# Logistic Regression



**Instructions:**

Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

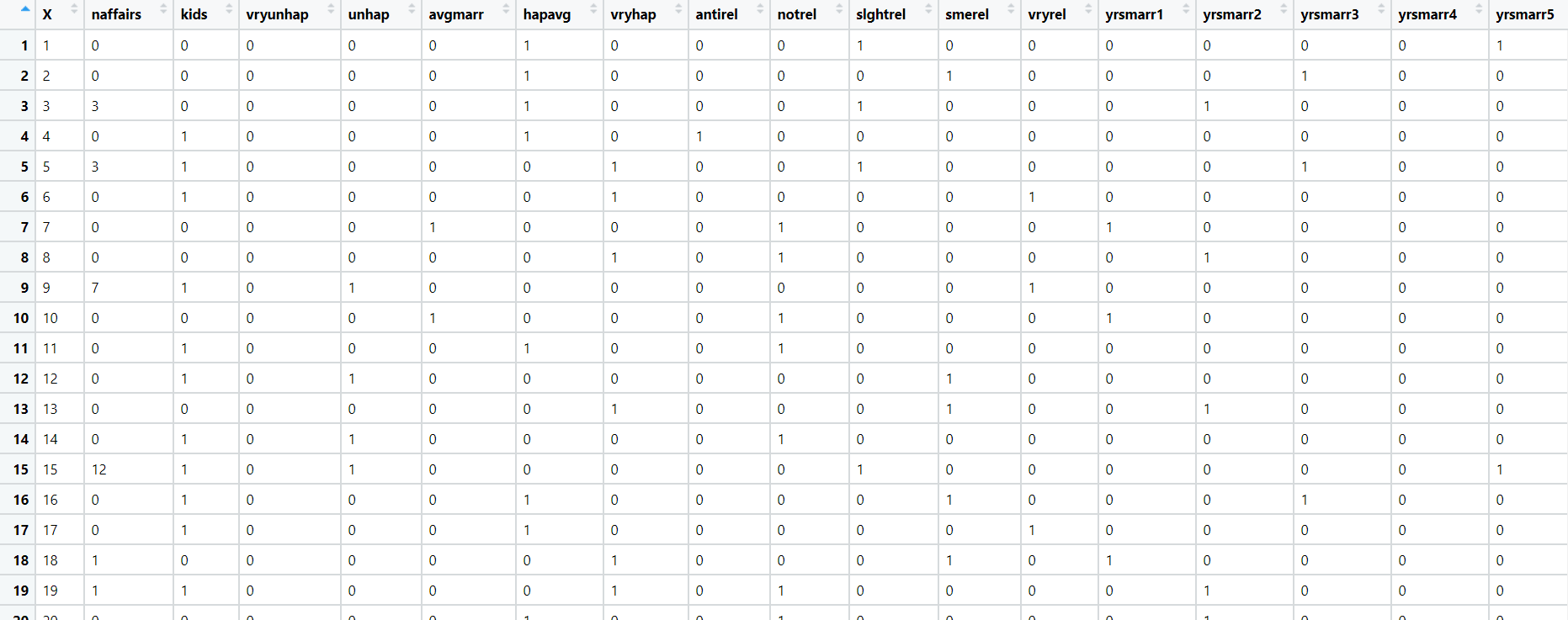
**Name: \_Prajay B. Urkude Batch ID:** \_16092021

**Topic: Logistic Regression**

A screenshot of a cell phone

Description automatically generatedProblem Statement: -

1. A psychological study has been conducted by a team of students at a university on married couples to determine the cause of having an extra marital affair. They have surveyed and collected a sample of data on which they would like to do further analysis. Apply Logistic Regression on the data to correctly classify whether a given person will have an affair or not given the set of attributes. Convert the naffairs column to discrete binary type before proceeding with the algorithm.



Ans: Business Objective:

To build the model to identify the given person have the external affair based on the given attributes.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Features** | **Description** | **Type** | **Relevance** |
| naffairs | External affair yes =1, no =0 | Quantitative, Nominal | Relevant |
| kids | Have Kids yes =1, no =0 | Quantitative, Nominal | Relevant |
| vryunhap | Veryunhappy marriage | Quantitative, Nominal | Relevant |
| unhap | Unhappy marriage | Quantitative, Nominal | Relevant |
| avgmarr | Average marriage | Quantitative, Nominal | Relevant |
| hapavg | Happy averahe marriage | Quantitative, Nominal | Relevant |
| vryhap | Very happy marriage | Quantitative, Nominal | Relevant |
| antirel |  | Quantitative, Nominal | Relevant |
| yrsmarr1S | One year complete yes =1 no =0 | Quantitative, Nominal | Relevant |
| yrsmarr2 | Two year complete yes =1 no =0 | Quantitative, Nominal | Relevant |
| yrsmarr3 | Three year complete yes =1 no =0 | Quantitative, Nominal | Relevant |
| yrsmarr4 | Four year complete yes =1 no =0 | Quantitative, Nominal | Relevant |
| yrsmarr5 | Five year complete yes =1 no =0 | Quantitative, Nominal | Relevant |
| yrsmarr6 | Six year complete yes =1 no =0 | Quantitative, Nominal | Relevant |

Steps For the Logistic Regression:

