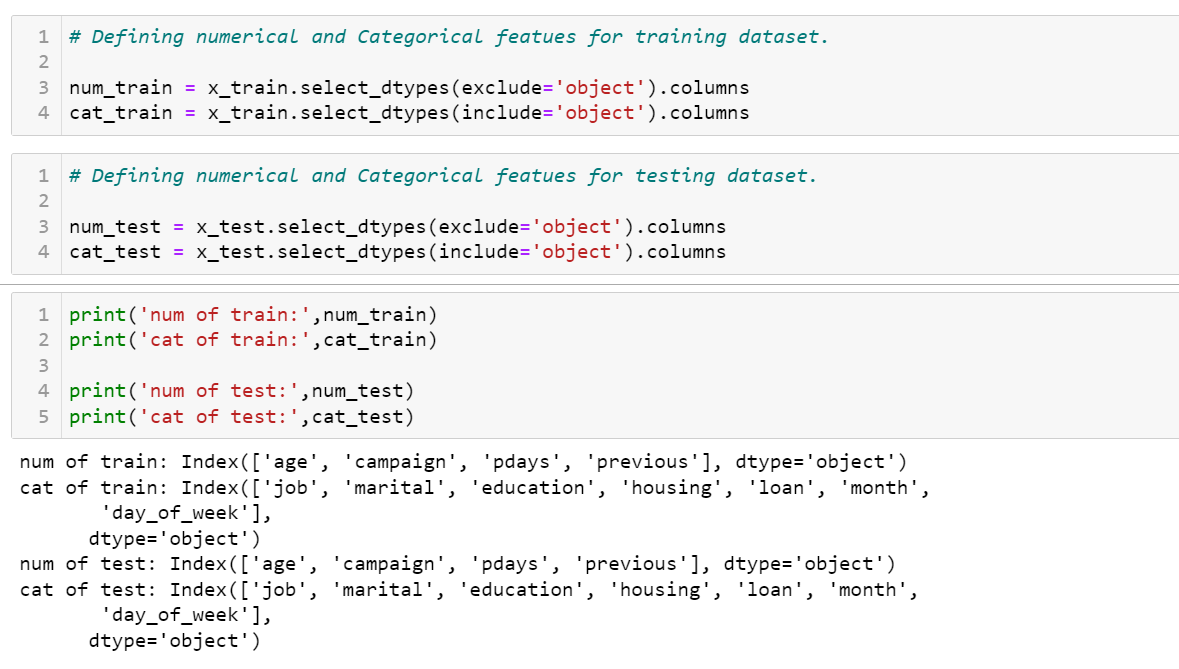
* X\_test



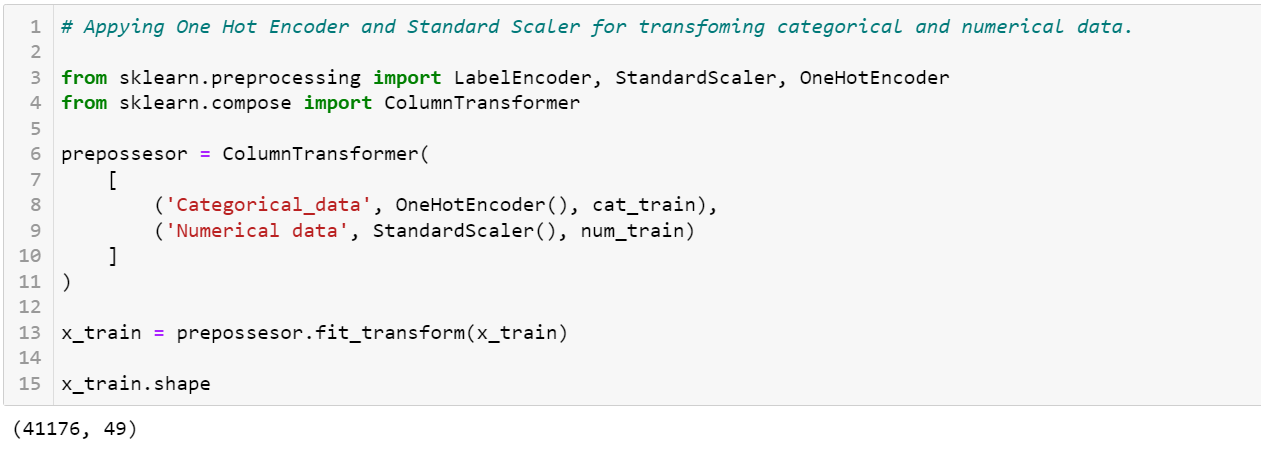
* Y\_train & Y\_test

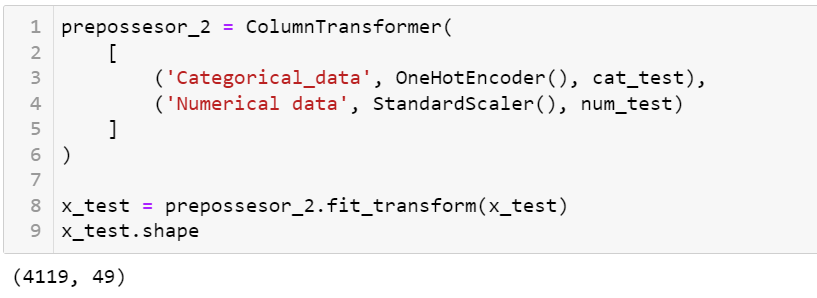


* Defining categorical and numerical data for training and testing dataset.

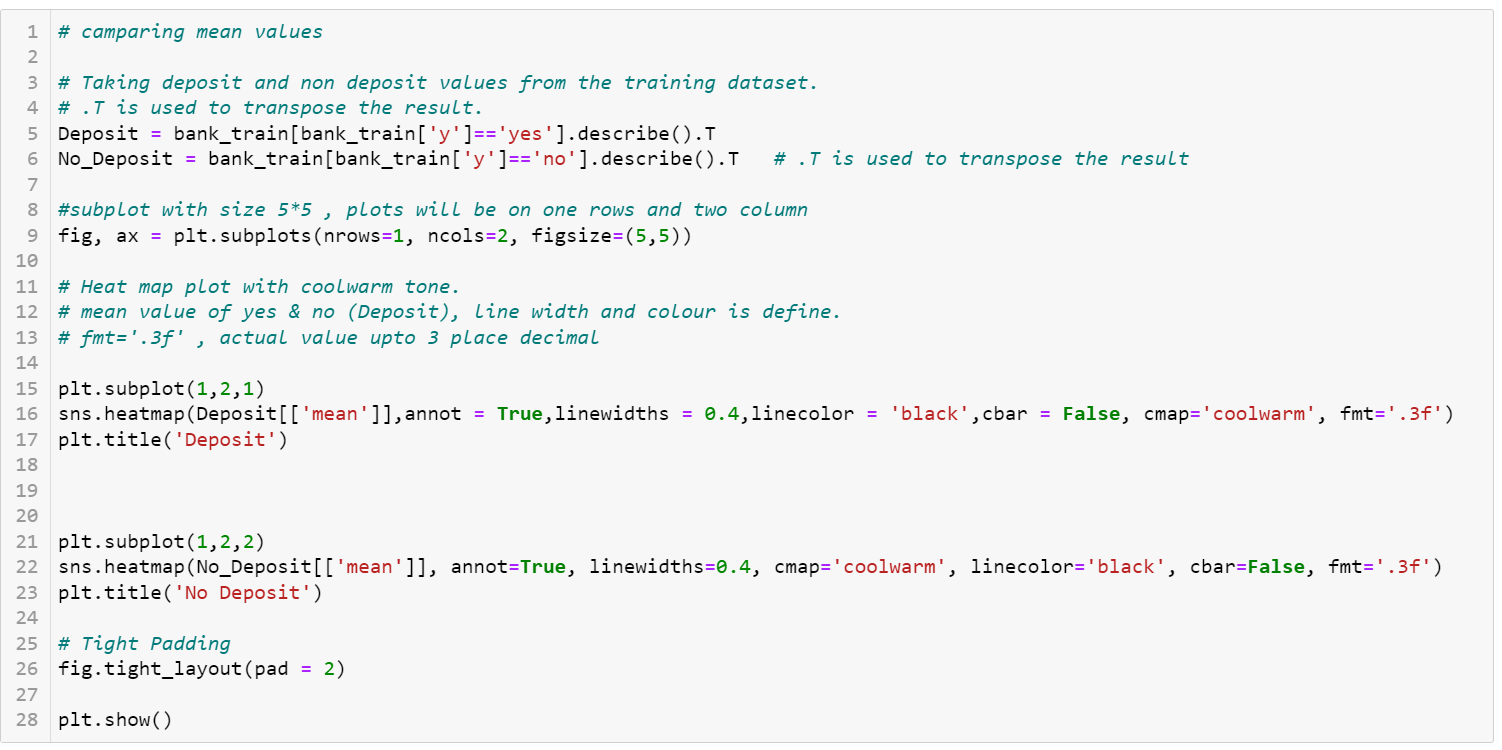


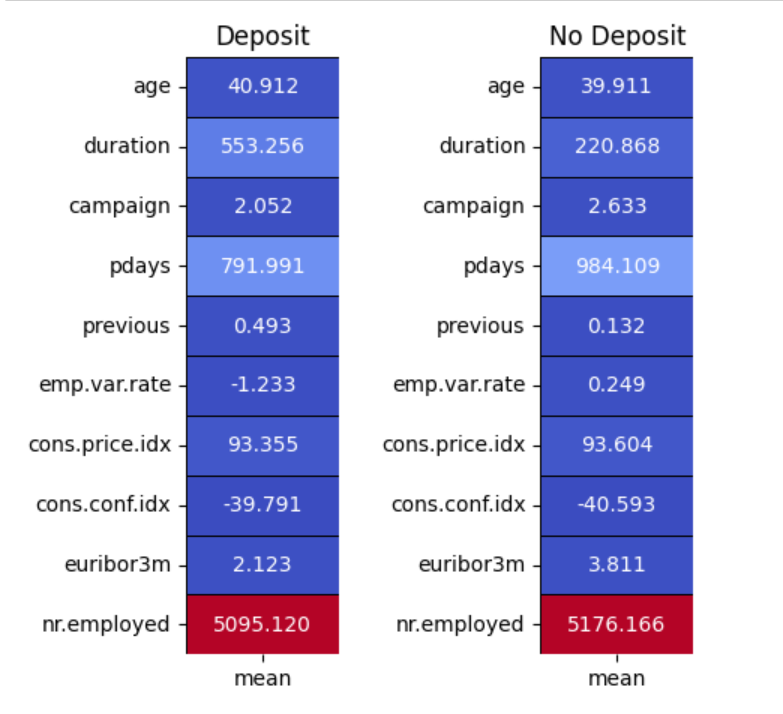
* Applying One Hot Encoding to transform categorical features into numerical representation and Standard Scaler to adjust the data to have standardized distribution.





* **Basic Visualization (Training dataset)**
* Comparing the mean value of y (Deposit) = ‘yes’ or ’no’ with Numerical features using heatmap.





1. Age: mean age of customer who made deposit is 41, which is slightly higher than who did not make deposit (40).
2. Duration: Average length of time that customer who subscribe / not subscribe to the deposit has spent with marketing campaign.

Higher the duration more the chances for the customer to subscribe for deposit.

1. Campaign: Average no of contact made for made to client who made a deposit is 2.05 whereas who did not make deposit is 2.63.

Average client who made the deposit were contacted fewer times during the market campaign compare to those who did not make deposit.

1. Pdays: Average no of days since the customer was last contacted from previous campaign.

Customer who made deposit were contacted approx. 198 days earlier in comparison to customer who did not make deposit.

1. Previous: Average no of contacts or interaction the customer had with the bank prior to the current campaign.

Customer who made the deposit has higher no of previous contact with bank before the current campaign.

1. Employment variation rate: Rate of change in employment levels over a specific period during economic growth.

Employment rate of for customers who made a deposit is approx. five times higher than the customer who did not make deposit.

1. Consumer Price Index: Average change over the time in the prices paid by urban consumer for a market basket of consumer goods and services.

Customer who made the deposit has slightly lower CPI compare to consumer who did not make deposit.

1. Consumer confidence index: Measure of consumer confidence in economy.

Customer who did not make the deposit tend to have lower confidence in the economy compare to those who made deposit.

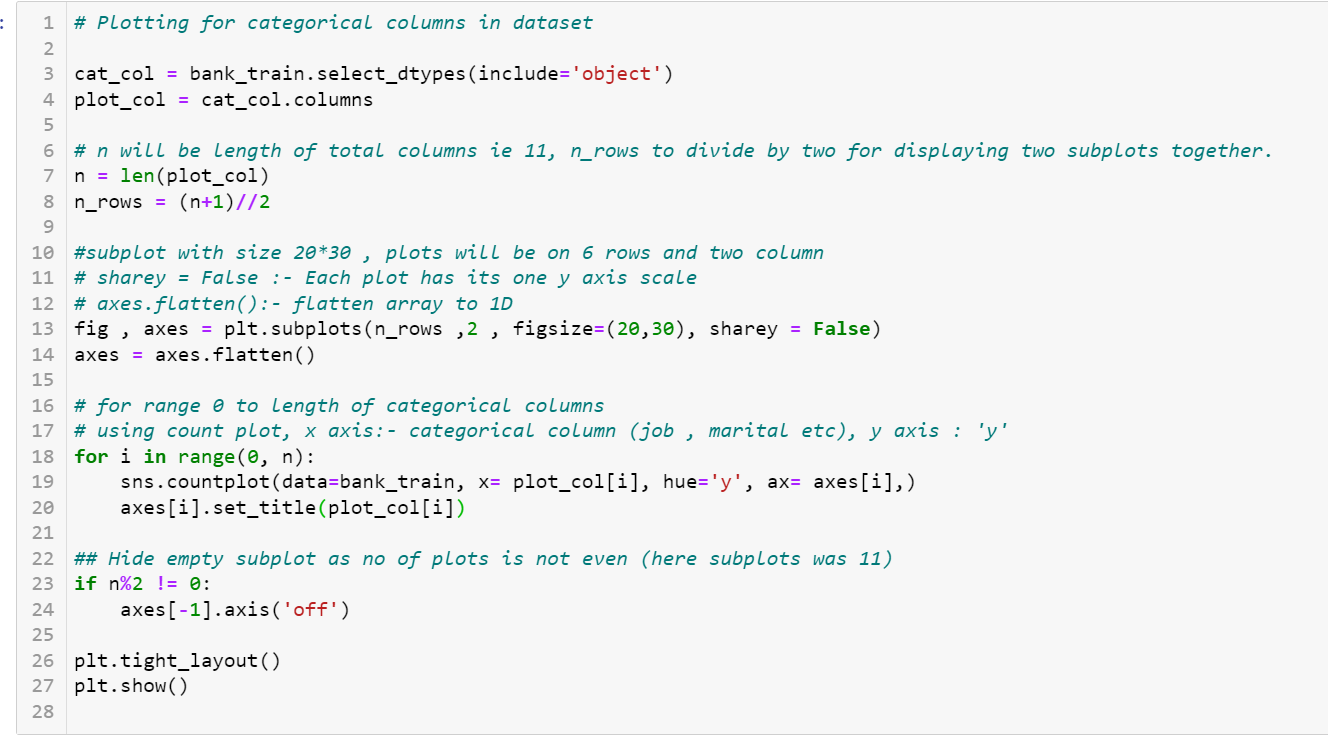
1. Euribor3m: Euro Interbank offered rate for 3 month period, Interest rate for which European bank offers to lend funds to one another.

