Assignment 4

Logistic regression

Analysis report

Introduction to Artificial Intelligence

Gitansh Mittal

301200517

Analysis

* Four columns that needed to drop.

1. PassengerId: It is just the count for rows.
2. Cabin: as it has a lot of null values.
3. Name: as the everyone person has a different name and it can be converted to float.
4. Ticket: ticket number also not important.

* Analyze both plots and write a conclusion from each plot in your written  
  response.

A screenshot of a computer

Description automatically generated with medium confidence

The above graph is between the Passenger classes and survived. The graph representations are

1. Graph shows that the passenger with 1st class had a better survival chance.
2. The 1st  class passenger’s survival rate is more than 50 % as around 70 not-survived and around 140 survived.
3. The 2nd class passengers’ survival rate is less than 50% as around 100 not-survived and around 80 survived.
4. The 3rd class passengers survival is also less than 50% as more than 350 not-survived and around 130 survived.

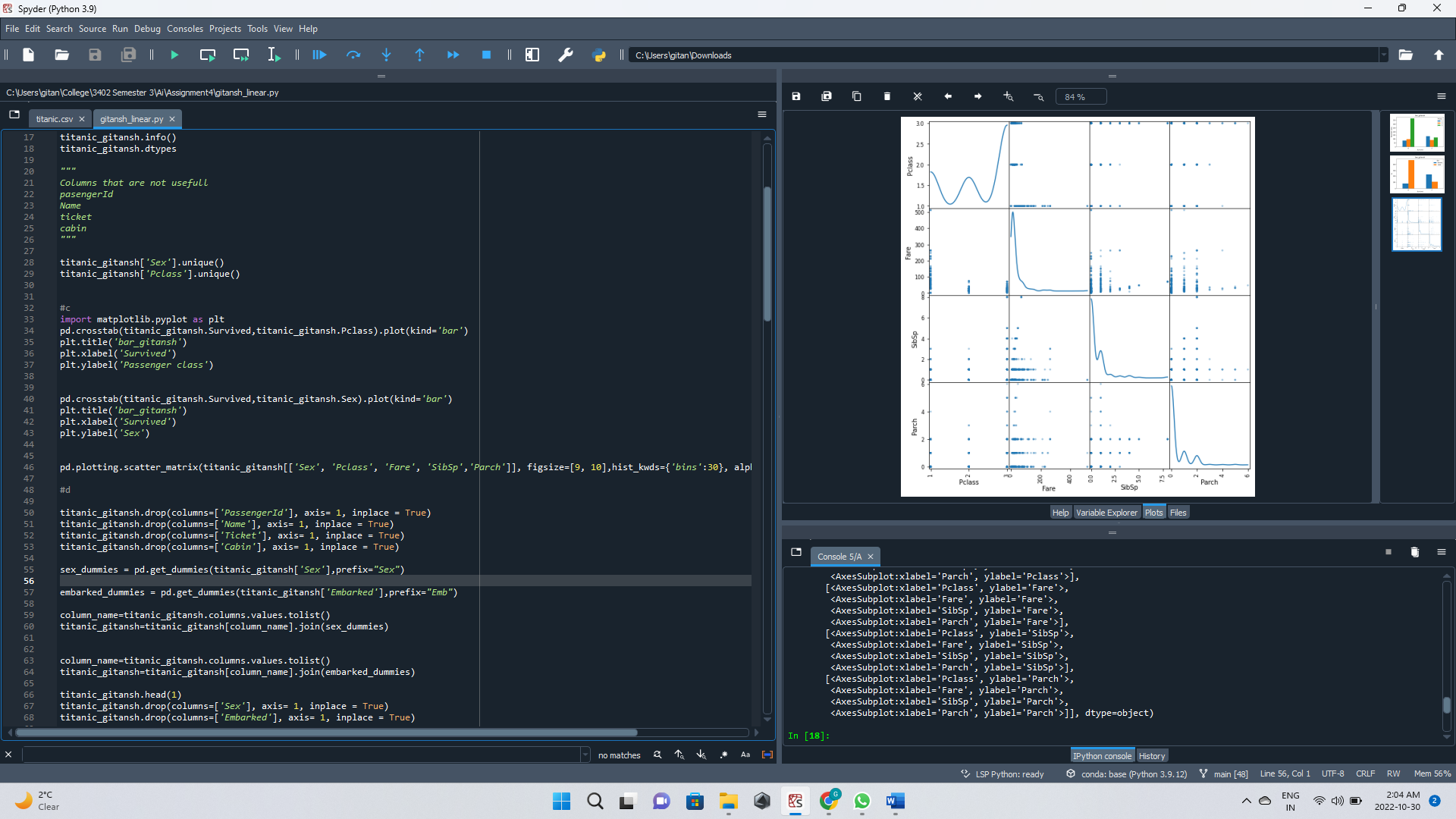
A screenshot of a computer

Description automatically generated with medium confidence

The above graph is between the Sex and survived. The graph representations are:

1. The survival rate is more of females than male.
2. Around 70 females not-survived and more than 200 survived.
3. More than 440 males not-survived and around 100 survived.

* Use pandas scatter matrix to plot the relationships between the number of  
  survived a the following features (attributes) : Gender, Passenger class, Fare,  
  Number of siblings/spouses aboard, Number of parents/children aboard. Analyze  
  the output and write some conclusions in your written response



A picture containing chart

Description automatically generated

Conclusion:

1. The data in each columns used in graph has a relation with another as no one box is empty.
2. Some shows the direct proportion and some are inverse proportion.

* Form the histogram generated focus on the “Port of Embarkation” and  
  write in your written response some highlights (Conclusions)

Graphical user interface