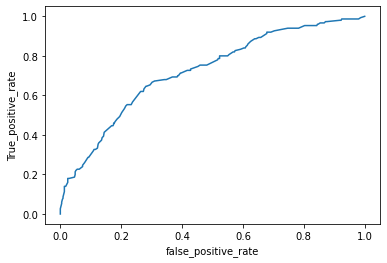
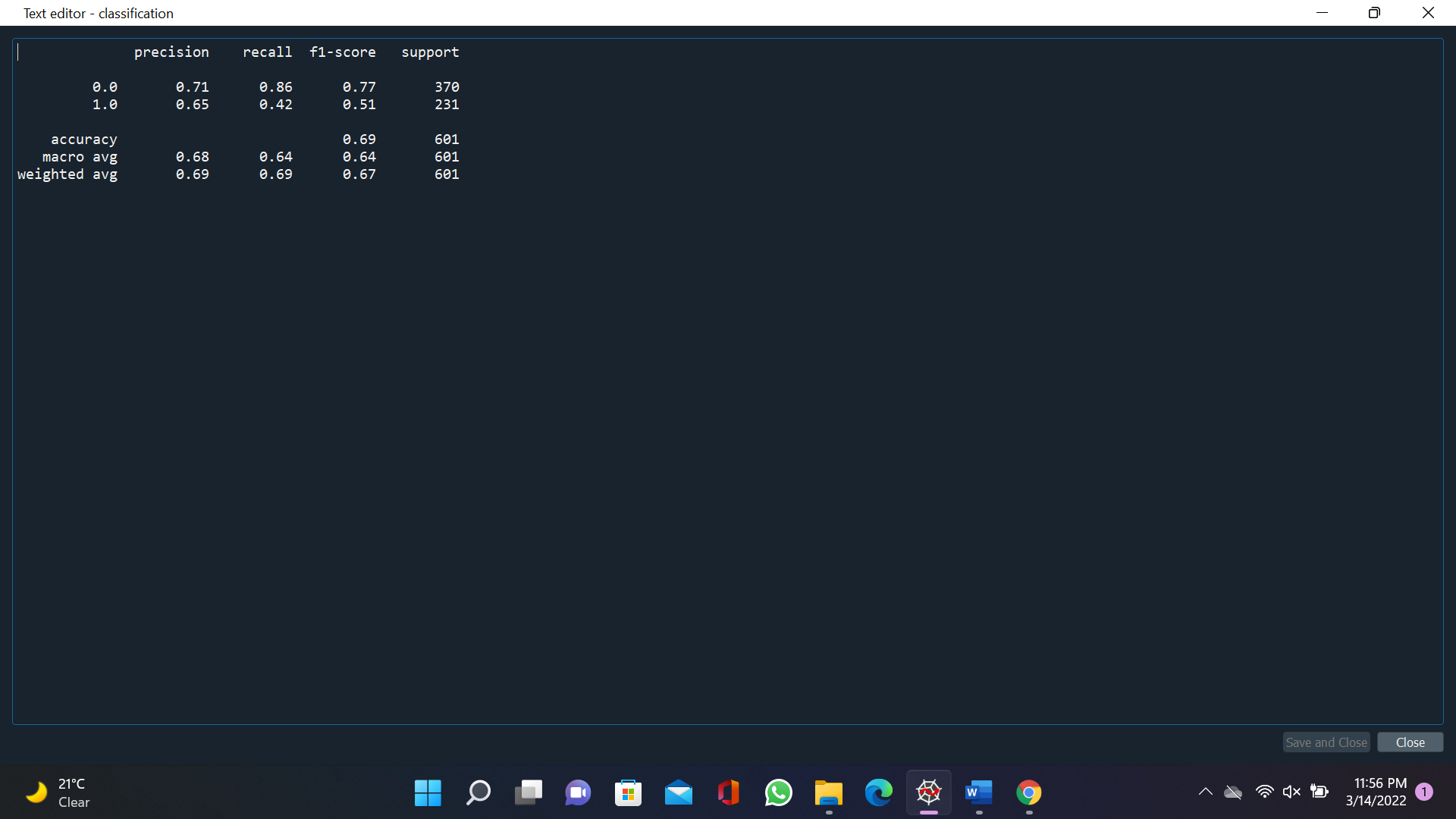
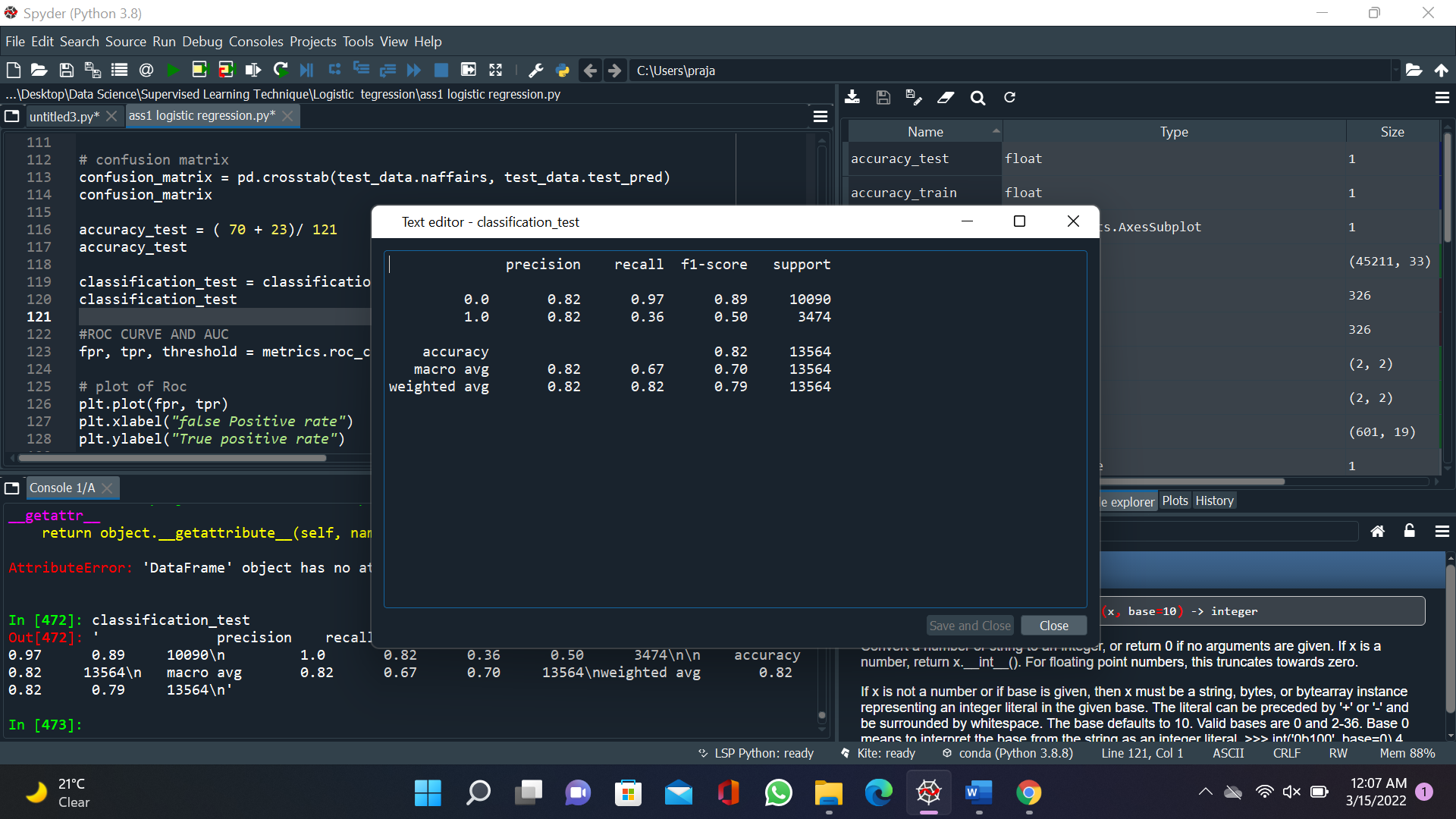
* Import the required libraries. Pandas, numpy, seaborn, matplotlib, scipy.

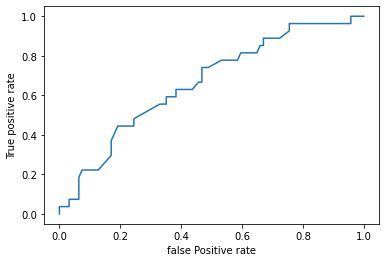
Seborn is the library for the advanced data visualization.

scipy is the library for the scientific calculation and for plotting the Q-Q plot.