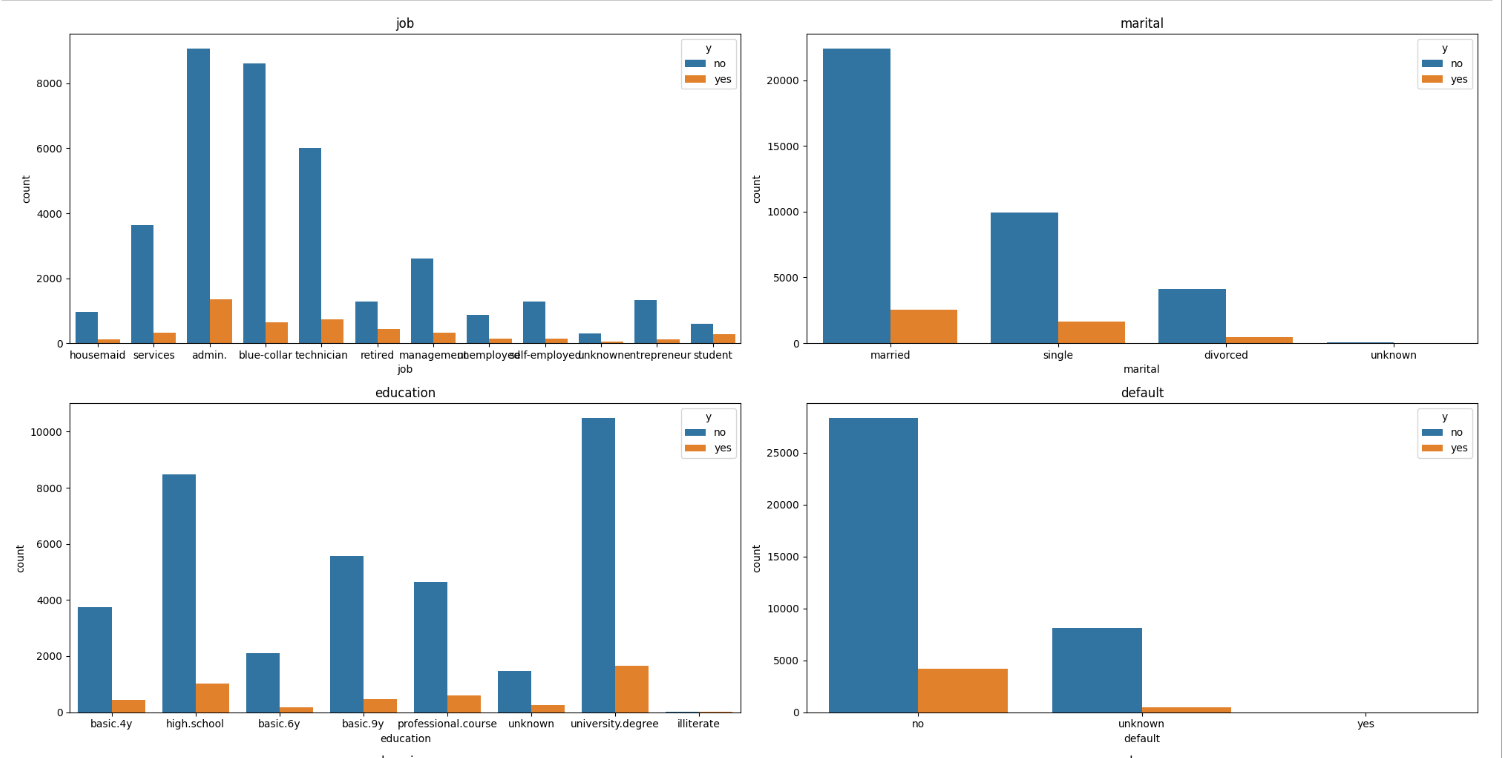
Customer who made deposit has lower Euribor 3 month rate campare to who did not make deposit.

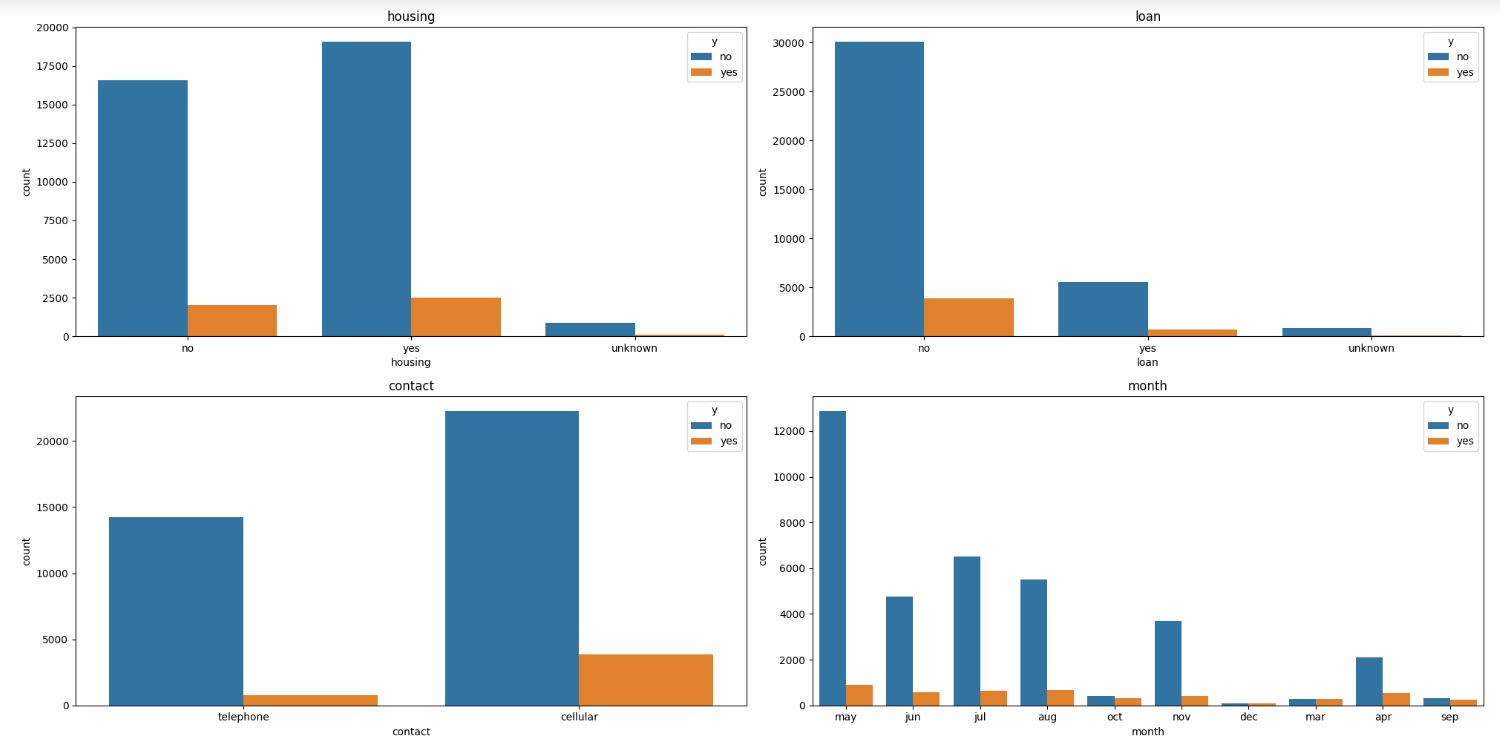
1. No of Employees: Total no of people employed in the country.

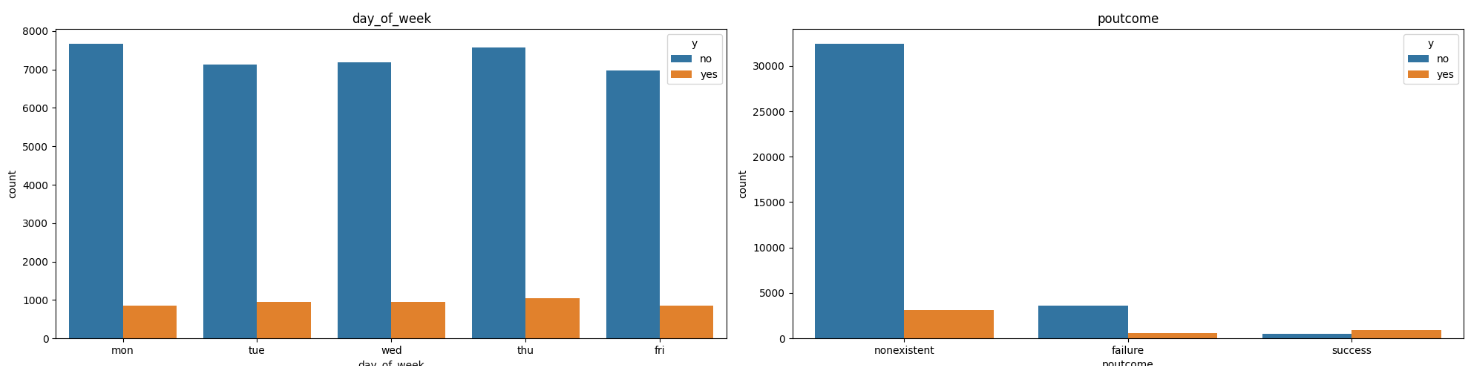
Customer with deposit were associated with a lower no of employees compare to the customers who did not make the deposit, indicating the potential difference in employment level between two groups.

* Visualization for categorical features wrt “y” using countplot.