Description automatically generated

A picture containing text, crossword puzzle

Description automatically generated

This graph shows the data of each column respect to the number of passengers.

1. Cherbourg was Port of embarkment for around 180 passengers.
2. Southampton was Port of embarkment for more than 600 passengers.
3. Queenstown was Port of embarkment for less than 100 passengers.
4. The passengers that not-survived are around 600.
5. Around 210,180 and 570 person has 1st ,2nd and 3rd passenger class respectively.
6. The age is also described in number of passengers in that age group.
7. Almost same is with the data of columns in Sibsp, fare, Parch.
8. Next two graphs shows that around 580 were males and 320 were females.
9. Cherbourg was Port of embarkment for around 180 passengers.
10. Southampton was Port of embarkment for more than 600 passengers.
11. Queenstown was Port of embarkment for less than 100 passengers.

* Note these results in your writer report and recommended the best split  
  scenario.

According to me test size 0.15 means 15% and 85% is best split as the values of each fold in this split in a array has minimum range. As shown in the screen shot of list [loop\_scores]

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

Graphical user interface, text

Description automatically generated

Text

Description automatically generated

* Compare the accuracy on the test data with the accuracy generated using  
  the training data.

Train data results:

1. Accuracy score at threshold 0.5: 0.8186195826645265
2. Cross\_val\_score (mean): 0.8138248847926267

Test data results:

1. Accuracy score at threshold 0.5: 0.7723880597014925
2. Cross val score(mean): 0.7344729344729344

Text

Description automatically generated

Graphical user interface

Description automatically generated

Text

Description automatically generated

Graphical user interface, text

Description automatically generated

Text

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Text

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

* Compare the values of accuracy, precision and re-call generated at the  
  threshold 0.5 and 0.75 .