From statsmodel library import formulae package.

Statsmodel:- Statsmodels is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration. An extensive list of result statistics are available for each estimator.

* Load the data
* Doing the univariate analysis and Exploratory data analysis.
* Checking the head i.e., top 5 rows of the datasets
* Checking the columns names of the datasets
* Checking the null values if any available in dataset.
* Checking the duplicate values in the datasets
* Checking the information i.e., datatypes of the datasets
* Exploratory data analysis. mean, median, mode, count, min, max etc.
* Check the distribution of the data.
* Dropping the unwanted column which is not useful for the analysis.
* Converting the nonnumerical data into numerical data by using one hot encoding or Label Encoder or pandas get\_dummies function as per the requirement
* Converting the continuous data into discrete form if necessary.
* Model Building
* Define the model by using the logit function and fir it into the data by taking naffairs as the dependent variable and rest of the column as the independent variables.
* Find the summary2 and summary of the model .
* Prediction to be done on the given datasets by using the independent variables.
* Calculate the false positive ration, True positive ratio, threshold value by using the roc\_curve function of the metrics package of sklearn and out of all the value find out the optimal threshold value and for the given datasets we get the threshold value = 0.25
* Plot the roc curve by taking false positive ratio on the x-axis and True positive value on the Y axis.
* Find out the area under the curve which is come out 0.72.
* Filling all the cells with zeros by creating the extra column named pred.
* Append the 1 to the pred value if the prediction value is grater than the optimal threshold value else zero.
* Find out the classification report in which we can find out the accuracy of the models, precision, recall.
* Model building:
* Spilt the data into training datasets and testing datasets in the ratio 80:20.
* Fit the model on the training datasets and predict the model on the test datasets.
* Filling all the cells with zeros by creating the extra column named test\_pred.
* Append the 1 to the test\_pred value if the prediction value is greater than the optimal threshold value else zero.
* Find out the accuracy of the model manually by plotting the confisuon matrix
* Find out the classification report in which we can find out the accuracy of the models, precision, recall.
* Calculate the false positive ration, True positive ratio, threshold value by using the roc\_curve function of the metrics package of sklearn and plot the roc\_curve for the test datasets.



* Find out the area under the curve by using auc function it founds upto 0.67 for test datasets.
* Similarly, the prediction for the training datasets and finally calculate the accuracy and compare it with the test accuracyif the accuracy is same the model is perfectly fit.

1. In this time and age of widespread internet usage, effective and targeted marketing plays a vital role. A marketing company would like to develop a strategy by analyzing their customer data. For this, data like age, location, time of activity, etc. has been collected to determine whether a user will click on an ad or not. Perform Logistic Regression on the given data to predict whether a user will click on an ad or not.

A screenshot of a cell phone

Description automatically generated

Ans: Business Objective:

To build a model which predict whether the customer click on the add or not

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Features** | **Description** | **Type** | **Relevance** |
| Daily Time\_ Spent \_on\_Site | Daily time spent by the customer on site | Quantitative, Ratio | Relevant |
| Age | Age of the customer | Quantitative, Ratio | Relevant |
| Area\_Income | Income of the customer | Quantitative, Ratio | Relevant |
| Daily Internet Usage | Daily internet usage of the customer | Quantitative, Ratio | Relevant |
| Ad\_Topic\_Line | Topic which customer watch | Qualitative, Nominal | Irrelevant |
| City | City of the customer | Qualitative, Nominal | Irrelevant |
| Male | Male=1 female = 0 | Quantitative, Nominal | Relevant |
| Country | Country of the customer | Qualitative, Nominal | Irrelevant |
| Timestamp | Timestamp at which customer is on site | Quantitative, Nominal | Relevant |
| Clicked\_on\_Ad | Click on ad yes =1 no = 0 | Qualitative, Nominal | Relevant |

Steps For the Logistic Regression:

* Import the required libraries. Pandas, numpy, seaborn, matplotlib, scipy.

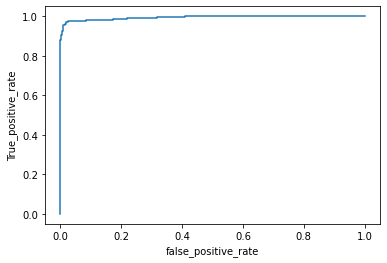
Seborn is the library for the advanced data visualization.

scipy is the library for the scientific calculation and for plotting the Q-Q plot.

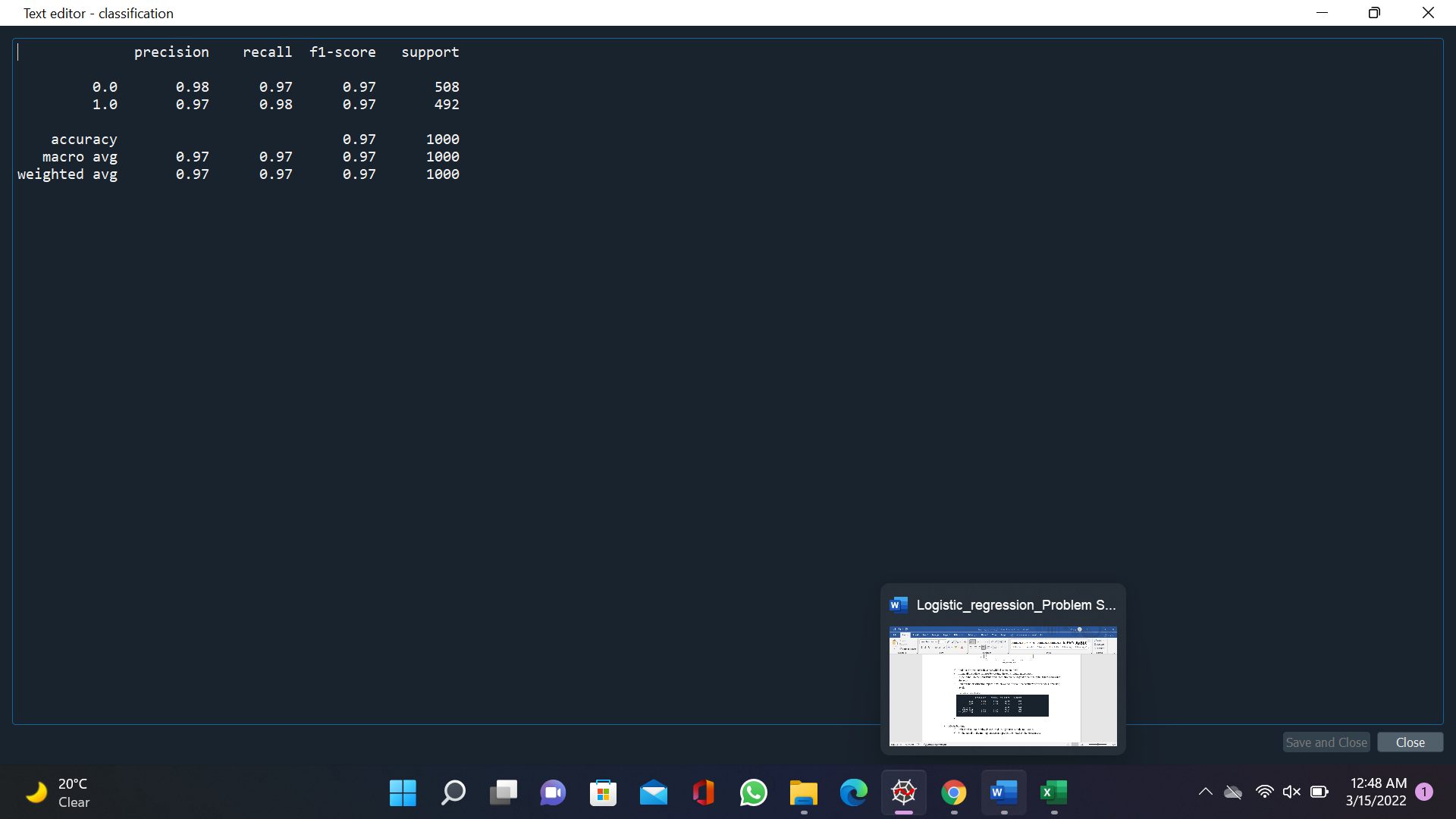
From statsmodel library import formulae package.

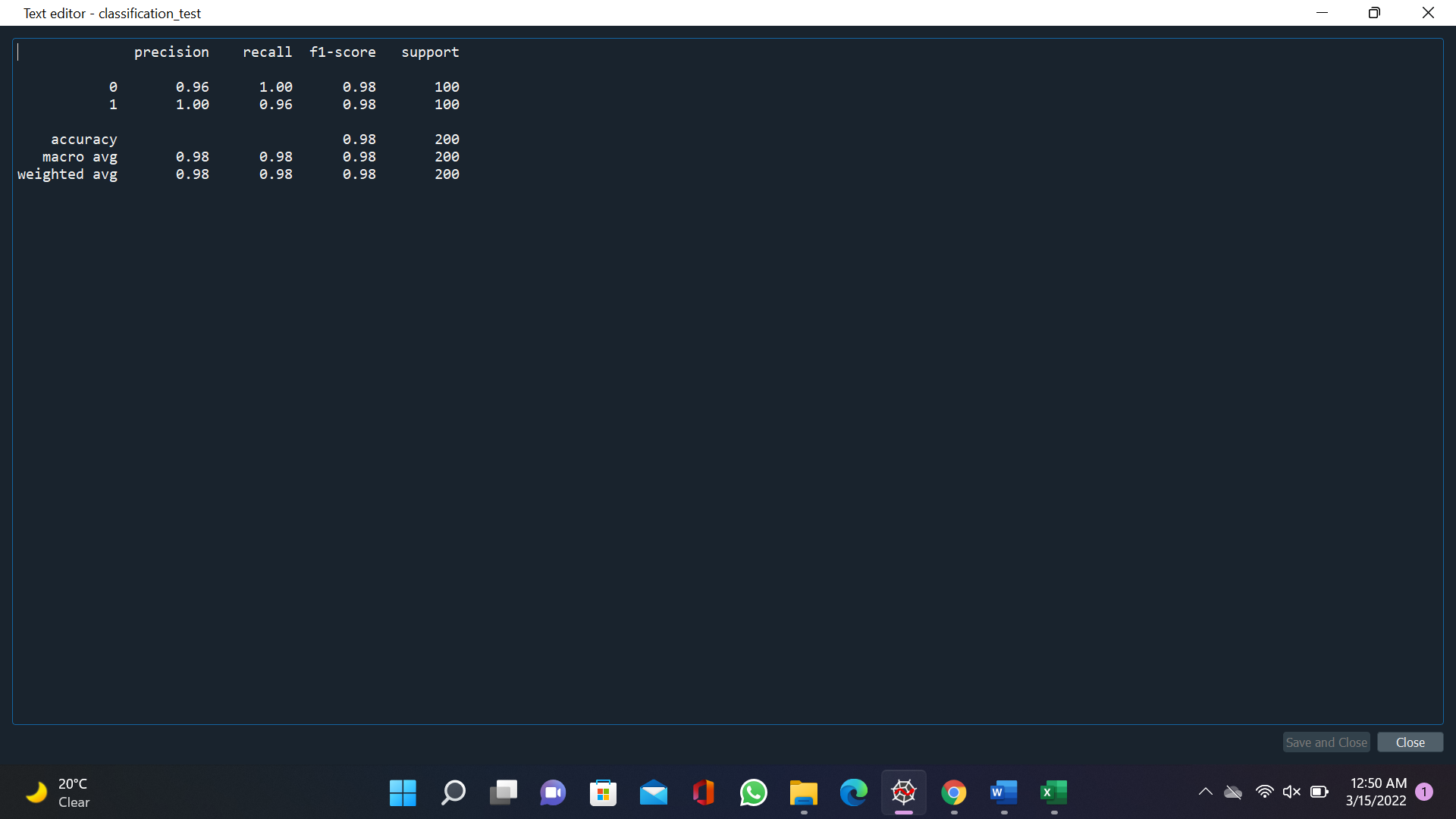
Statsmodel:- Statsmodels is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration. An extensive list of result statistics are available for each estimator.

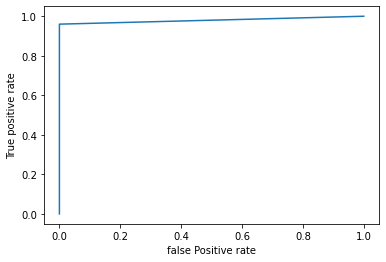
* Load the data
* Doing the univariate analysis and Exploratory data analysis.
* Checking the head i.e., top 5 rows of the datasets
* Checking the columns names of the datasets
* Checking the null values if any available in dataset.
* Checking the duplicate values in the datasets
* Checking the information i.e., datatypes of the datasets
* Exploratory data analysis. mean, median, mode, count, min, max etc.
* Check the distribution of the data.
* Dropping the unwanted column which is not useful for the analysis.
* Converting the nonnumerical data into numerical data by using one hot encoding or Label Encoder or pandas get\_dummies function as per the requirement
* Converting the continuous data into discrete form if necessary.
* Convert the timestamp data into the individual variables like month, day, hour, weekday by using to\_datetime function of pandas.
* Model Building
* Define the model by using the logit function and fir it into the data by taking naffairs as the dependent variable and rest of the column as the independent variables.
* Find the summary2 and summary of the model.
* Prediction to be done on the given datasets by using the independent variables.
* Calculate the false positive ration, True positive ratio, threshold value by using the roc\_curve function of the metrics package of sklearn and out of all the value find out the optimal threshold value and for the given datasets we get the threshold value = 0.497
* Plot the roc curve by taking false positive ratio on the x-axis and True positive value on the Y axis.



* Find out the area under the curve which is come out 0.99.
* Filling all the cells with zeros by creating the extra column named pred.
* Append the 1 to the pred value if the prediction value is greater than the optimal threshold value else zero.
* Find out the classification report in which we can find out the accuracy of the models, precision, recall.

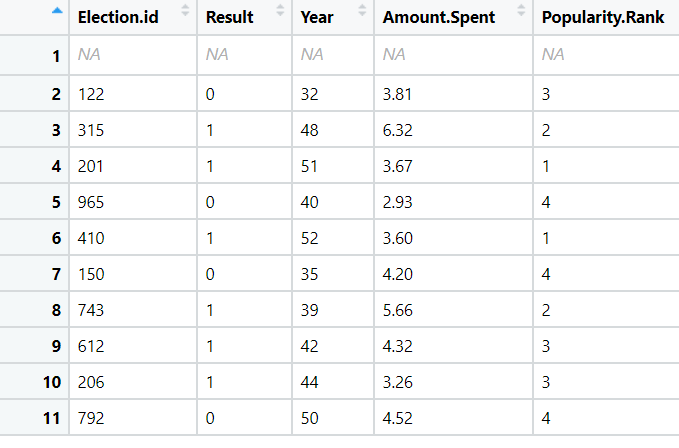


* Model building:
* Spilt the data into training datasets and testing datasets in the ratio 80:20.
* Fit the model on the training datasets and predict the model on the test datasets.
* Filling all the cells with zeros by creating the extra column named test\_pred.
* Append the 1 to the test\_pred value if the prediction value is greater than the optimal threshold value else zero.
* Find out the accuracy of the model manually by plotting the confusion matrix
* Find out the classification report in which we can find out the accuracy of the models, precision, recall.
* Calculate the false positive ration, True positive ratio, threshold value by using the roc\_curve function of the metrics package of sklearn and plot the roc\_curve for the test datasets.



* Find out the area under the curve by using auc function it founds upto 0.98 for test datasets.
* Similarly, the prediction for the training datasets and finally calculate the accuracy and compare it with the test accuracy if the accuracy is same the model is perfectly fit.

1. Perform Logistic Regression on the dataset to predict whether a candidate will win or lose the election based on factors like amount of money spent and popularity rank.



**Ans:- Business Objective:-**

To build a model to predict whether a candidate will win or lose the election based on factors like amount of money spent and popularity rank.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Type** | **Relevant** |
| Election-id | Id of the election | Quantitative, Nominal | Irrelevant |
| Result | Result win =1, loss = 0 | Quantitative, Nominal | Relevant |
| Year | No od years from which the party is established | Quantitative, Ratio | Irrelevant |
| Amount Spent | Amount spent for election | Quantitative, Ratio | Relevant |
| Popularity Rank | Rank of the popularity | Quantitative, Ratio | Relevant |

Steps For the Logistic Regression:

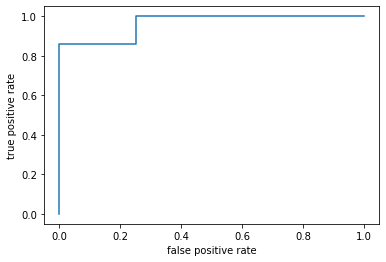
* Import the required libraries. Pandas, numpy, seaborn, matplotlib, scipy.

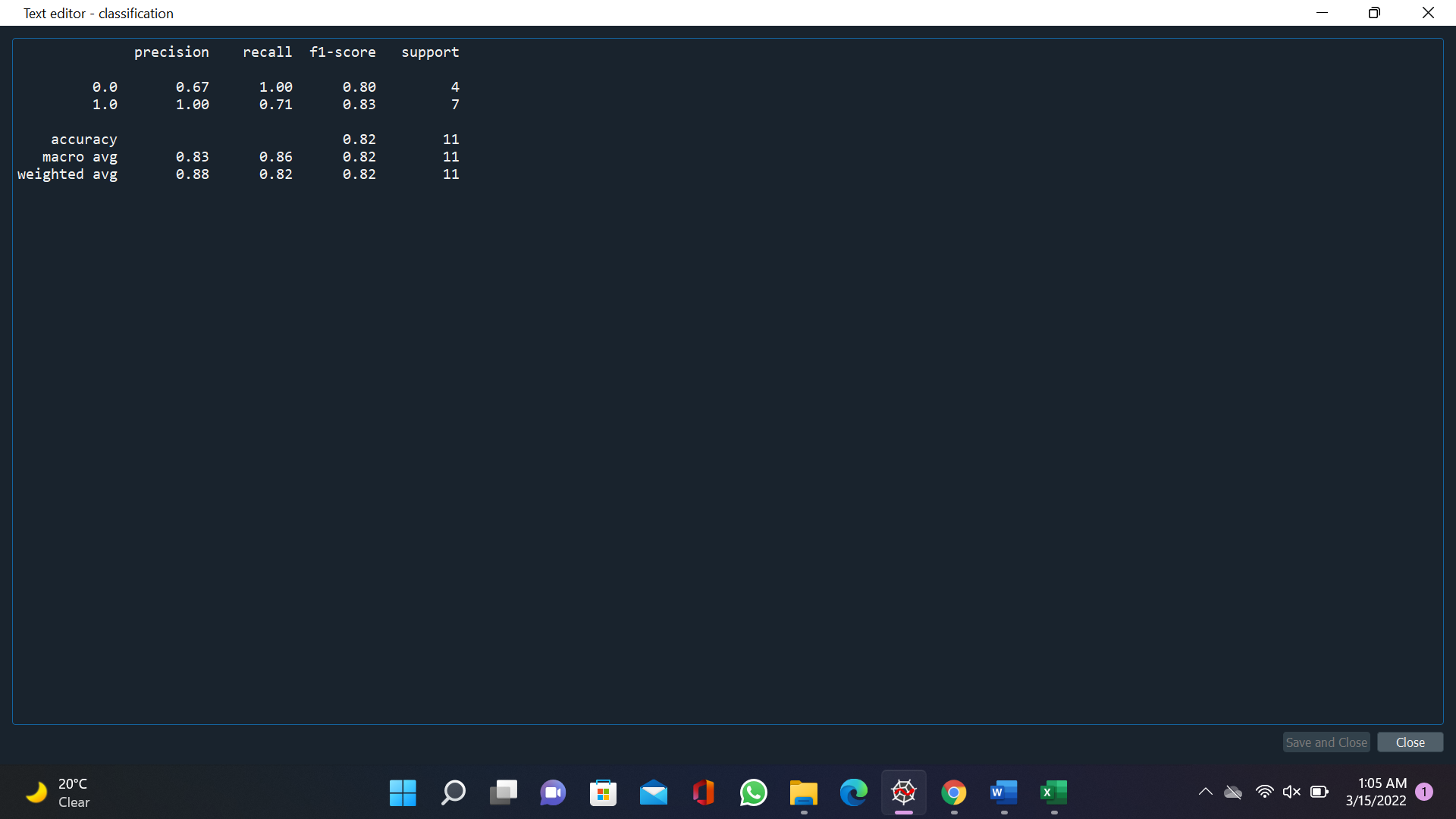
Seborn is the library for the advanced data visualization.

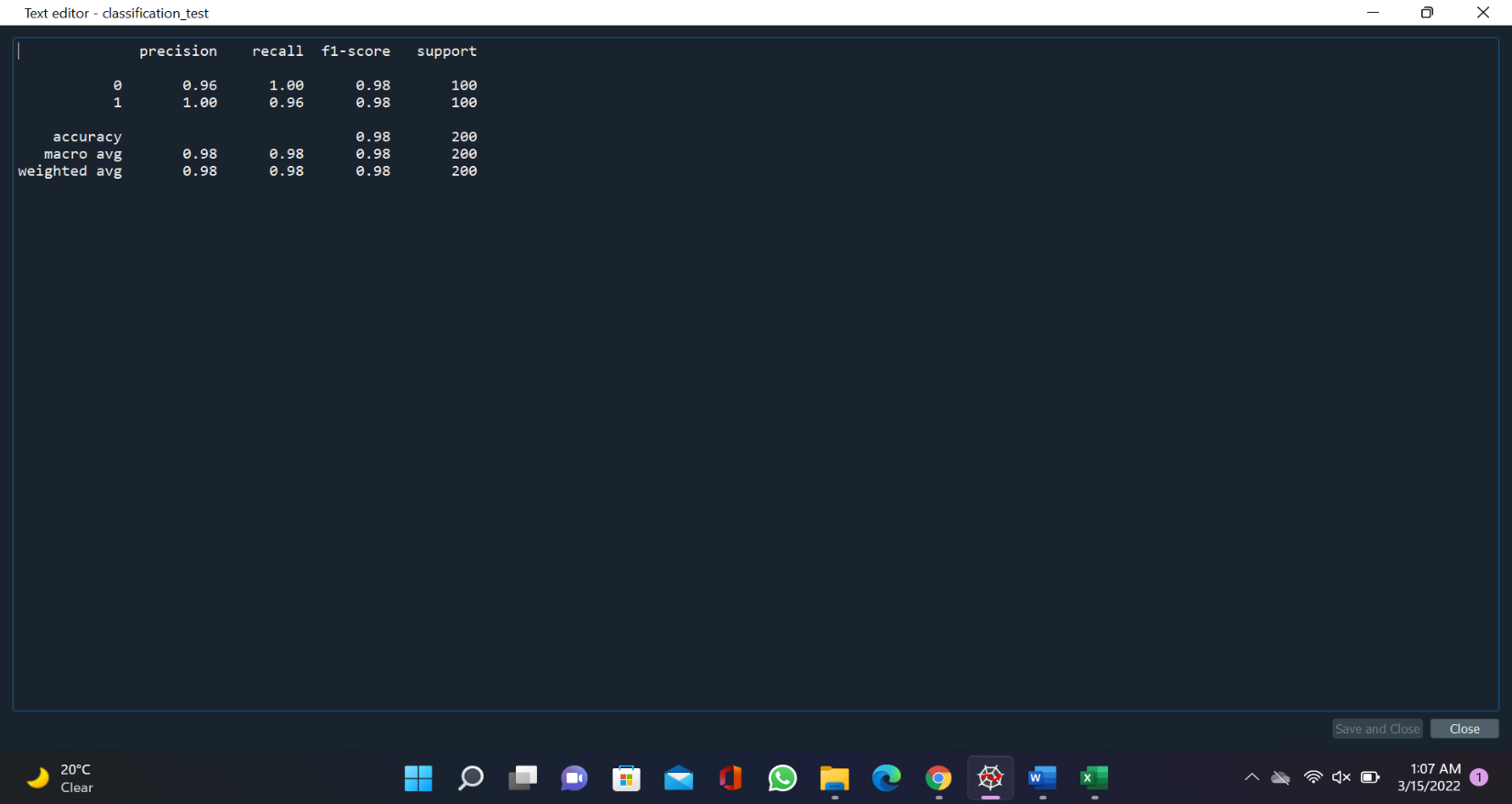
scipy is the library for the scientific calculation and for plotting the Q-Q plot.

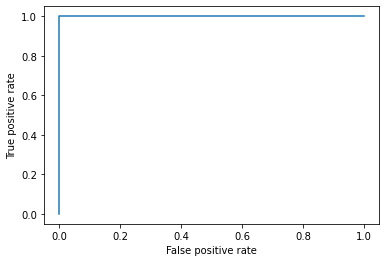
From statsmodel library import formulae package.

Statsmodel:- Statsmodels is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration. An extensive list of result statistics are available for each estimator.

* Load the data
* Doing the univariate analysis and Exploratory data analysis.
* Checking the head i.e., top 5 rows of the datasets
* Checking the columns names of the datasets
* Checking the null values if any available in dataset.
* Checking the duplicate values in the datasets
* Checking the information i.e., datatypes of the datasets
* Exploratory data analysis. mean, median, mode, count, min, max etc.
* Check the distribution of the data.
* Dropping the unwanted column which is not useful for the analysis.
* Converting the nonnumerical data into numerical data by using one hot encoding or Label Encoder or pandas get\_dummies function as per the requirement
* Converting the continuous data into discrete form if necessary.
* Convert the timestamp data into the individual variables like month, day, hour, weekday by using to\_datetime function of pandas.
* Model Building
* Define the model by using the logit function and fir it into the data by taking naffairs as the dependent variable and rest of the column as the independent variables.
* Find the summary2 and summary of the model.
* Prediction to be done on the given datasets by using the independent variables.
* Calculate the false positive ration, True positive ratio, threshold value by using the roc\_curve function of the metrics package of sklearn and out of all the value find out the optimal threshold value and for the given datasets we get the threshold value = 0.76
* Plot the roc curve by taking false positive ratio on the x-axis and True positive value on the Y axis.
* Find out the area under the curve which is come out 0.96.
* Filling all the cells with zeros by creating the extra column named pred.
* Append the 1 to the pred value if the prediction value is greater than the optimal threshold value else zero.
* Find out the classification report in which we can find out the accuracy of the models, precision, recall.



* Model building:
* Spilt the data into training datasets and testing datasets in the ratio 80:20.
* Fit the model on the training datasets and predict the model on the test datasets.
* Filling all the cells with zeros by creating the extra column named test\_pred.
* Append the 1 to the test\_pred value if the prediction value is greater than the optimal threshold value else zero.
* Find out the accuracy of the model manually by plotting the confusion matrix
* Find out the classification report in which we can find out the accuracy of the models, precision, recall.
* 
* Calculate the false positive ration, True positive ratio, threshold value by using the roc\_curve function of the metrics package of sklearn and plot the roc\_curve for the test datasets.



* Find out the area under the curve by using auc function it founds upto 0.98 for test datasets.
* Similarly, the prediction for the training datasets and finally calculate the accuracy and compare it with the test accuracy if the accuracy is same the model is perfectly fit.

1. It is vital for banks that customers put in long term fixed deposits as they use it to pay interest to customers and it is not viable to ask every customer if they will put in a long-term deposit or not. So, build a Logistic Regression model to predict whether a customer will put in a long-term fixed deposit or not based on the different variables given in the data. The output variable in the dataset is Y which is binary. Snapshot of the dataset is given below.

**A picture containing large

Description automatically generated**

**Ans: Business Objectives: -**

To build a model to predict whether a customer will put in a long-term fixed deposit or not based on the different variables given in the data.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Type** | **Relevance** |
| Age | age of customer | Quantitative, Ration | Relevant |
| Job | type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown') | Quantitative, Nominal | Relevant |
| Marital status | marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed) | Quantitative, Nominal | Relevant |
| education | : (categorical "unknown","secondary","primary","tertiary") | Quantitative, Ordinal | Relevant |
| Default | has credit in default? (binary: "yes","no") | Quantitative, Nominal | Relevant |
| Balance | average yearly balance, in euros (numeric) | Quantitative, Nominal | Relevant |
| Housing | has housing loan? (Binary: "yes","no") | Quantitative, Nominal | Relevant |
| Loan | has personal loan? (Binary: "yes","no") | Quantitative, Nominal | Relevant |
| Contact | ontact communication type (categorical: "unknown","telephone","cellular") | Quantitative, Nominal | Relevant |
| Day | last contact day of the month (numeric) | Quantitative, Nominal | Relevant |
| Month | ast contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec") | Quantitative, Nominal | Relevant |
| Duration | last contact duration, in seconds (numeric) | Quantitative, Nominal | Relevant |
| campaign | number of contacts performed during this campaign and for this client (numeric, includes last contact) | Quantitative, Nominal | Relevant |
| Pdays | number of days that passed by after the client was last contacted from a previous campaign (numeric, | Quantitative, Nominal | Relevant |
| Previous | number of contacts performed before this campaign and for this client (numeric) | Quantitative, Nominal | Relevant |
| poutcome | outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success') | Quantitative, Nominal | Relevant |
| y | has the client subscribed a term deposit? (binary: "yes","no") | Quantitative, Nominal | Relevant |

Steps For the Logistic Regression:

* Import the required libraries. Pandas, numpy, seaborn, matplotlib, scipy.

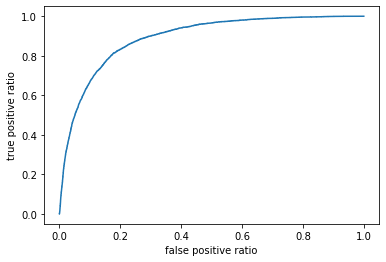
Seborn is the library for the advanced data visualization.

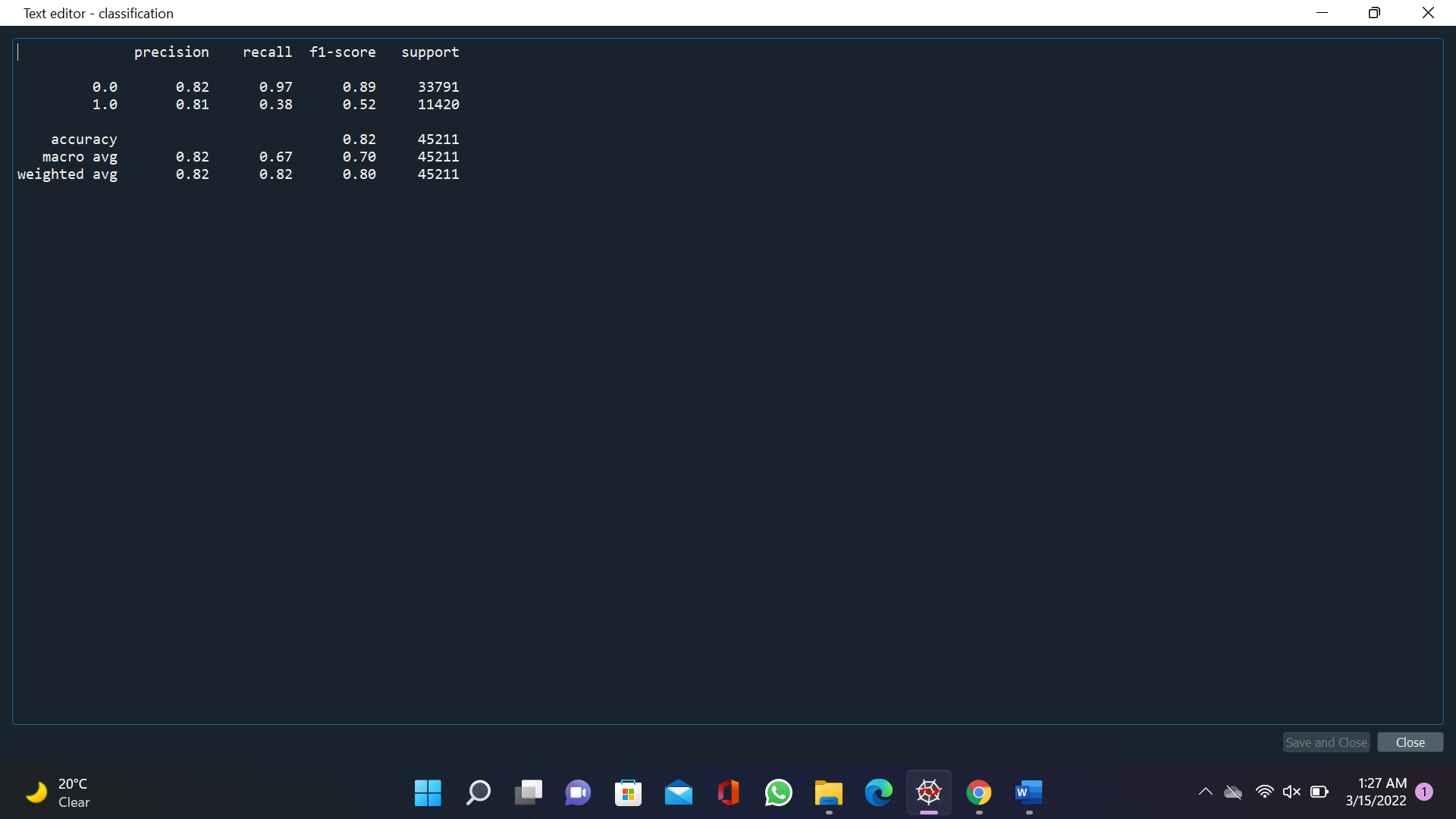
scipy is the library for the scientific calculation and for plotting the Q-Q plot.

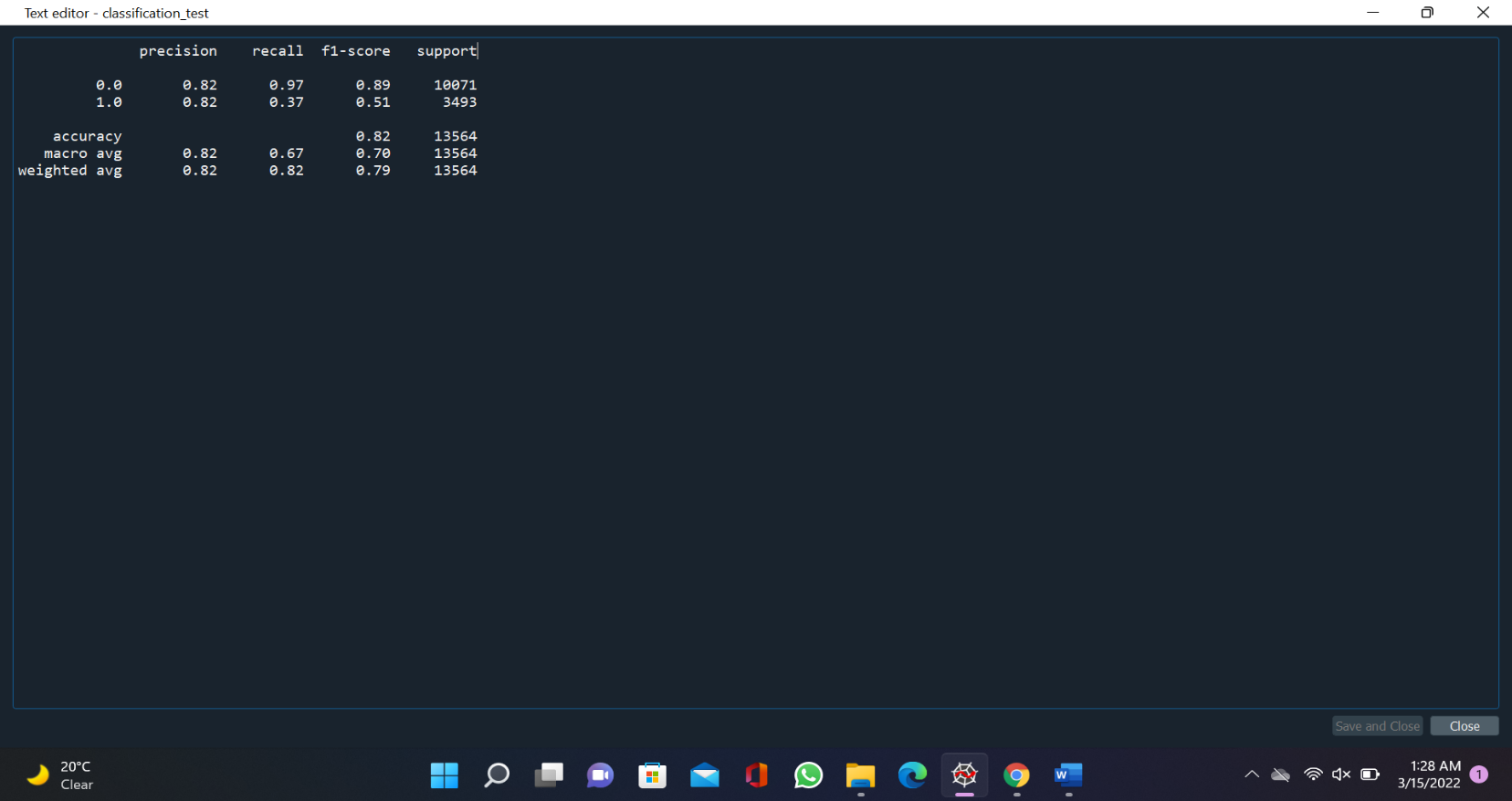
From statsmodel library import formulae package.

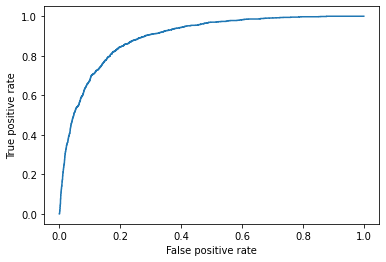
Statsmodel:- Statsmodels is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration. An extensive list of result statistics are available for each estimator.

* Load the data
* Doing the univariate analysis and Exploratory data analysis.
* Checking the head i.e., top 5 rows of the datasets
* Checking the columns names of the datasets
* Checking the null values if any available in dataset.
* Checking the duplicate values in the datasets
* Checking the information i.e., datatypes of the datasets
* Exploratory data analysis. mean, median, mode, count, min, max etc.
* Check the distribution of the data.
* Dropping the unwanted column which is not useful for the analysis.
* Converting the nonnumerical data into numerical data by using one hot encoding or Label Encoder or pandas get\_dummies function as per the requirement
* Converting the continuous data into discrete form if necessary.
* Convert the timestamp data into the individual variables like month, day, hour, weekday by using to\_datetime function of pandas.
* Model Building
* Define the model by using the logit function and fir it into the data by taking naffairs as the dependent variable and rest of the column as the independent variables.
* Find the summary2 and summary of the model.
* Prediction to be done on the given datasets by using the independent variables.
* Calculate the false positive ration, True positive ratio, threshold value by using the roc\_curve function of the metrics package of sklearn and out of all the value find out the optimal threshold value and for the given datasets we get the threshold value = 0.12
* Plot the roc curve by taking false positive ratio on the x-axis and True positive value on the Y axis.



* Find out the area under the curve which is come out 0.90.
* Filling all the cells with zeros by creating the extra column named pred.
* Append the 1 to the pred value if the prediction value is greater than the optimal threshold value else zero.
* Find out the classification report in which we can find out the accuracy of the models, precision, recall.
* Model building:
* Spilt the data into training datasets and testing datasets in the ratio 80:20.
* Fit the model on the training datasets and predict the model on the test datasets.
* Filling all the cells with zeros by creating the extra column named test\_pred.
* Append the 1 to the test\_pred value if the prediction value is greater than the optimal threshold value else zero.
* Find out the accuracy of the model manually by plotting the confusion matrix
* Find out the classification report in which we can find out the accuracy of the models, precision, recall.



*  Calculate the false positive ration, True positive ratio, threshold value by using the roc\_curve function of the metrics package of sklearn and plot the roc\_curve for the test datasets.
* Find out the area under the curve by using auc function it founds upto 0. for test datasets.
* Similarly, the prediction for the training datasets and finally calculate the accuracy and compare it with the test accuracy if the accuracy is same the model is perfectly fit.