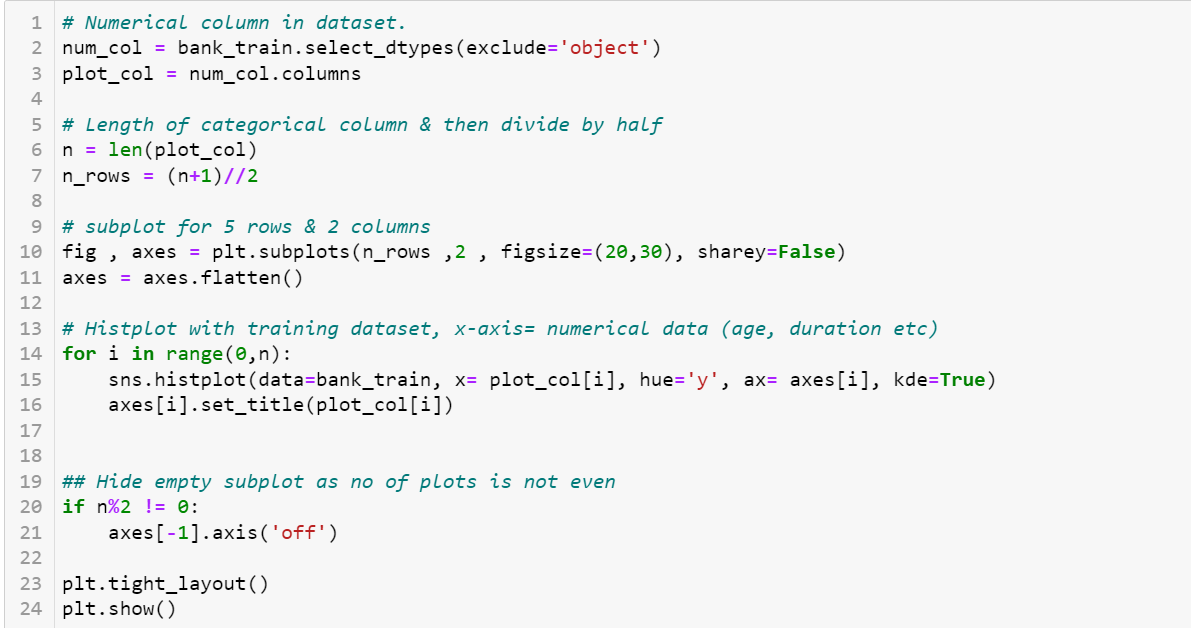


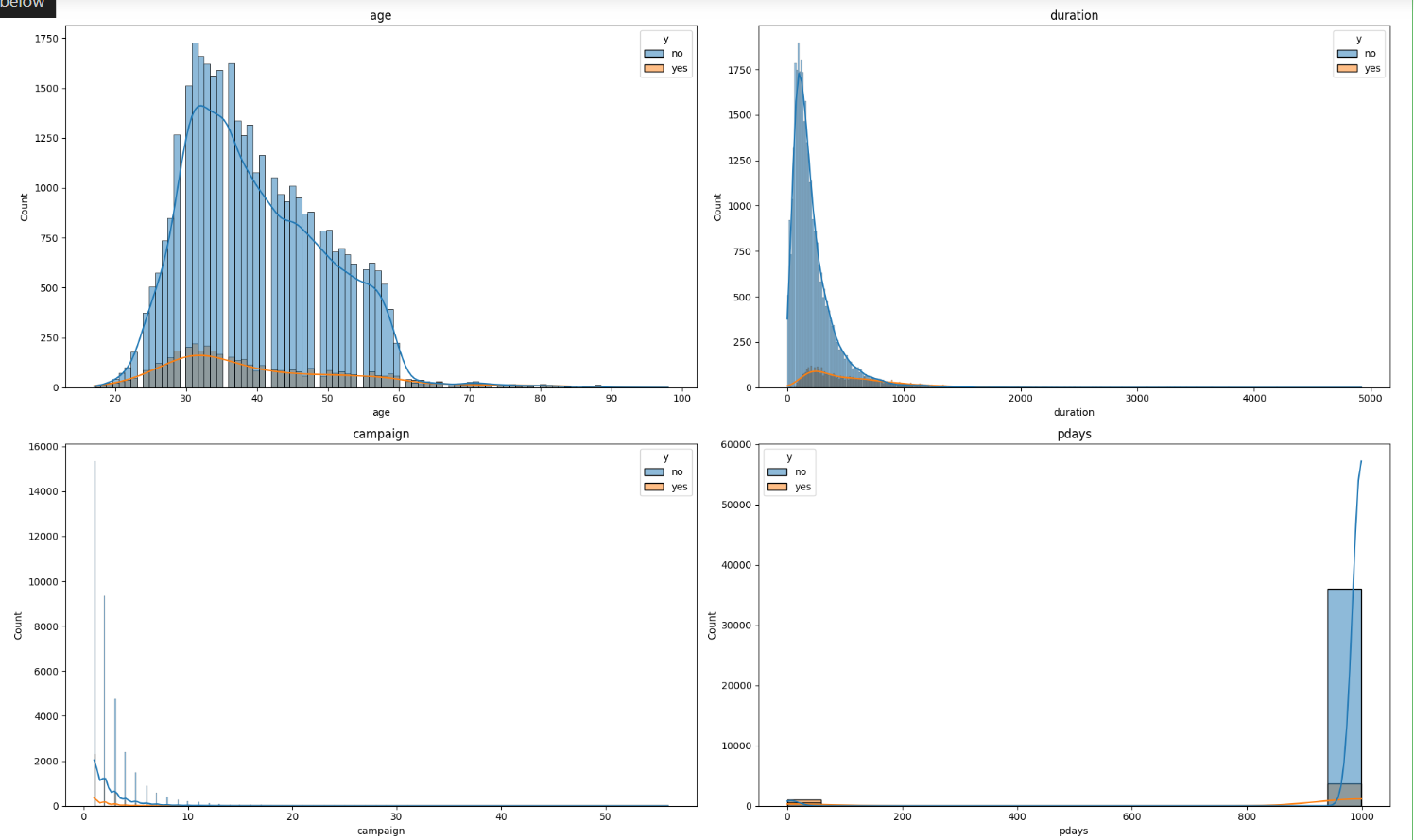


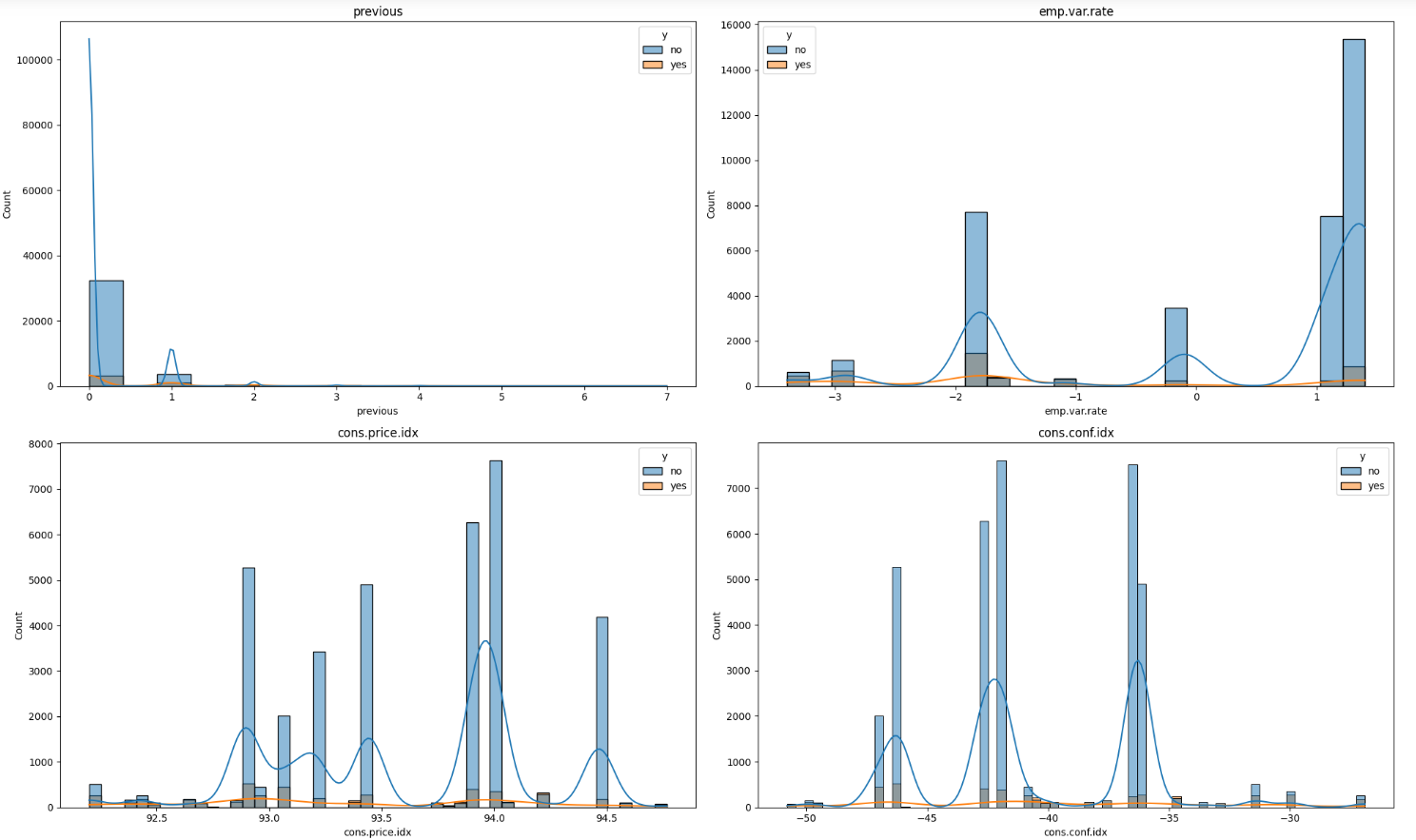


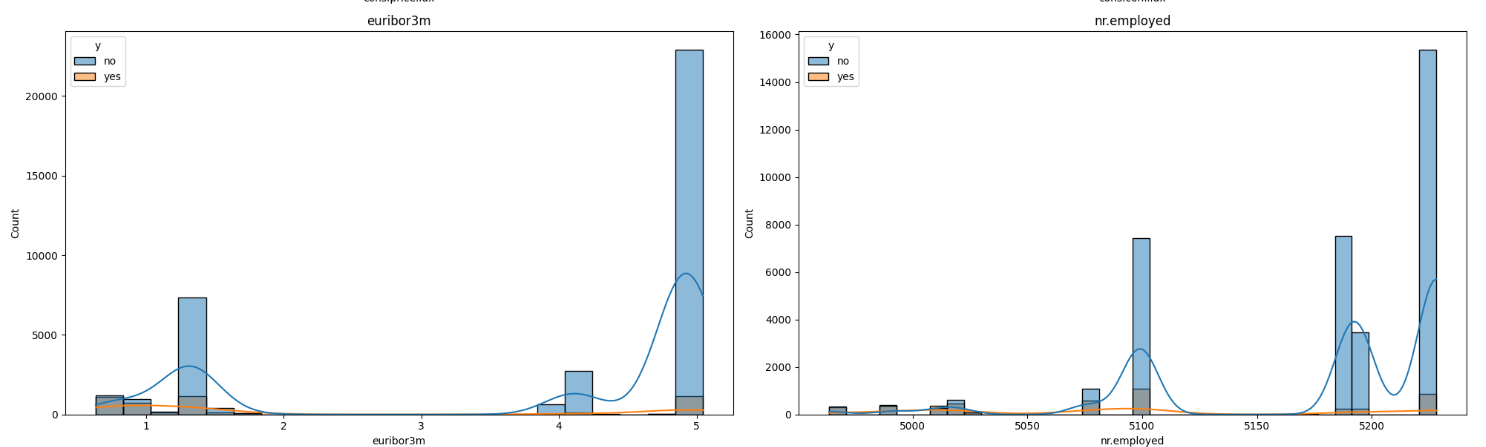


* Visualization for Numerical features wrt “y” using histplot.









**Evaluating Machine Learning Model and compare Accuracy score.**

* **Receiver operating characteristic (ROC) curve.**

Receiver operating characteristic ROC curve is the graphical representation of performance of binary classification model.

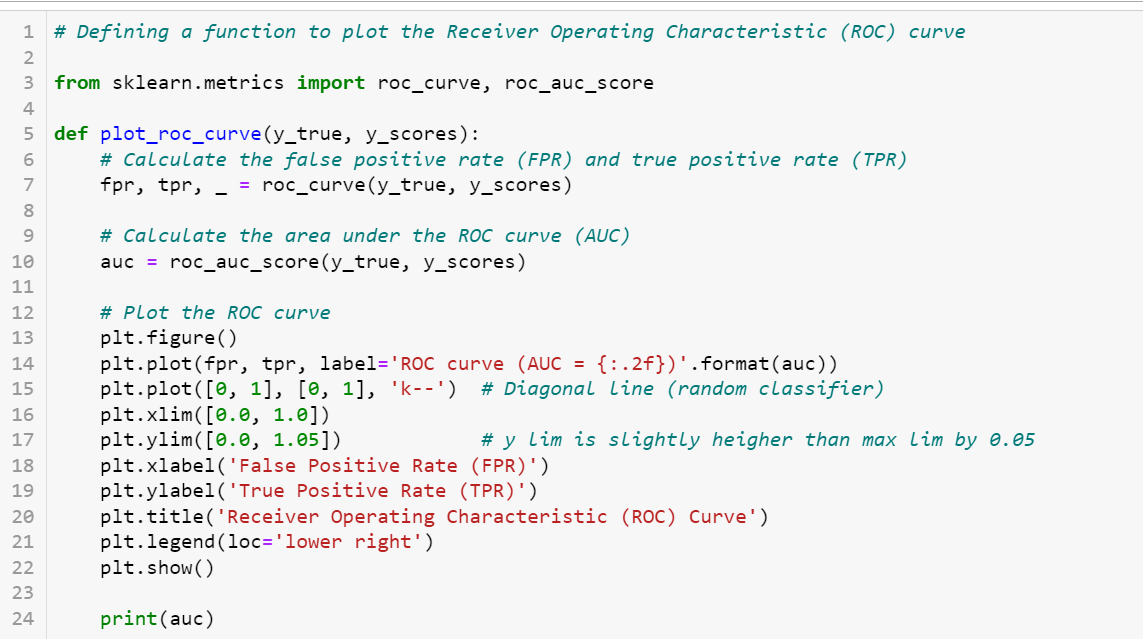
This curve has two parameters.

FPR:- False positive rate.

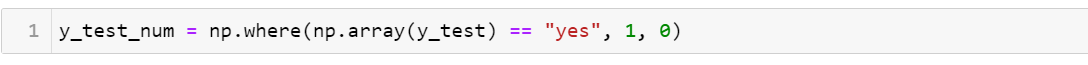
TPR:- True positive rate.

AUC:- Area under the Roc curve. It is the measure of the model.

AUC ranges from 0 to 1, 1 being the prediction is 100% true.



Converting y\_test data from yes and no to 1 & 0



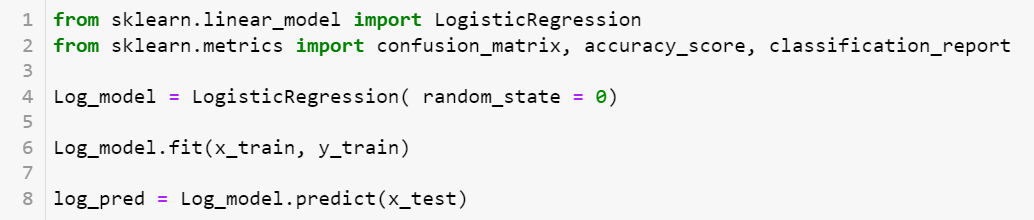
* **Machine Learning models**
* **Logistic Regression:-**

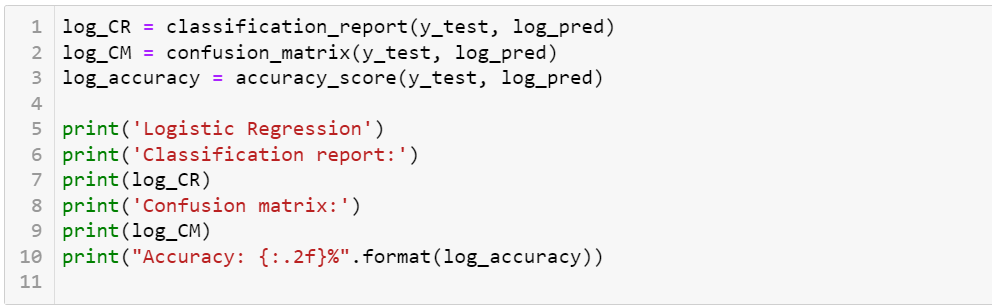
Logistic regression is a type of machine learning algorithm mainly used for binary classification task. (yes/no, pass/fail, 1/0, etc).

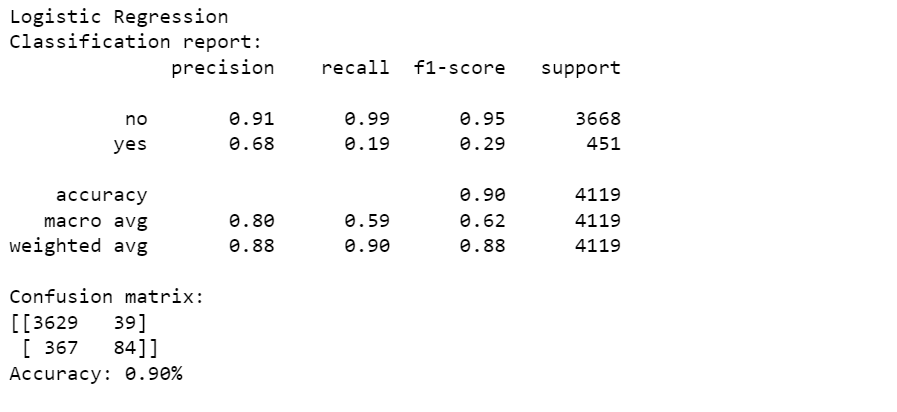
Relationship of input here should be non-linear.

Training data should be adequate. (not all False or all True)

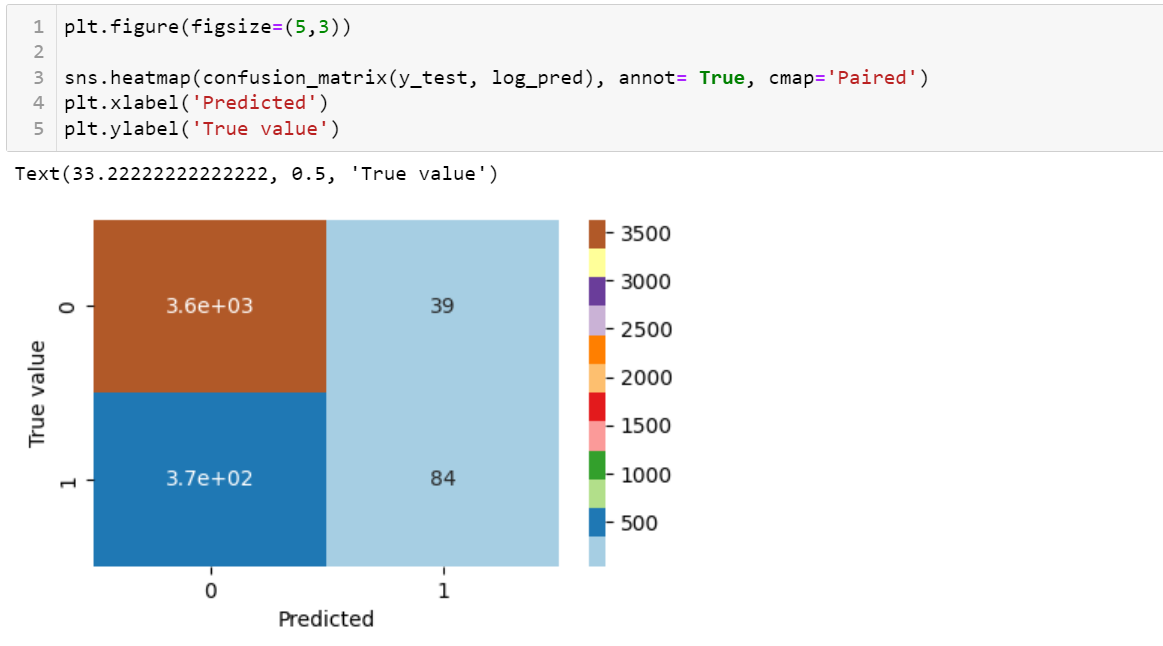
Real limitation on logistic regression is that the output must be discrete.



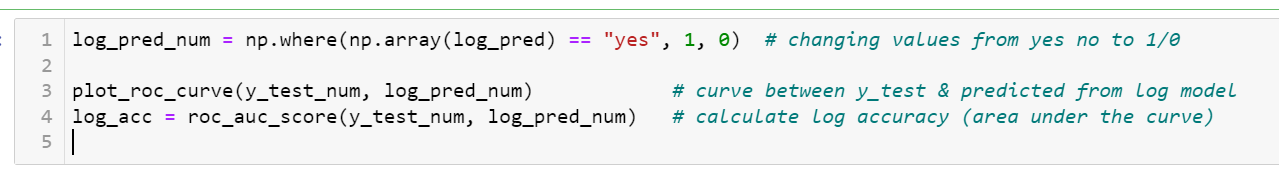


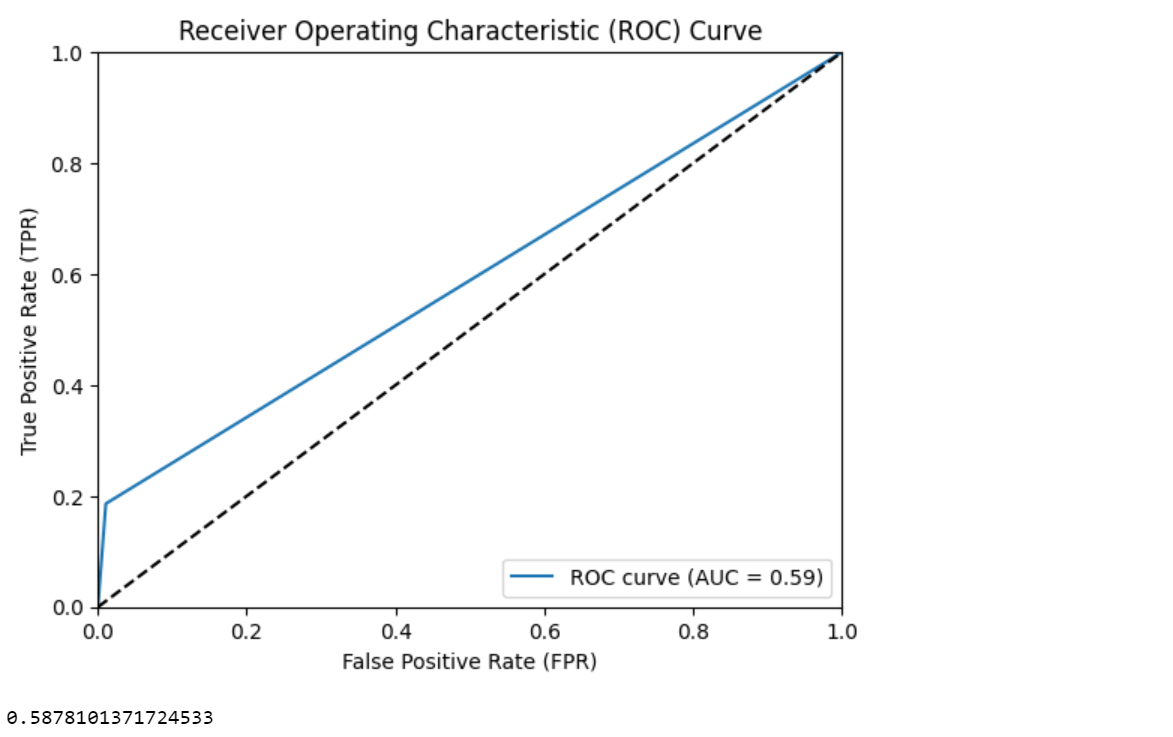


Confusion Metrix between actual and predicted value.

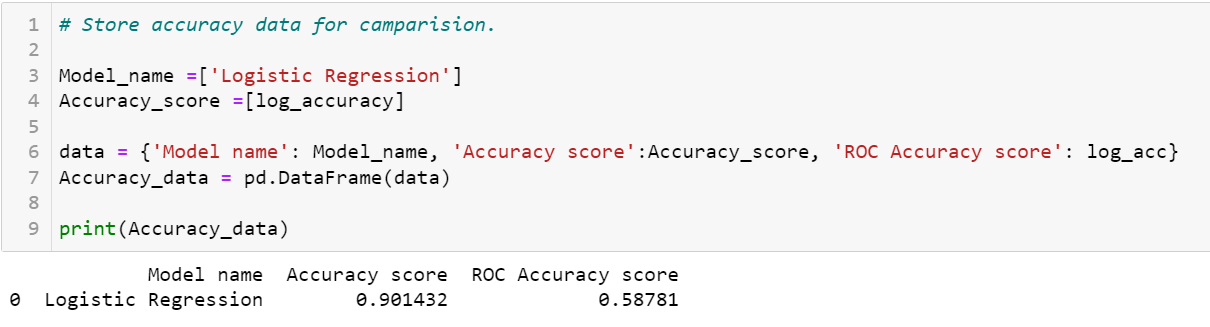


ROC Curve





Store accuracy result for final comparison



* **Decision Tree :-**

Decision tree is a supervised learning algorithm used for both classification and regression.

It takes features from root nodes and take all the features with decision until leaf node.

It has two type of splitting criteria:-

Entropy or ginni index (Purity split) & Information gain (feature DT split).

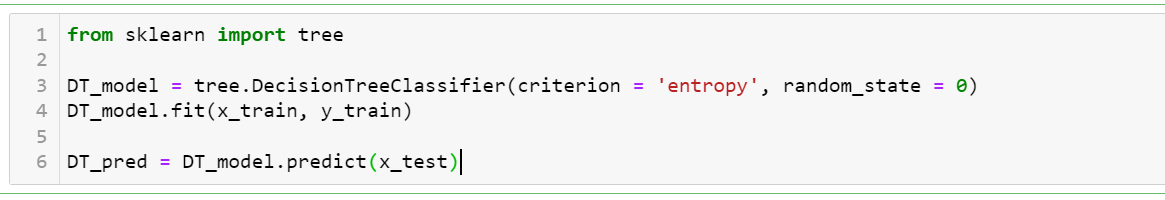
Entropy range is between 0-1, ginni index is 0-0.5

Information gain is used when the output is not discrete. It uses variance as a measure of spread.

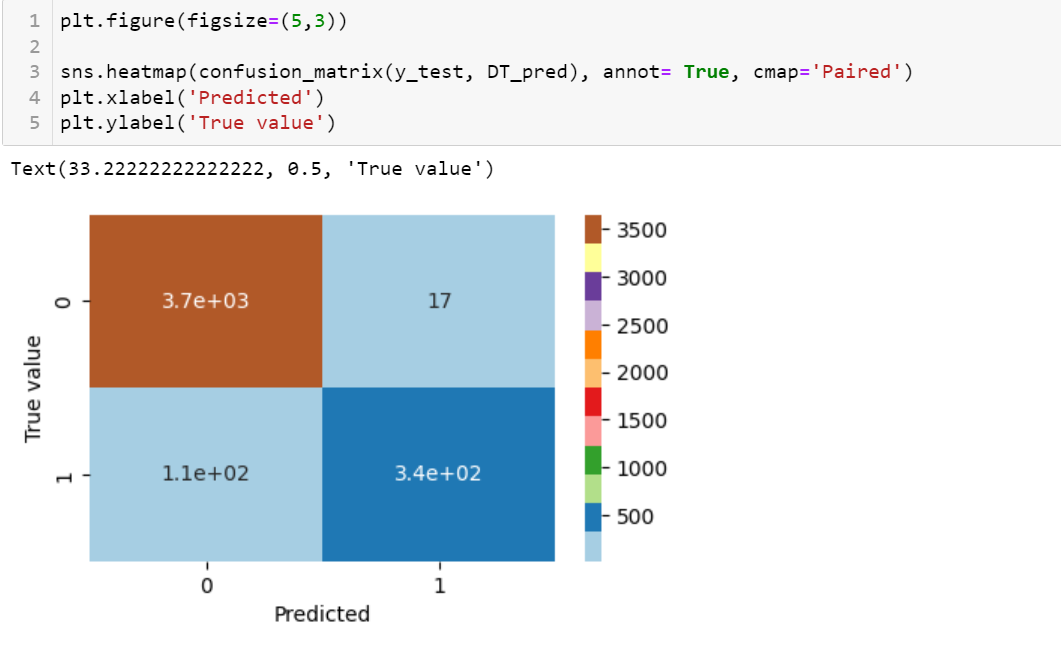
In this model I have selected criteria to be entropy.

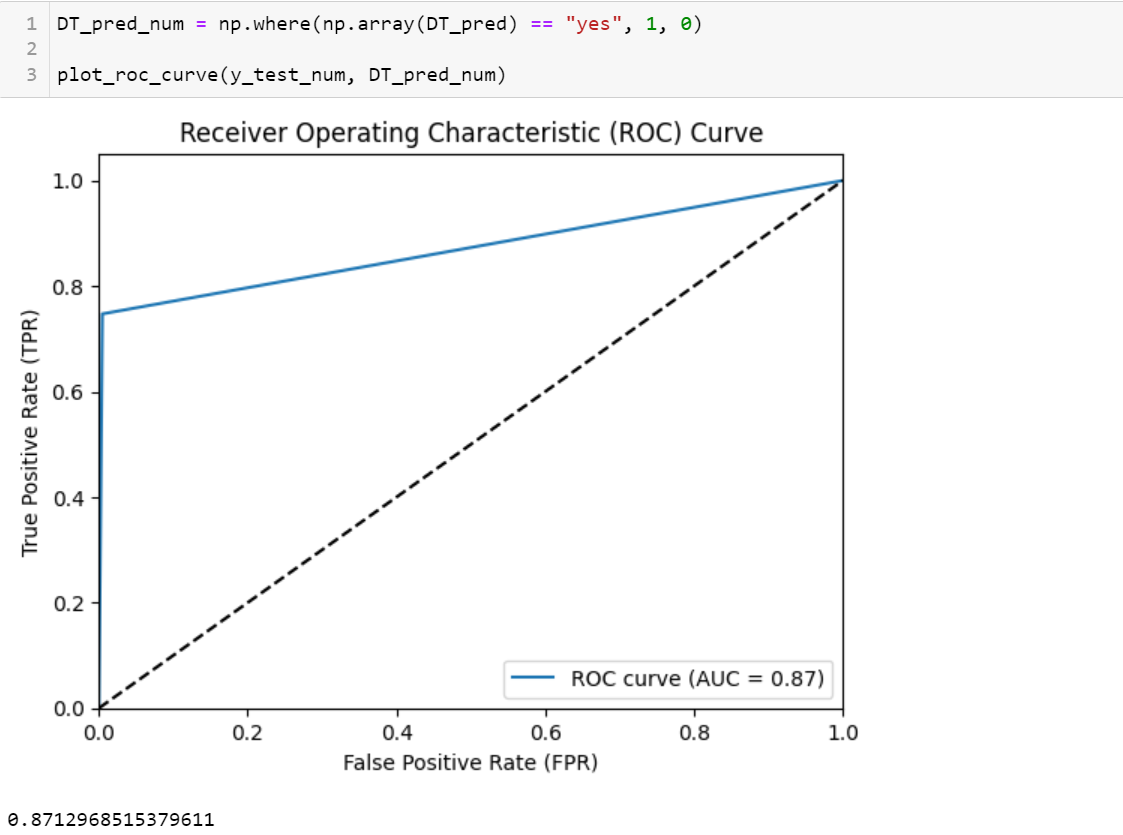
Entropy 0:- Pure split, sure about the event to occur or not.

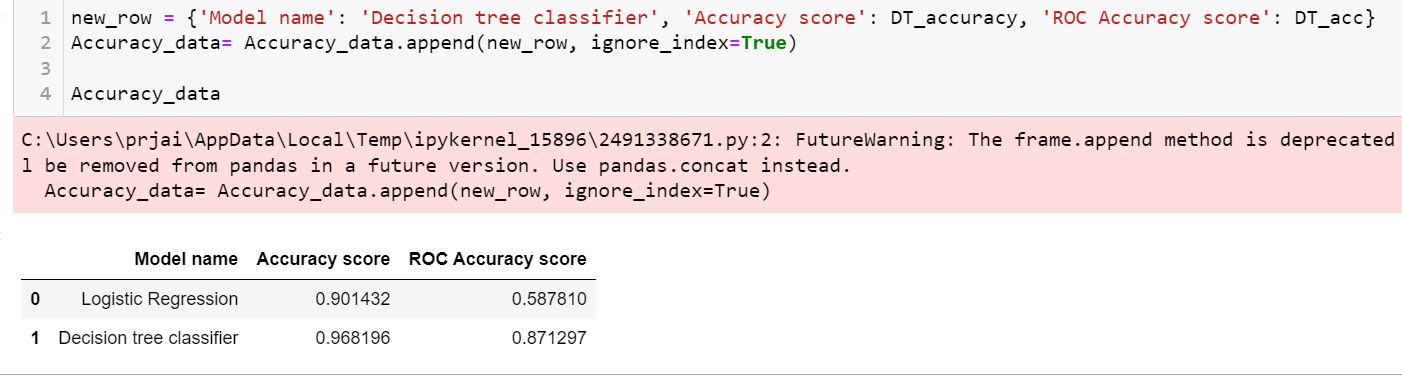
Entropy 1:- Total uncertainty.







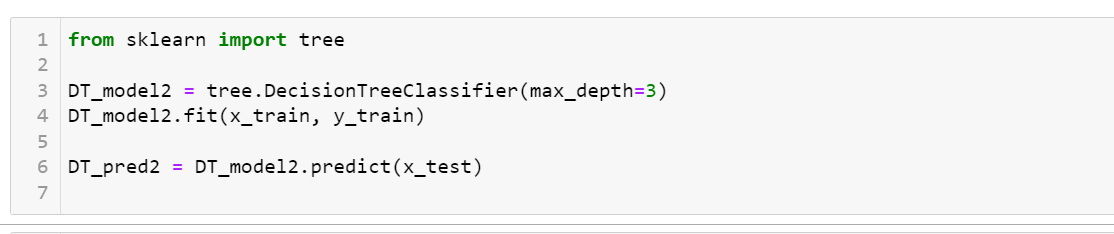


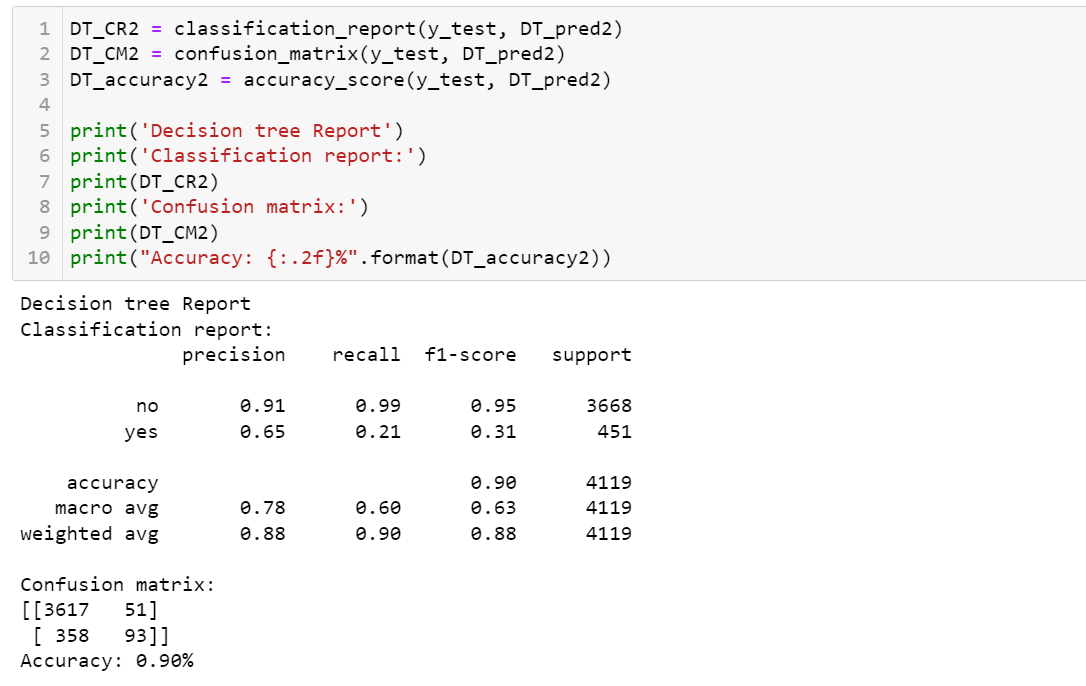


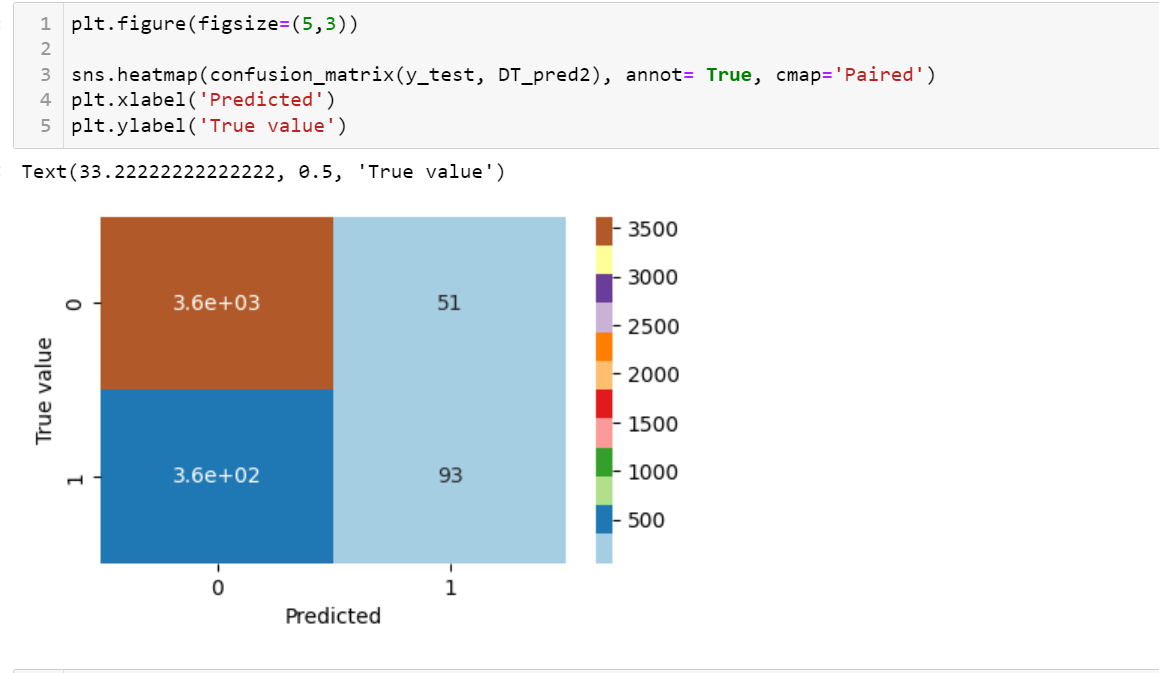
* **Decision Tree with max depth of 3 :-**

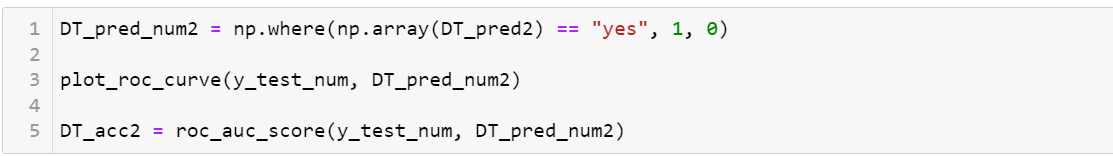
Limitation of decision tree is overfitting, or greedy approach. Hence depth of the tree can be limited, by which we have some uncertainties and don’t have to learn everything.

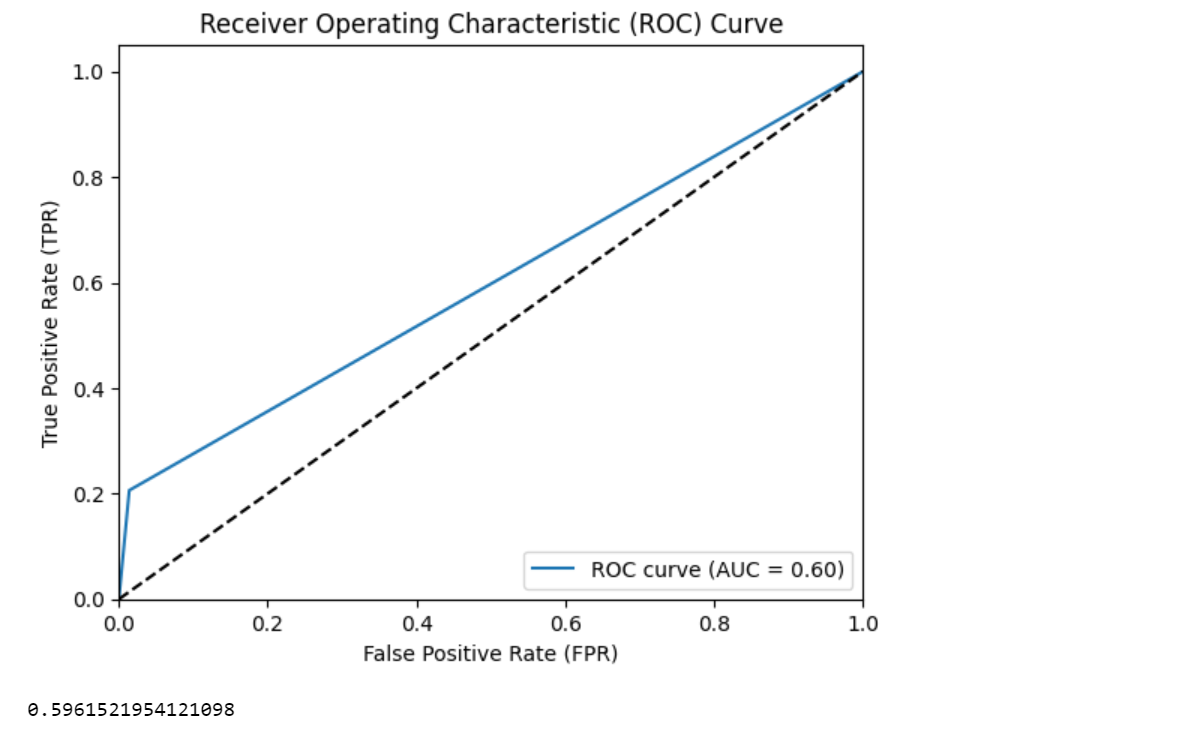
Max depth is introduced to prevent tree to be over complicated, and when the maximum depth is reached the leaf node is assigned to the most common occurring value in the path.

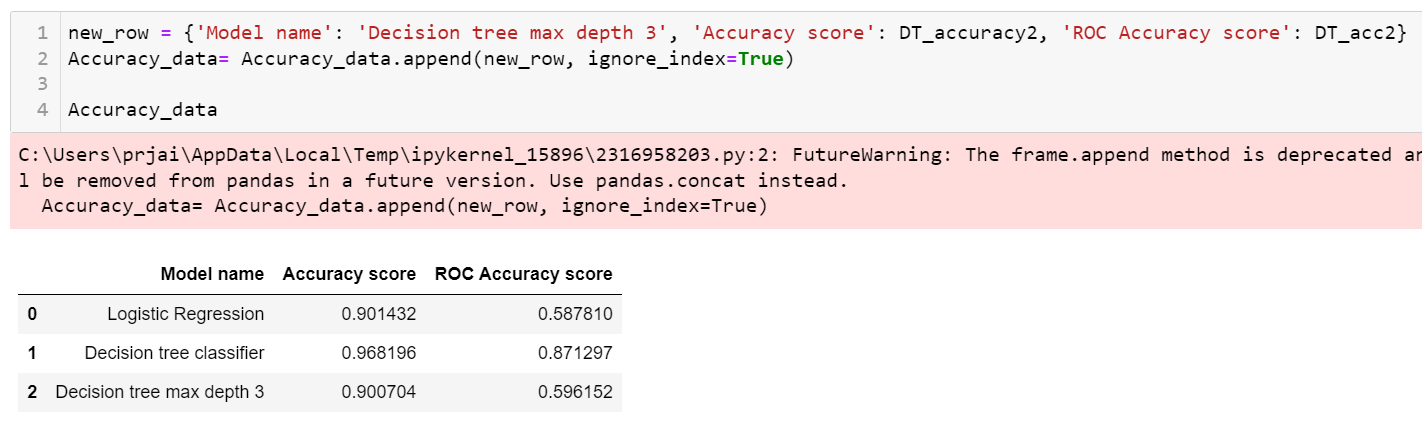


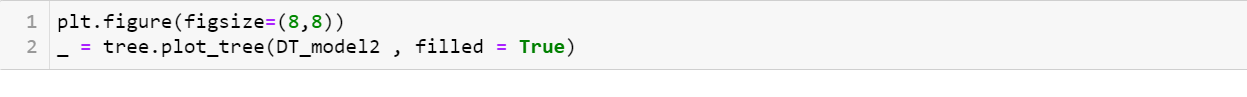


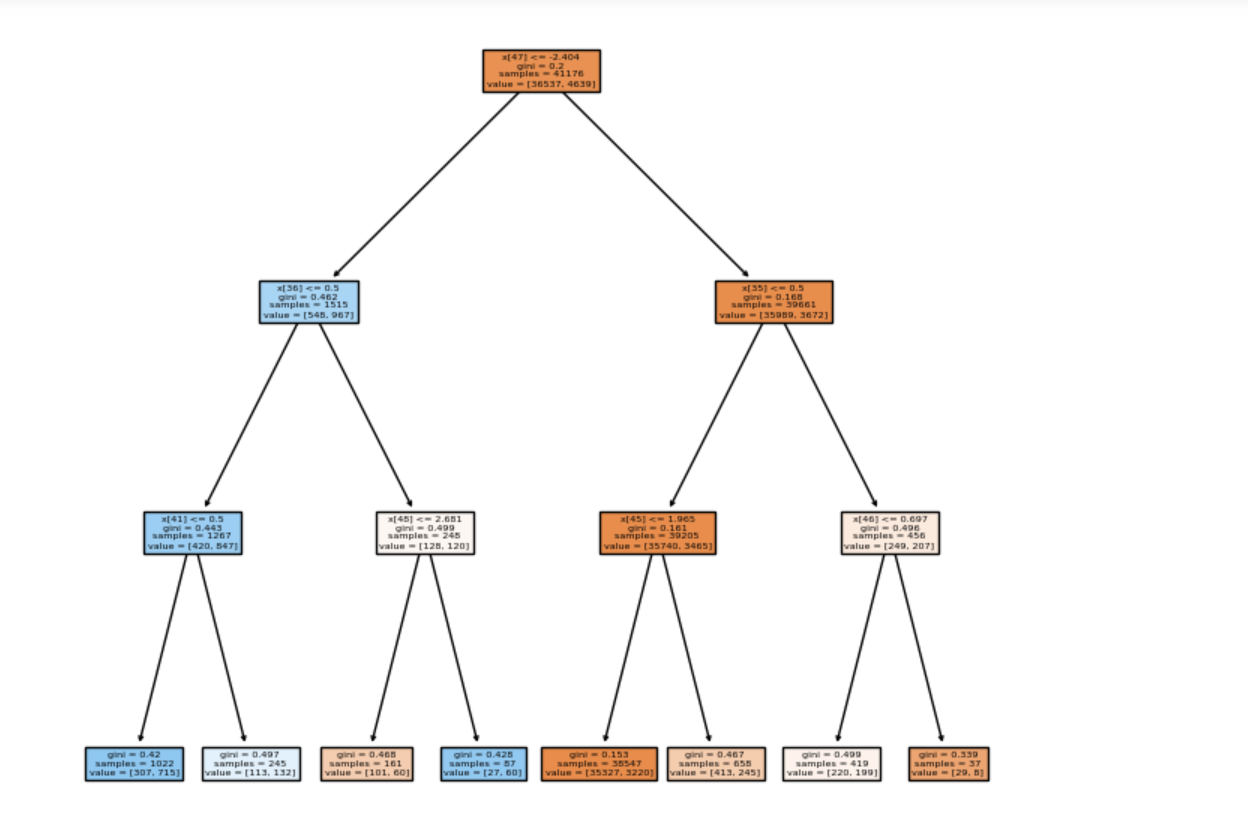












Accuracy is reduced by 7 when max depth is set.

Decision tree visualization is way too complex without setting any depth.

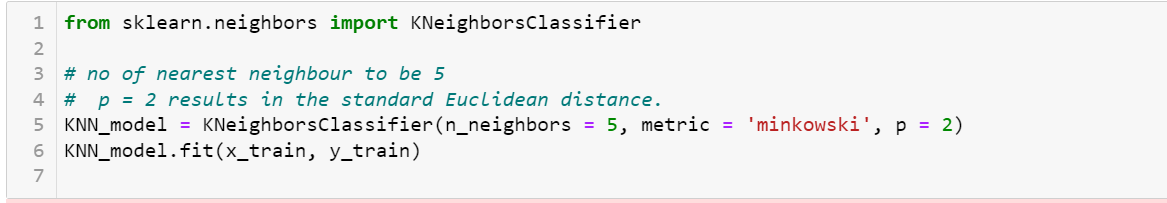
* **KNN - K Nearest Neighbour :-**

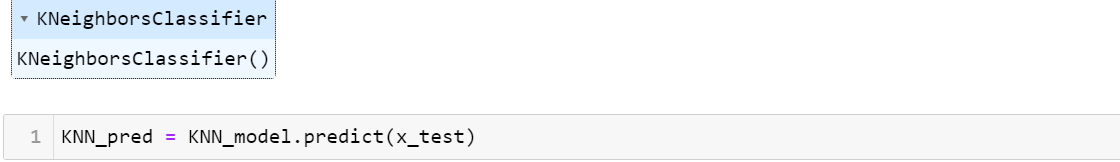
KNN is a simple algorithm that stores all the available cases and classify new cases bases on a similarity measurement. It is based on the idea that similar data points tend to belong to the same class or have similar values.

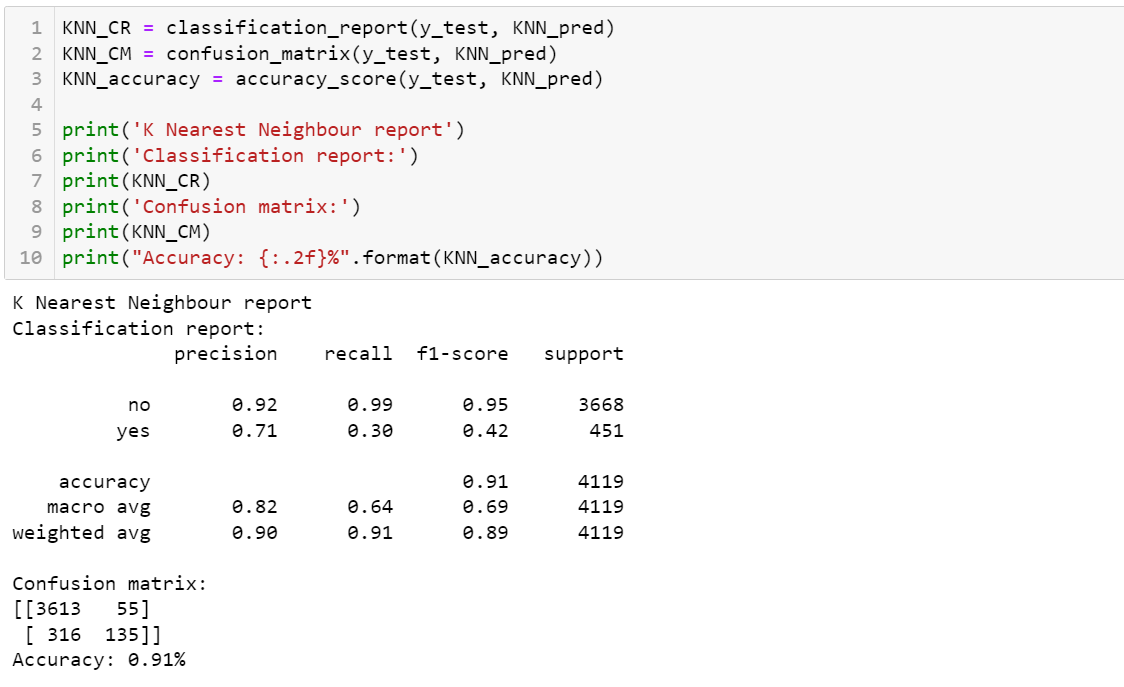
It is also known as lazy learning algorithm, as it stores the training dataset (rather than learning) , and predicts new data bases on similarity.

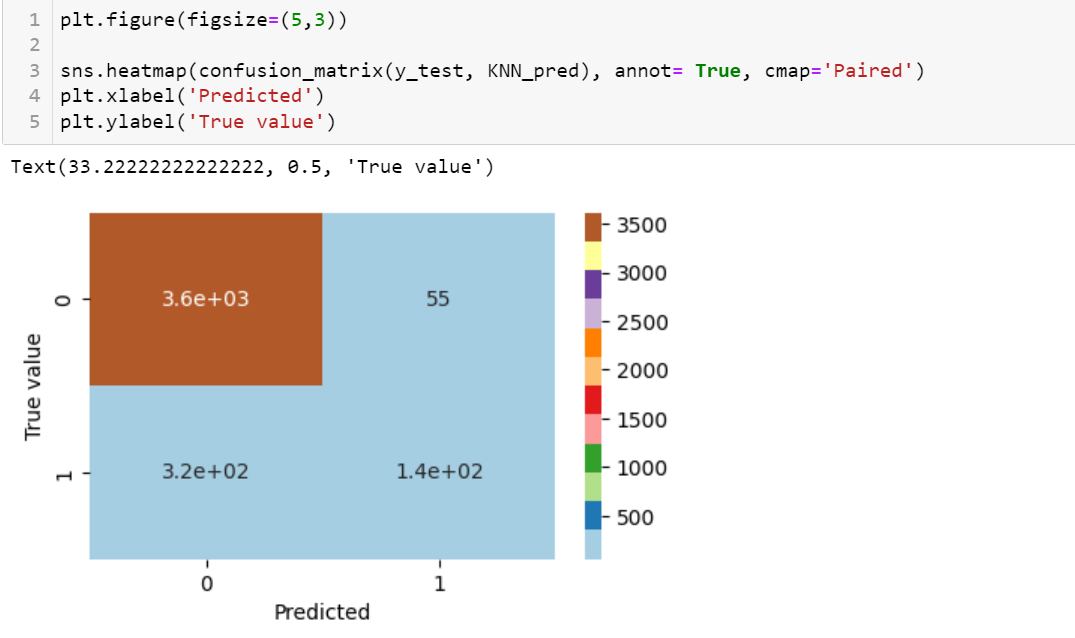
Measure of similarity is distance (Euclidean distance: - distance between two points)

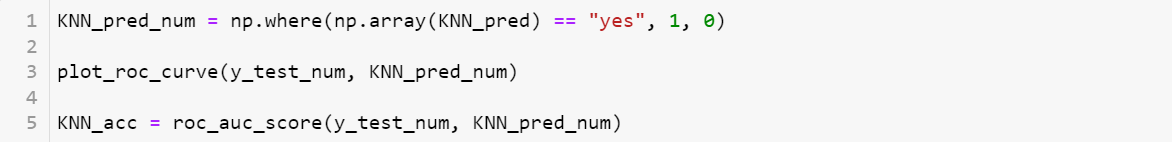
Major Drawback:- It uses the concept of majority voting rather than probability of each class.

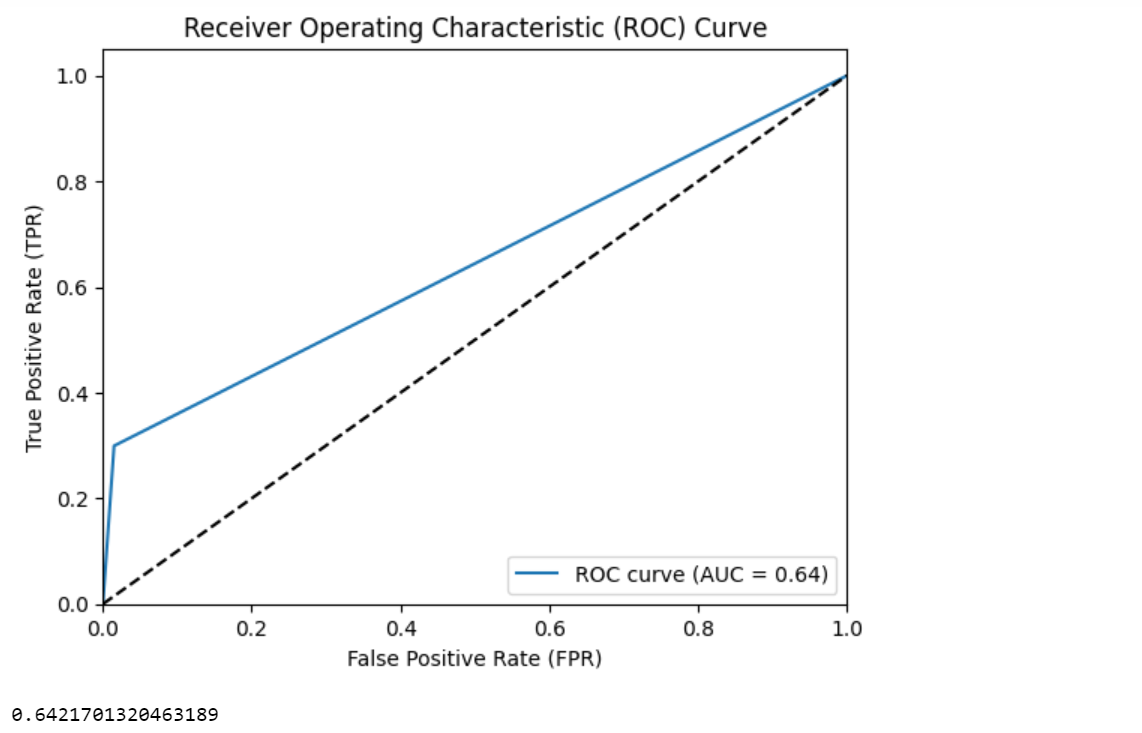


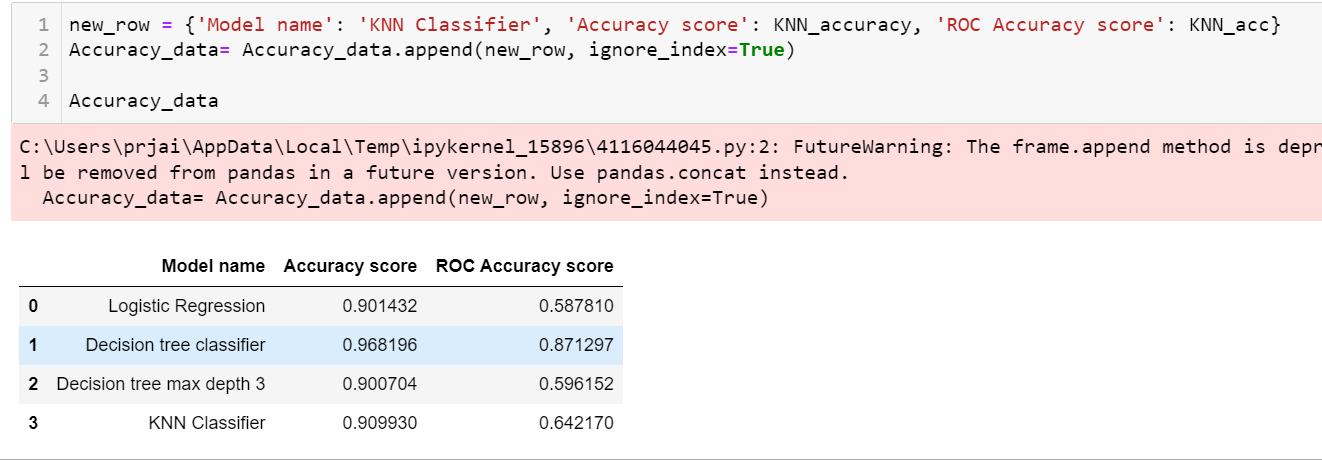






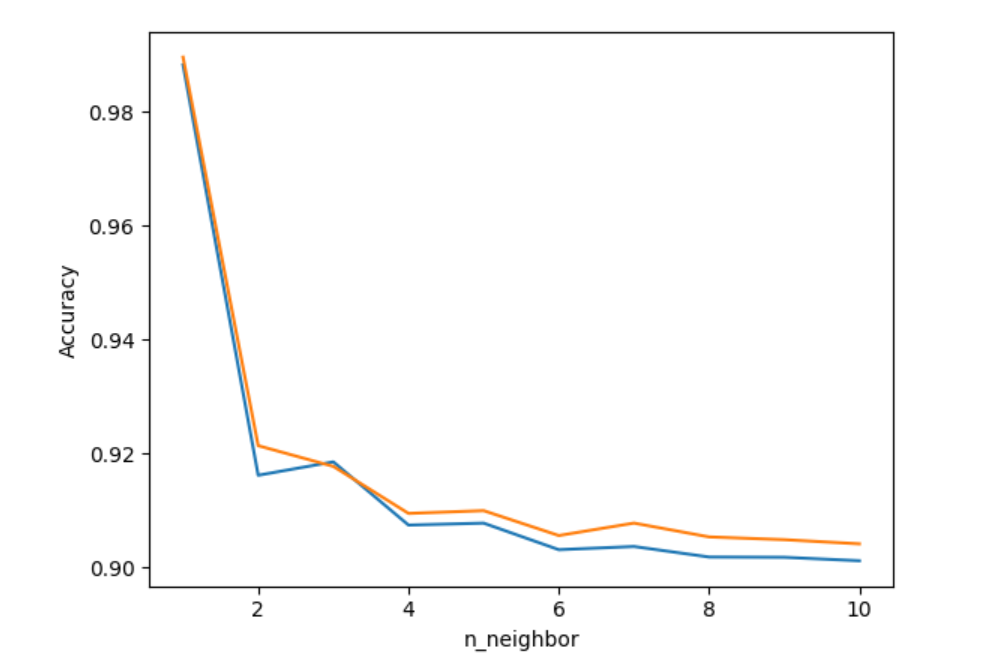






Performance of KNN for different no of nearest neighbour.



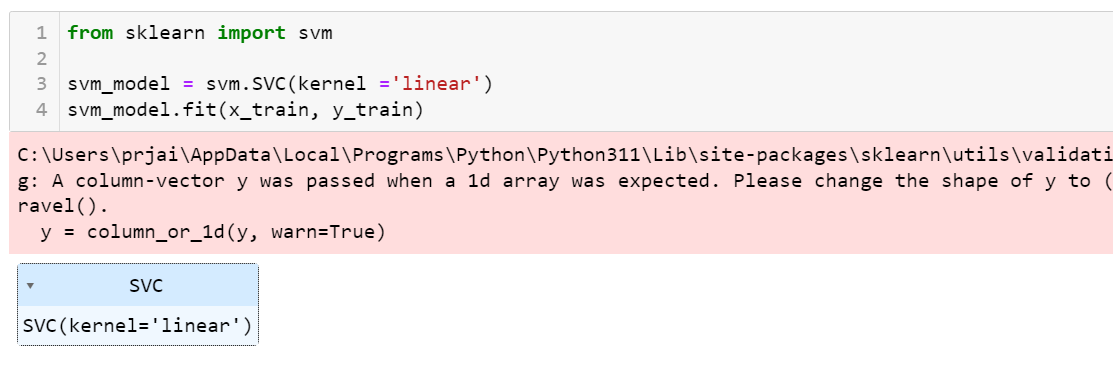


* **SVM Support Vector Machine :-**

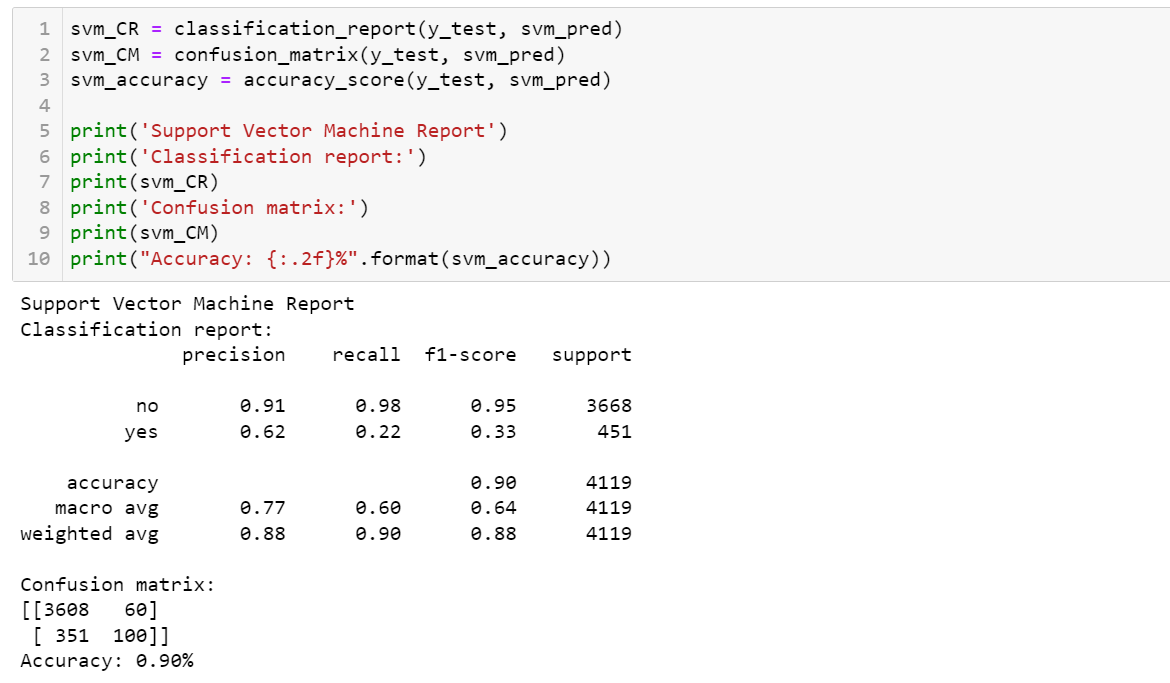
Support Vector Machine is a rule-based approach. SVM finds the support vector that best separates the data point to different classes in high dimensional plane.

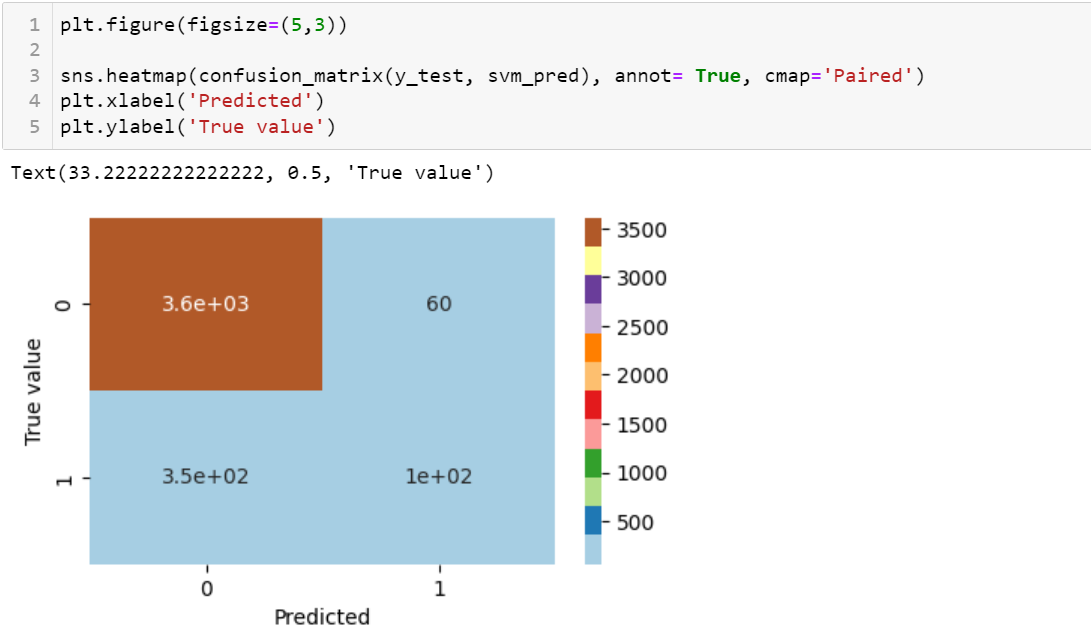
Linear SVM works on a approach to maximize the margin , distance between support vector and nearest data points from each class.

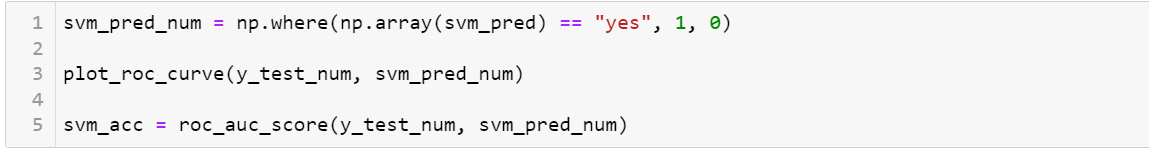
Best fit line is the one that passes through the nearest point known as support vectors.

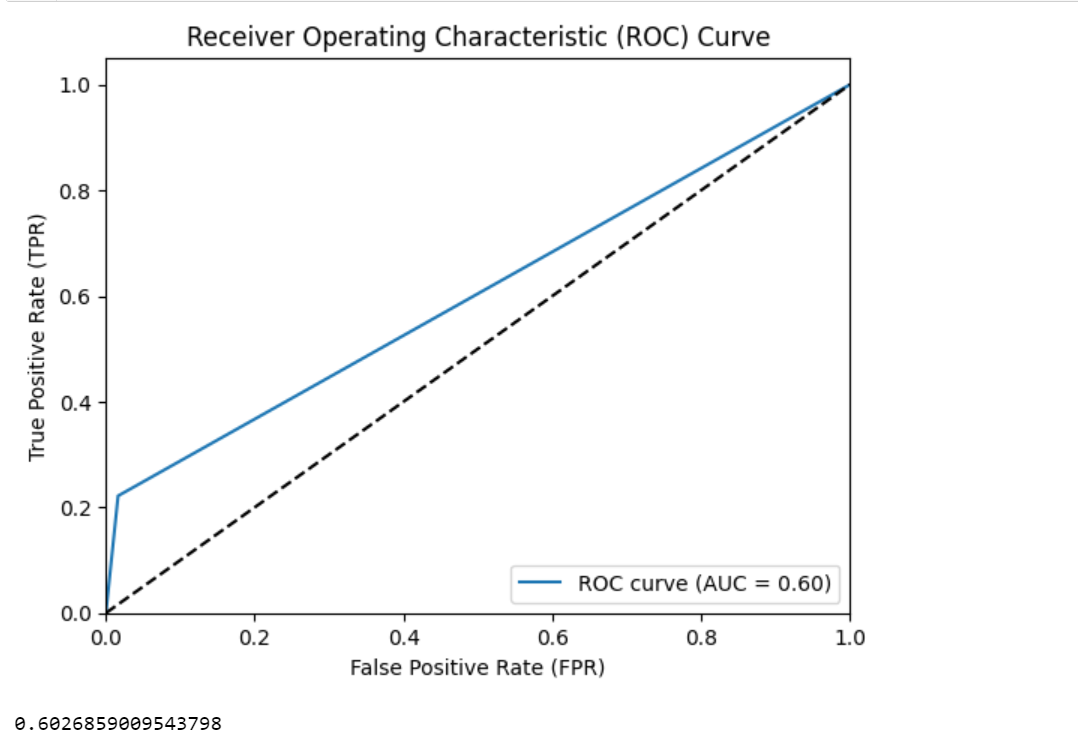


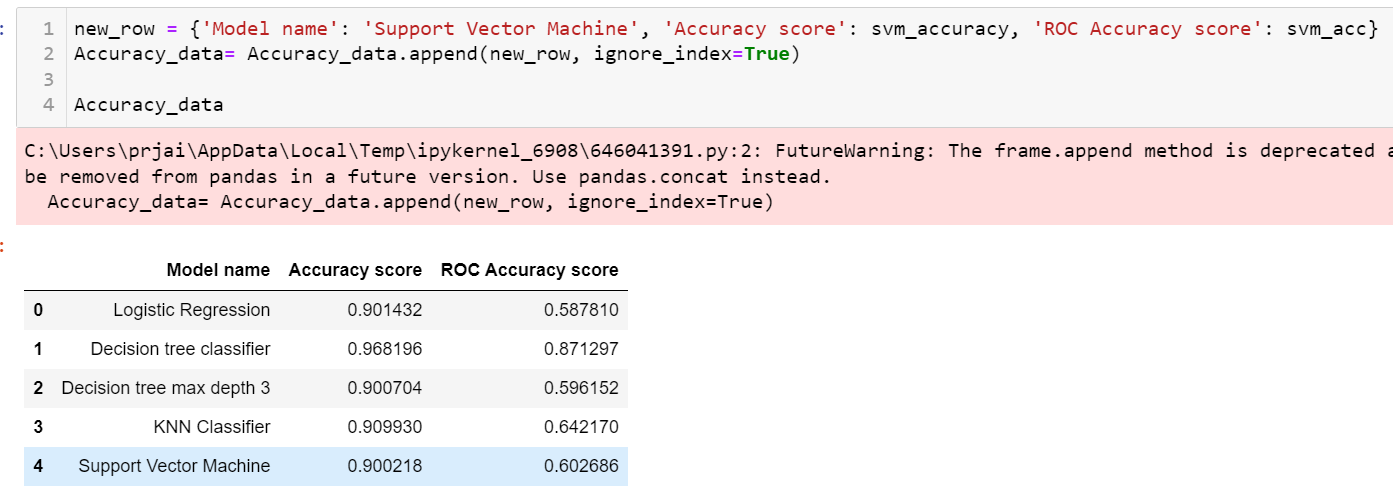












* **Gaussian Naïve Bayes :-**

Naïve Bayes is a probabilistic classifier which is bases on Bayes theorem which is a fundamental concept in probability theory.

Naïve Bayes works under a couple of key assumptions:

The classes are mutually exclusive and exhaustive.

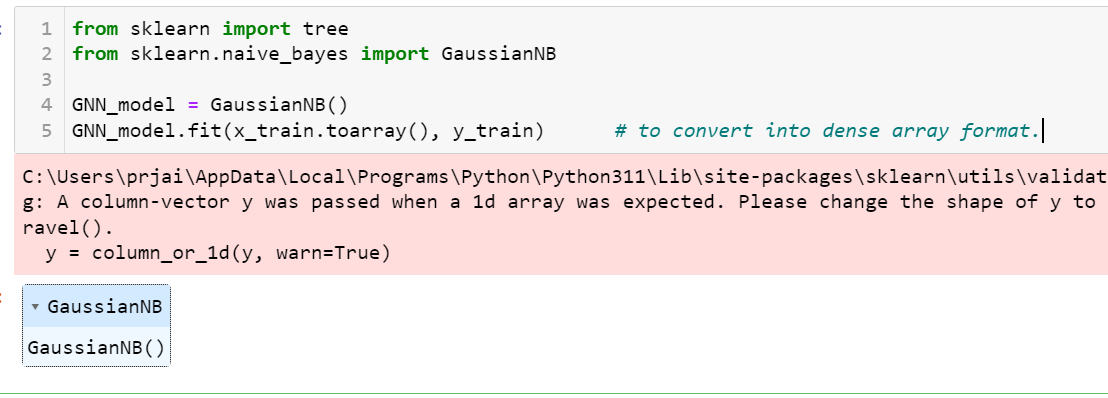
The features are independent given the class.

Every object is must be classified to only 1 feature.

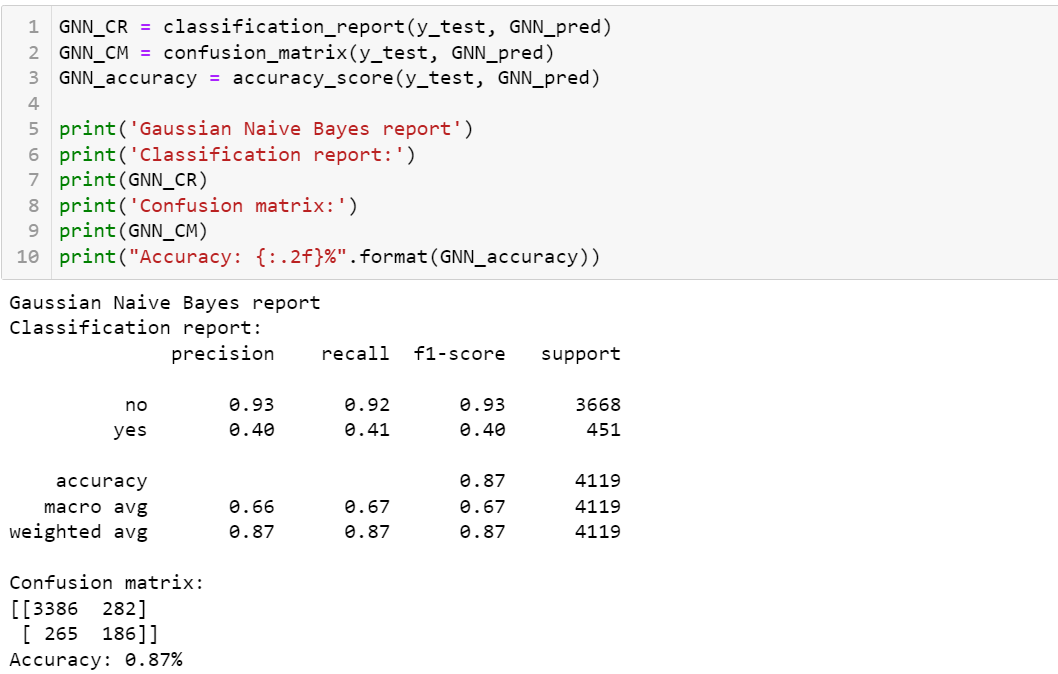
Limitations of naïve bayes includes that, It relies on all attributes being categorical, (difficult in real life situation) &

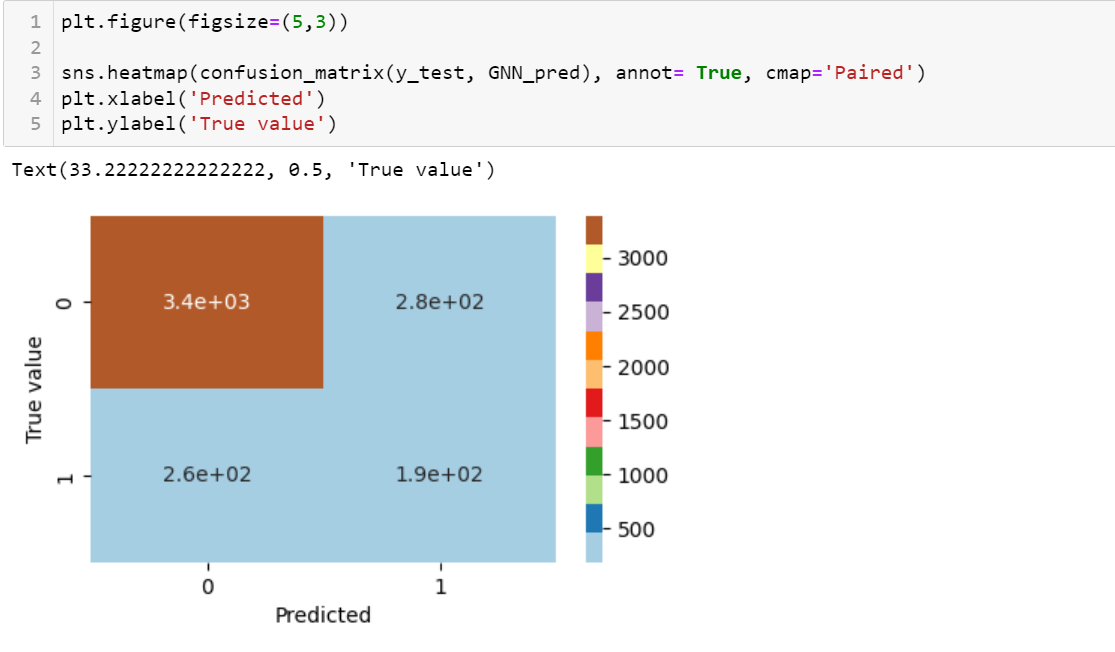
If data is less then it estimates very poorly.

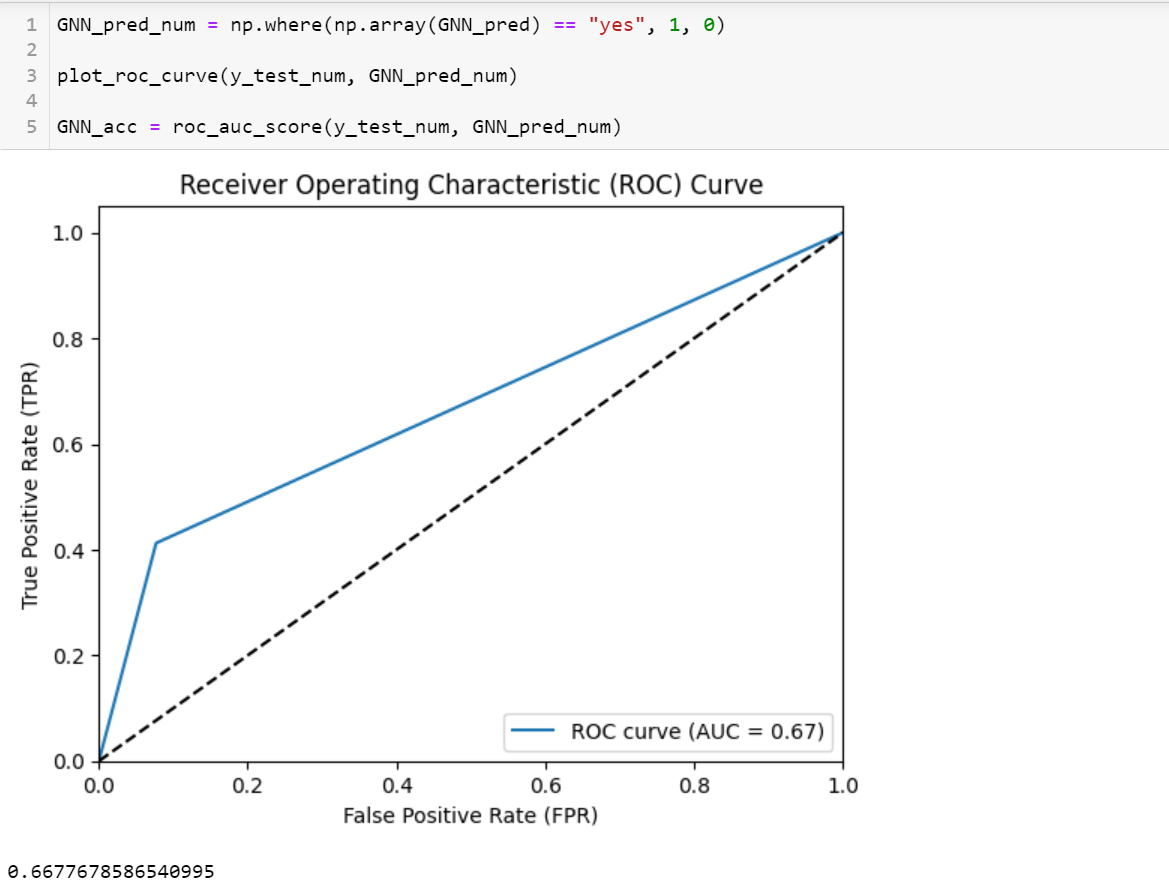
Gaussian Naïve Bayes assumes the features follow normal distribution.

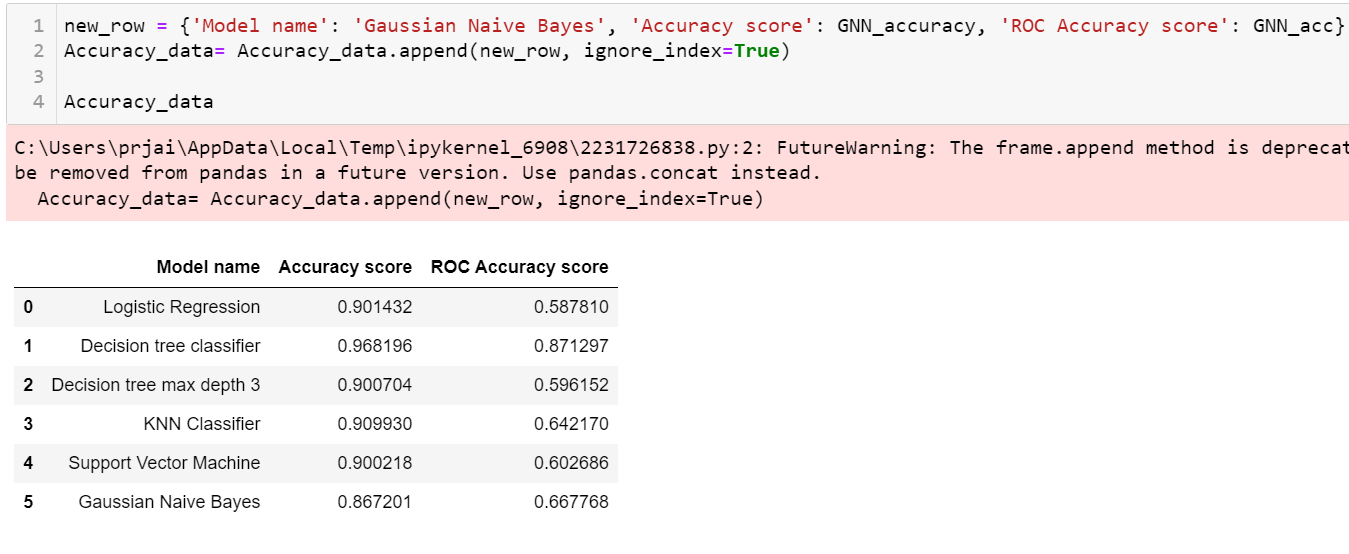








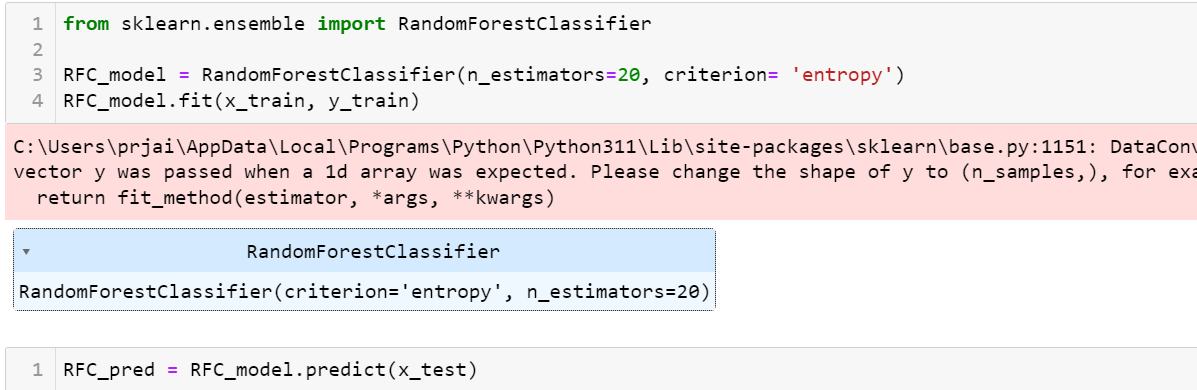


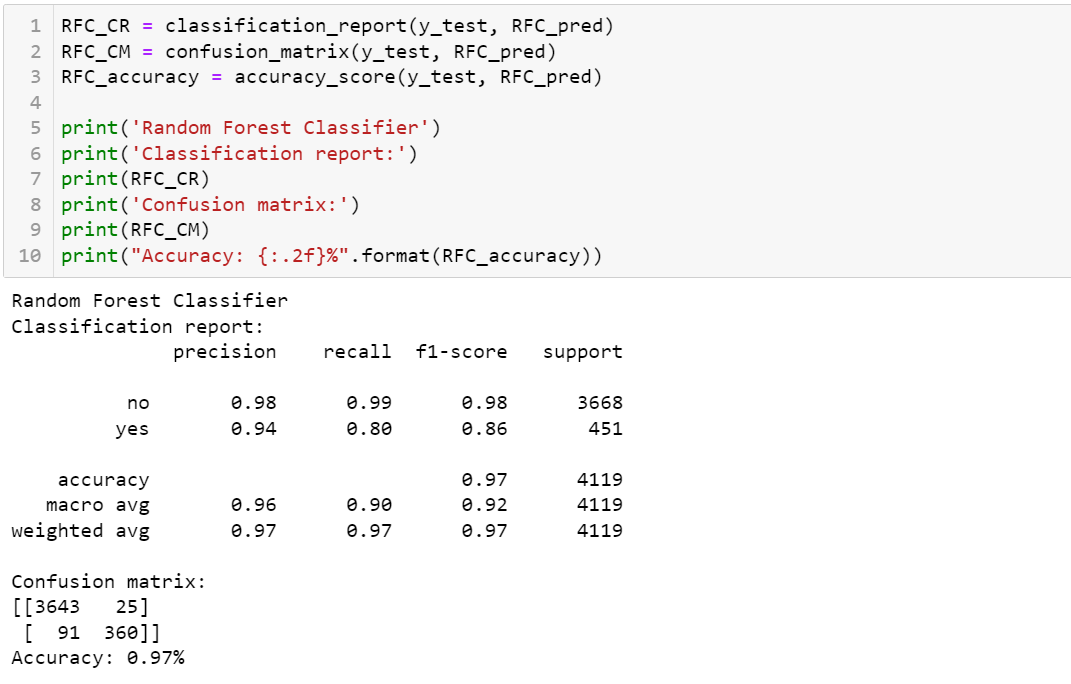


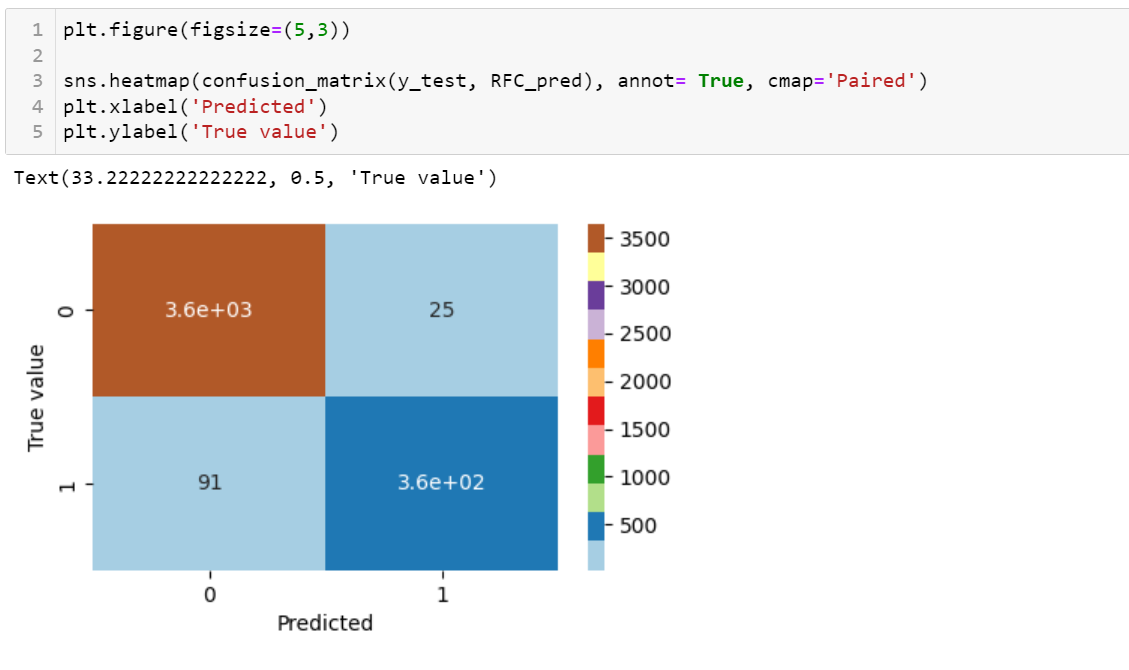
* **Random Forest :-**

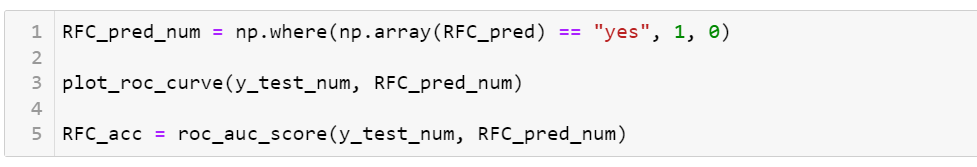
Random forest is an ensemble of decision tree, where each tree is trained on randomly selected features. Instead of searching the most important feature while splitting the node, it searches for the best features among the random subset of features.

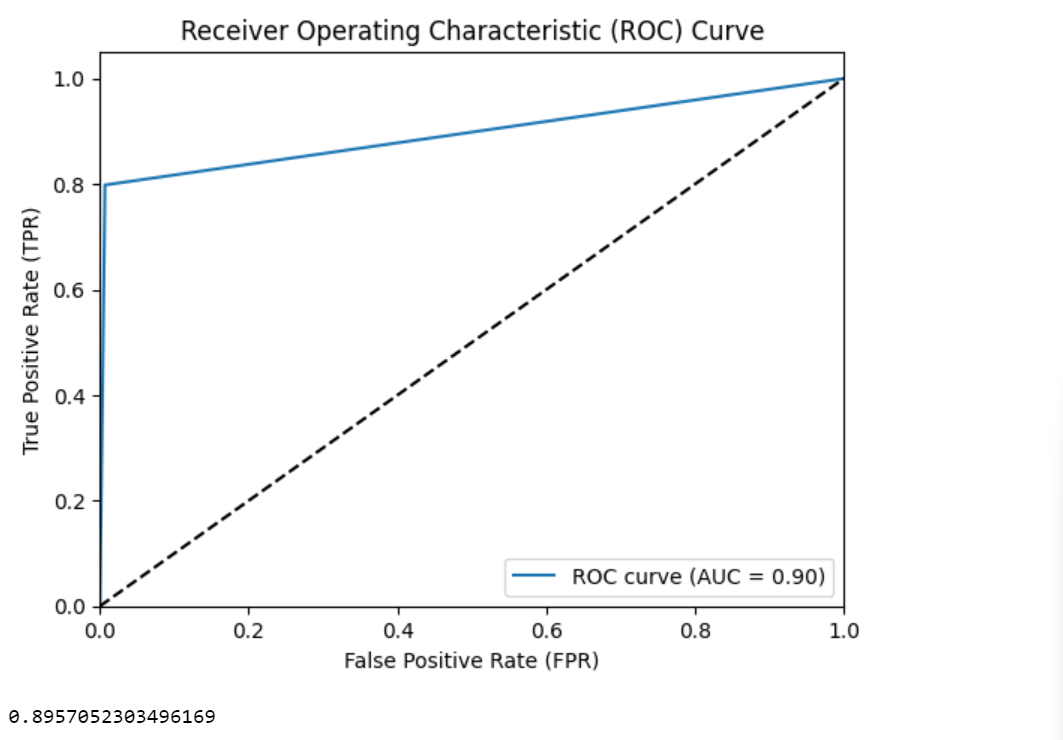
This randomness introduces variability among individual among individual trees, to reduce the risk of overfitting.

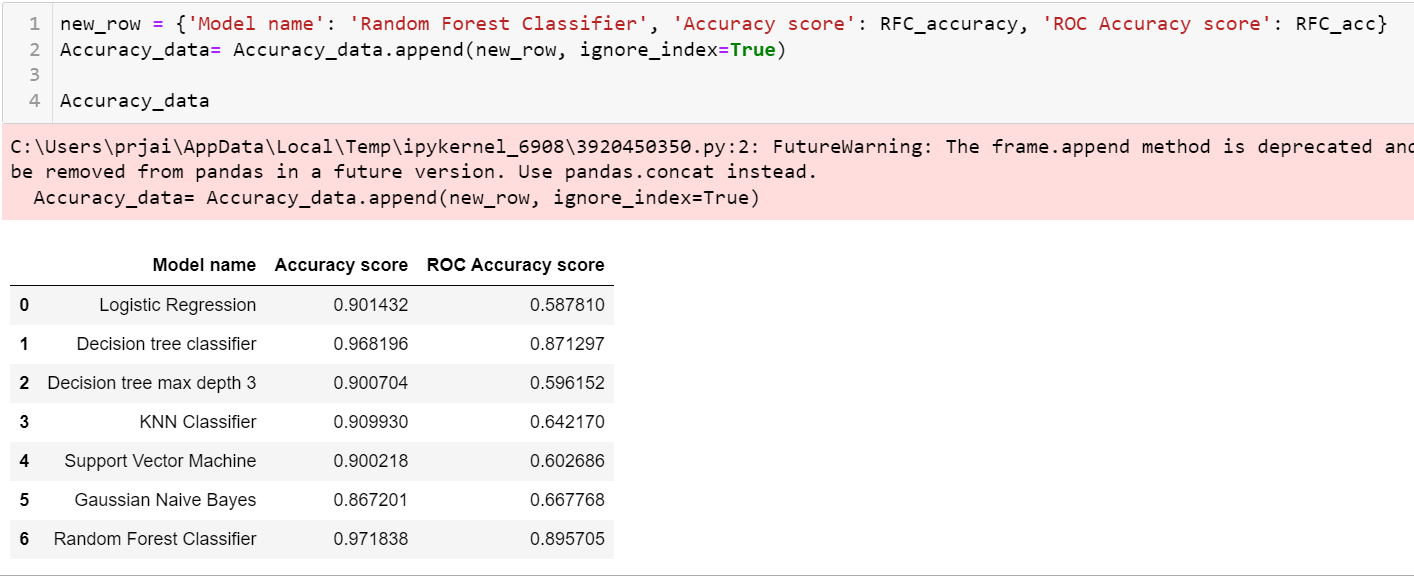










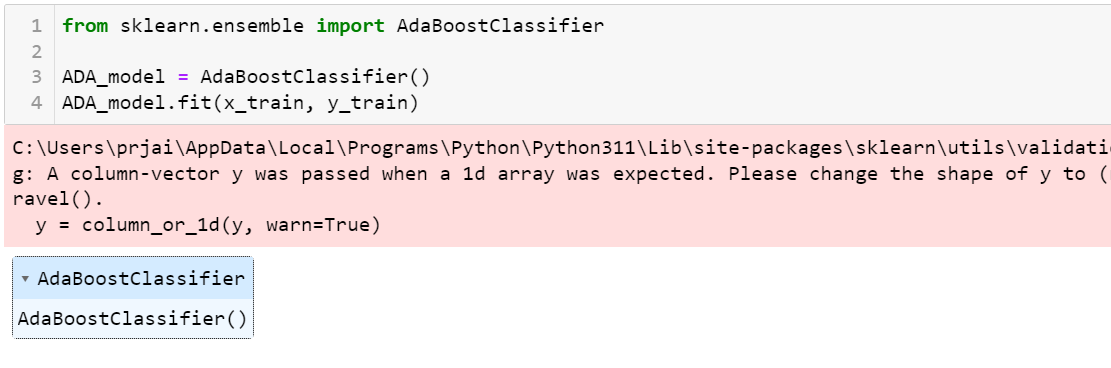


* **AdaBoost Classifier :-**

AdaBoost is short for Adaptive Boosting. It builds a strong classifier by combining multiple poorly performing classifier so that you will get high accuracy strong classifier.

It improves it performance by focusing on mistakes made by previous learners.

Any machine learning algorithm can be used as base classifier if it accepts weights on the training set.

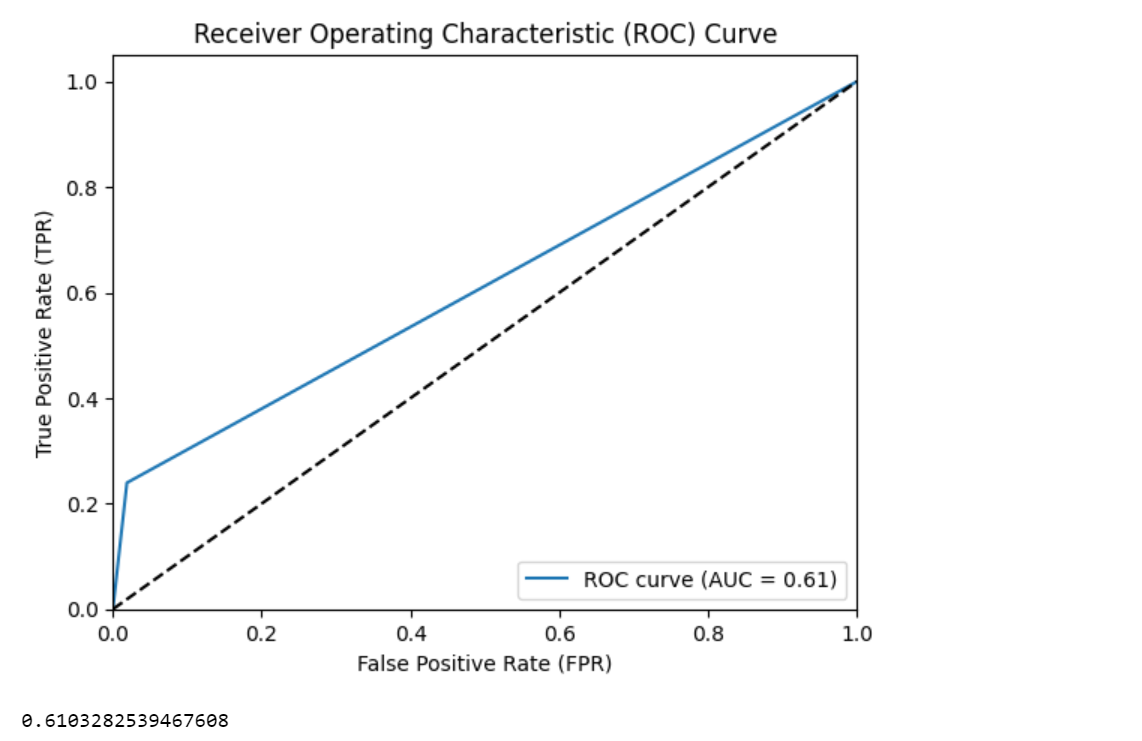


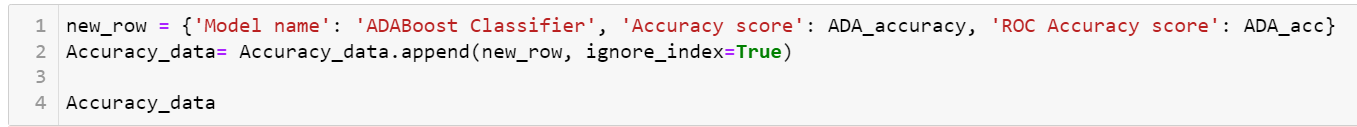


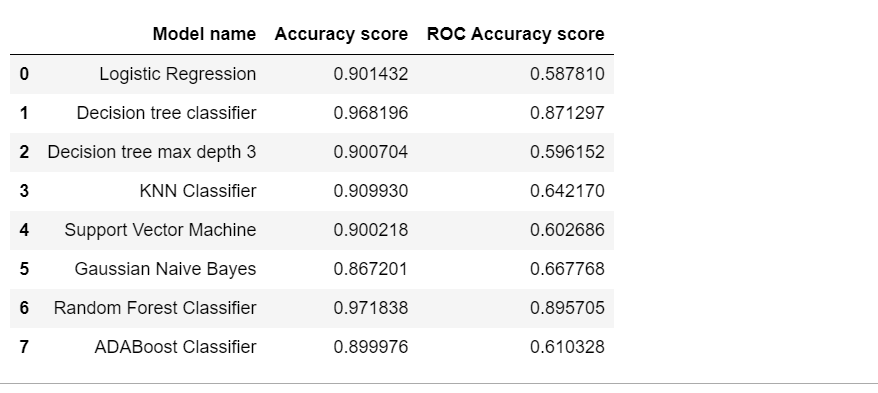












**Conclusion**

Based on the evaluation of various classification models using the provided bank marketing dataset, it is evident that the Random Forest Classifier outperforms other models in terms of both accuracy and ROC accuracy score, achieving accuracy score of 97.18% and ROC accuracy score of 89.57%. This indicates the robustness and effectiveness of a Random Forest approach in predicting whether clients will make deposit or not based on demographic and marketing related features. Although some models such as the Decision Tree Classifier and ADABoost classifier also demonstrate strong performance, they fall short compared to the Random Forest Classifier. Therefore, Random Forest is recommended for practical applications in assisting the bank’s marketing strategies to target potential depositors more accurately and efficiently.