At threshold 0.5:

Accuracy score is 0.7723880597014925

Precision = TP / (TP+ FP) = 132/(132+36) = 0.7857142857=0.79

Recall = TP/ (TP+FN) = 132/(132+25) = 0.84.7673312=0.84

At threshold 0.75:

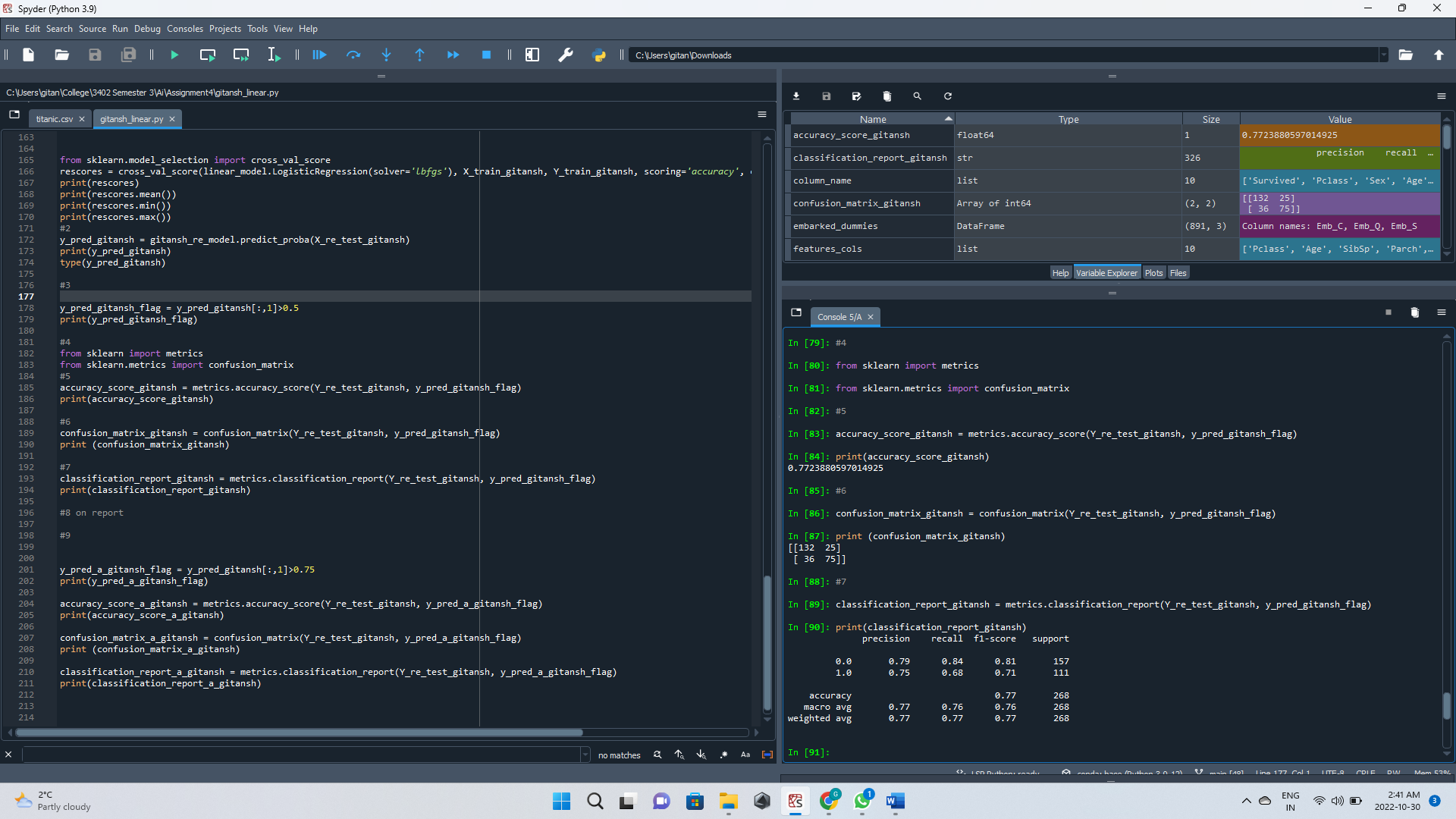
Accuracy score: 0.7425373134328358

Precision = TP / (TP+ FP) = 151/(151+63) = 0.7056074766=0.71

Recall = TP/ (TP+FN) = 151/(151+6) = 0.0.9617834395=0.96

Text

Description automatically generated



At threshold 0.75:

Text

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence