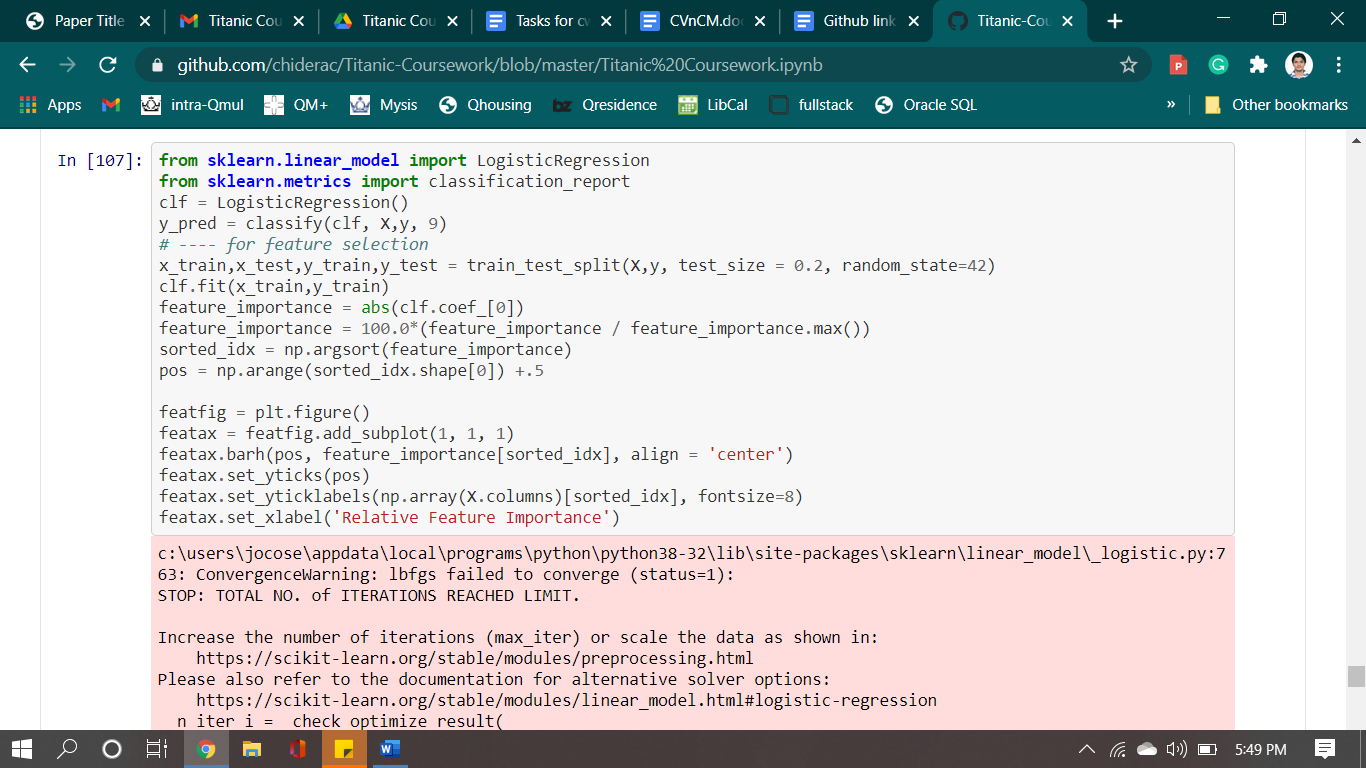
Testing and Analysis results

Finally, the number of features selected and cleaned for modelling are 8, which are: - Age, Fare, Pclass, Isalone, Sex, Embarked, Title and FamilySize. The model has been split using K-fold cross validation before the result of confusion matrix, accuracy and F-score can be analysed. The results of three classifer used are: -

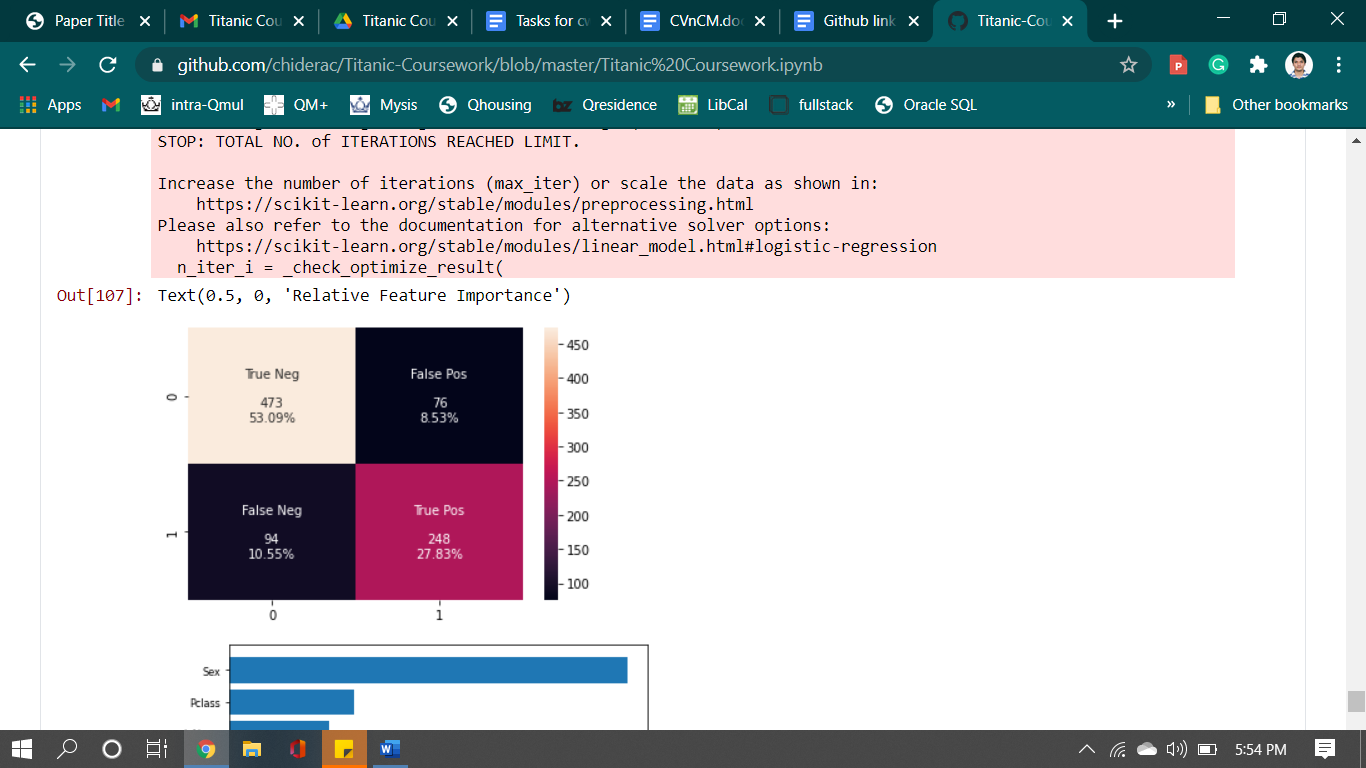
**Logistic Regression**

The first method used is Logistic Regression due to its straightforward and easy to understand approach. The classifier has been imported from sklearn library. The classifier is applied on the Kfold cross validated data multiple times using the predefined function classify (). The function prints the confusion matrix and gives comparison metrics scores like accuracy, precision, recall and f-score based on the comparison metrics drawn.



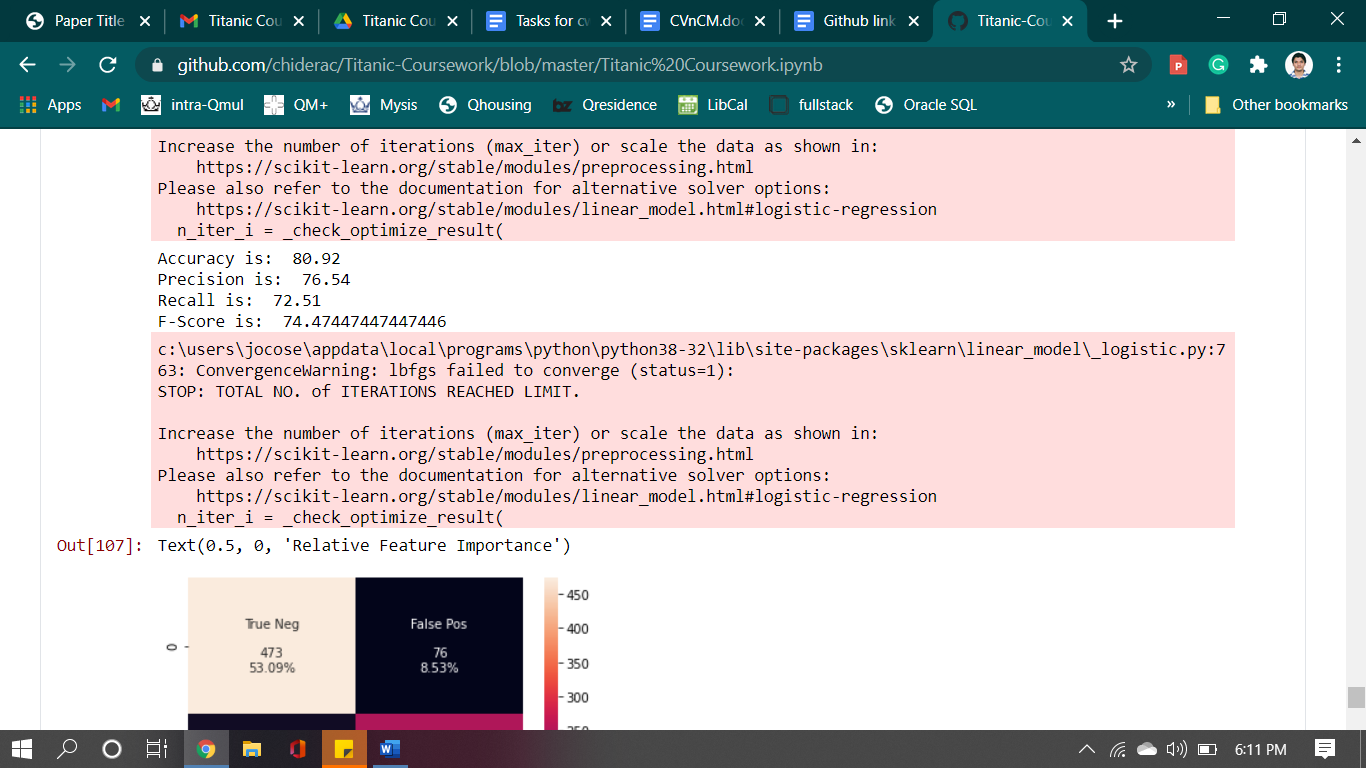
code snippet for training and testing data using the classify function with Logistic Regression.

The optimum folds of nine are used for cross validation by running the model several times and observing the comparison metrics through trial-and-error technique. The output received after the training and testing are as follows: -



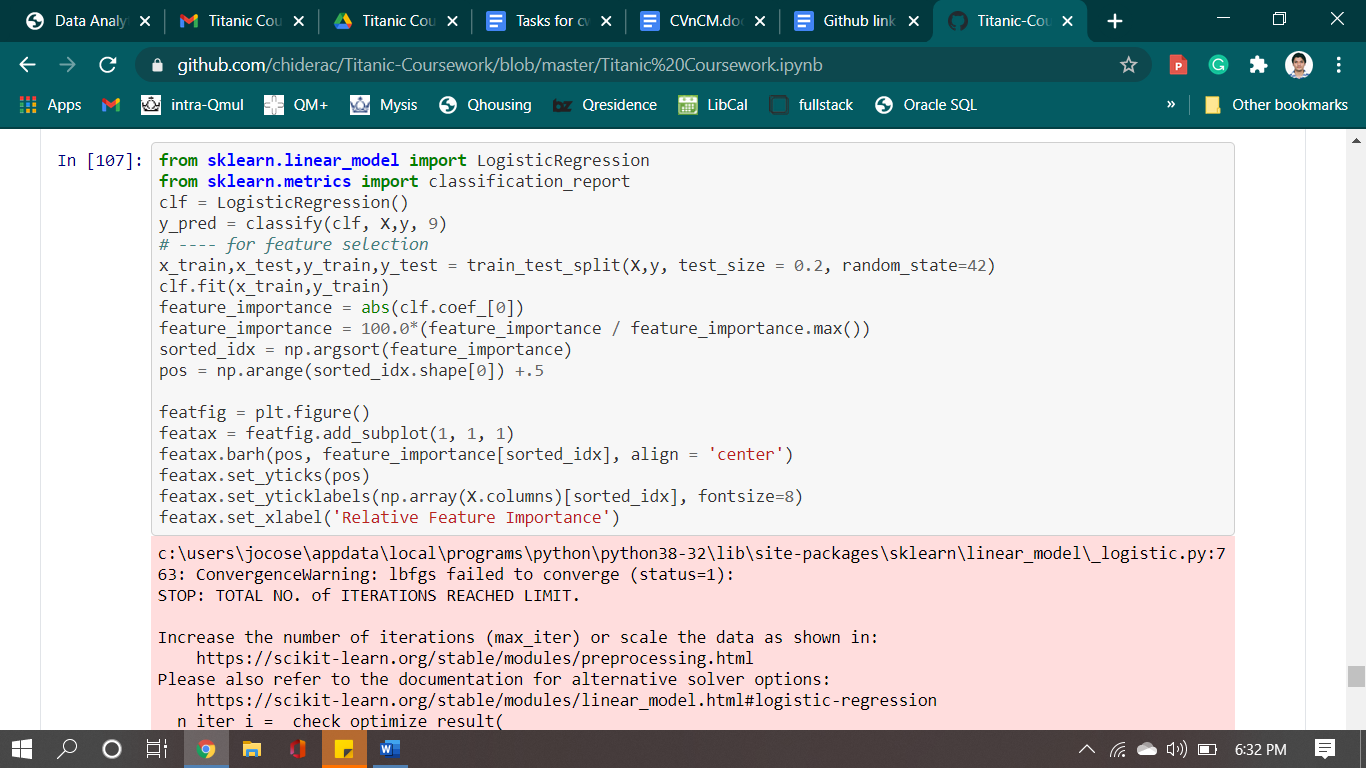
Confusion Matrix after Logistic Regression modelling

As it can be seen from the matrix that the number of true predictions is 473(53.09% of 891) in case of perished and 248(27.83% of 891) in case of survival. Further the number of false predictions is 76 in case of survived and 94 in case of perished. Based on the above matrix, comparison metrics are printed out



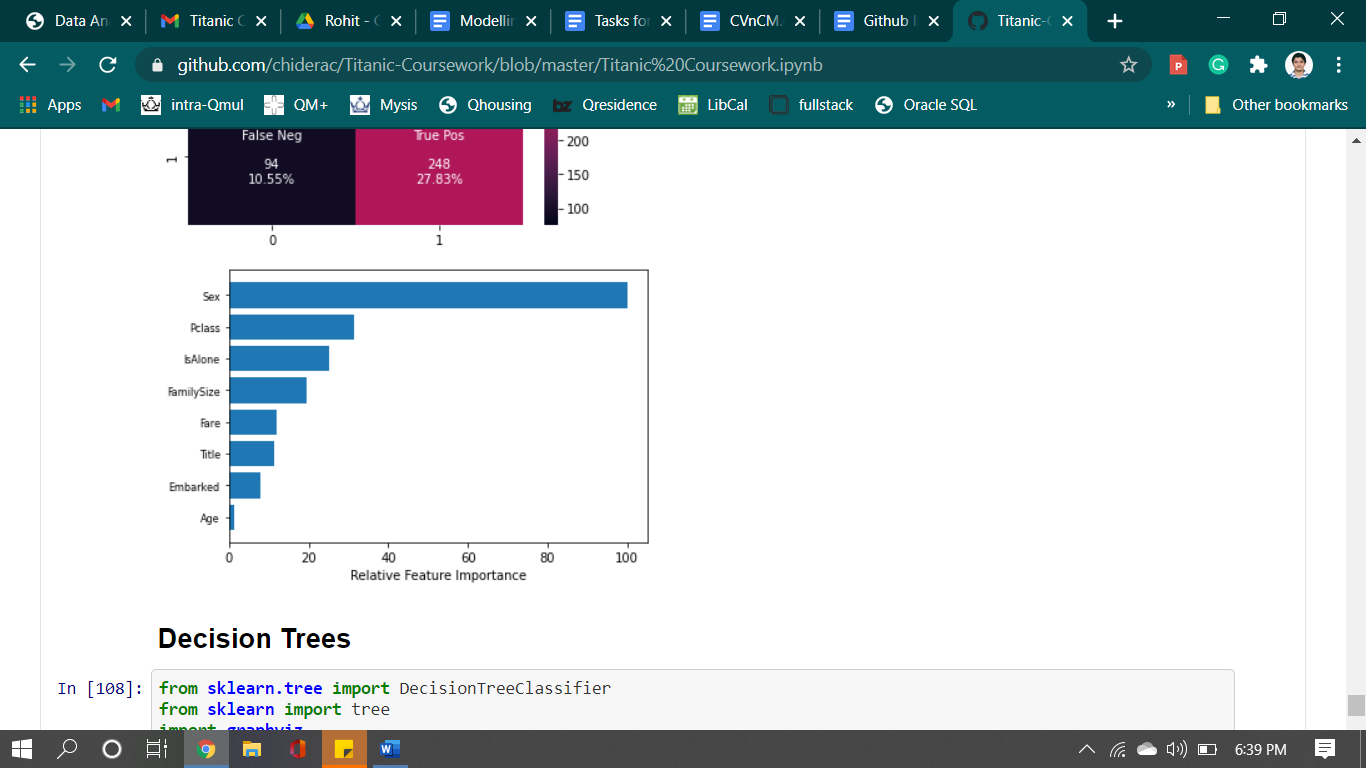
Overall Comparison metrics value after Logistic Regression modelling of the Titanic dataset

Further analysis of the model is done without using the cross validation to understand the real working of the model. The data is split in 80 to 20 ratios where 80% (712 X 8 input values and 712 X 1 output values) is training data and 20% (179 X 8 testing input values and 179 X 1 testing out values for comparing) is testing data. Then, the model is trained using the Logistic Regression.



Code snippet (in continuation with previous code) to draw the graph of relative feature importance.

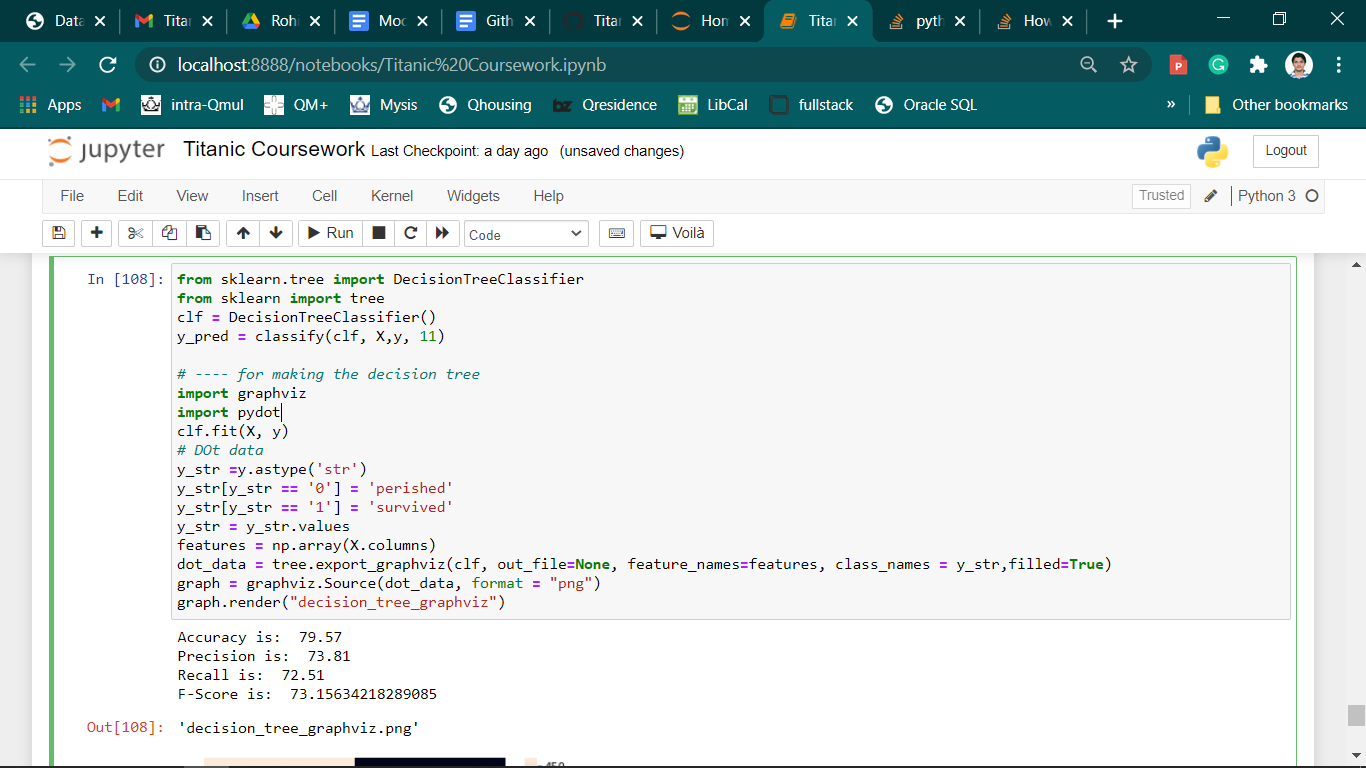
The output graph is :-



As it can be seen from the graph, the relative importance of sex is very high as compared to any of the other features. The age is least important feature for determining the survivability of a passenger as per the Logistic Regression modelling.

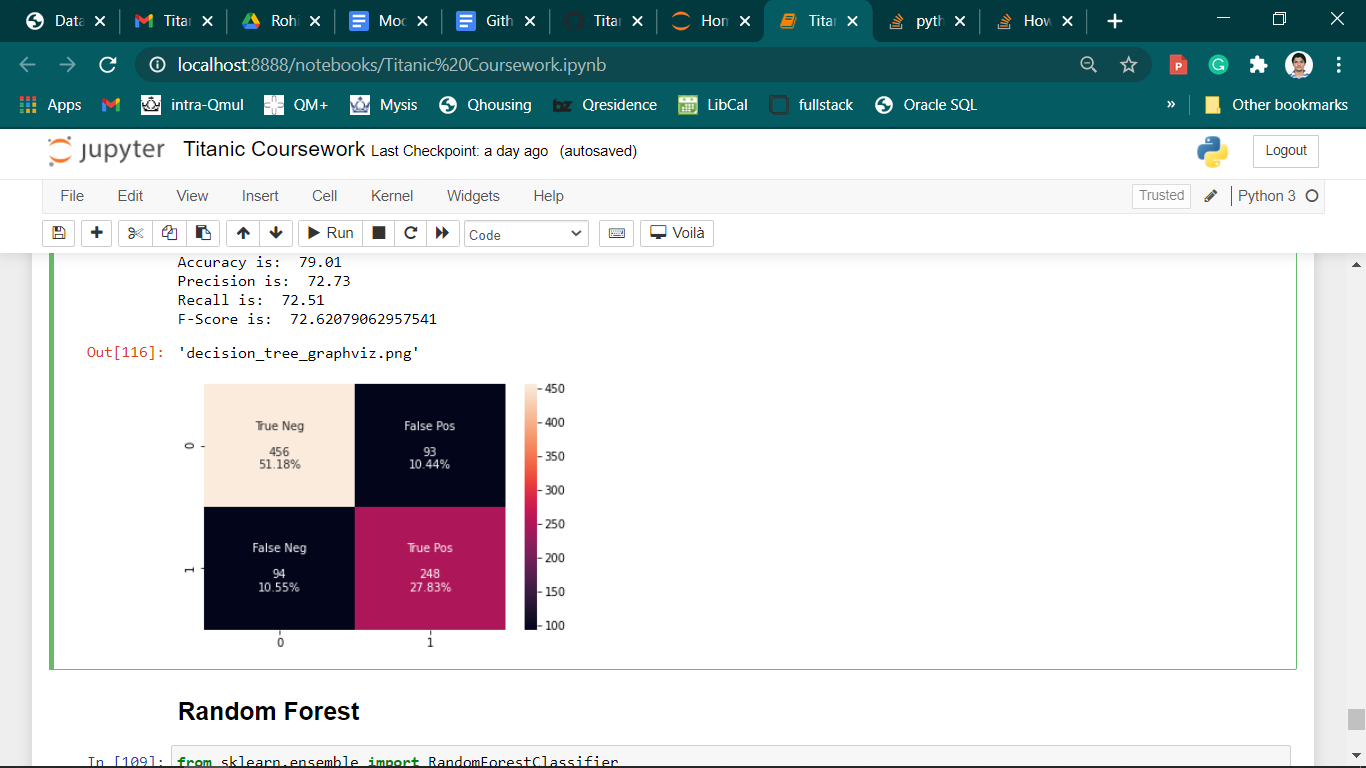
**Decision Tree**

As being stated above there could be a possibility of nonlinear relationship between the target value and any one of the predicting feature values. Therefore, to further improve the model, the data is trained and tested with Decision Tree algorithm in similar fashion as done in previous classifier. The data is split using cross validation with a kfold of 11 folds. This fold has been chosen via trial-and-error approach by observing theresults of comparison metrics.



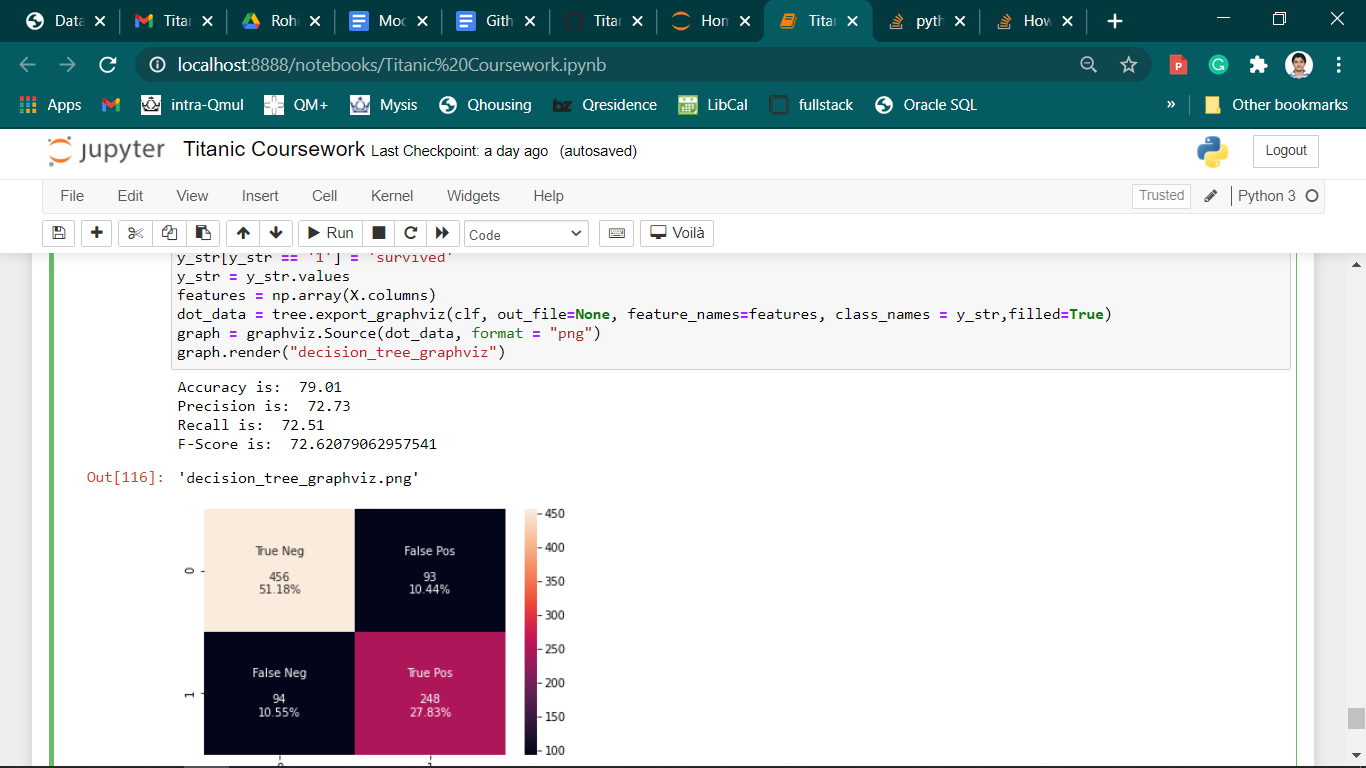
code snippet for training and testing data using the classify function with Decision Tree.

The confusion matrix obtained from the above testing is as follows: -



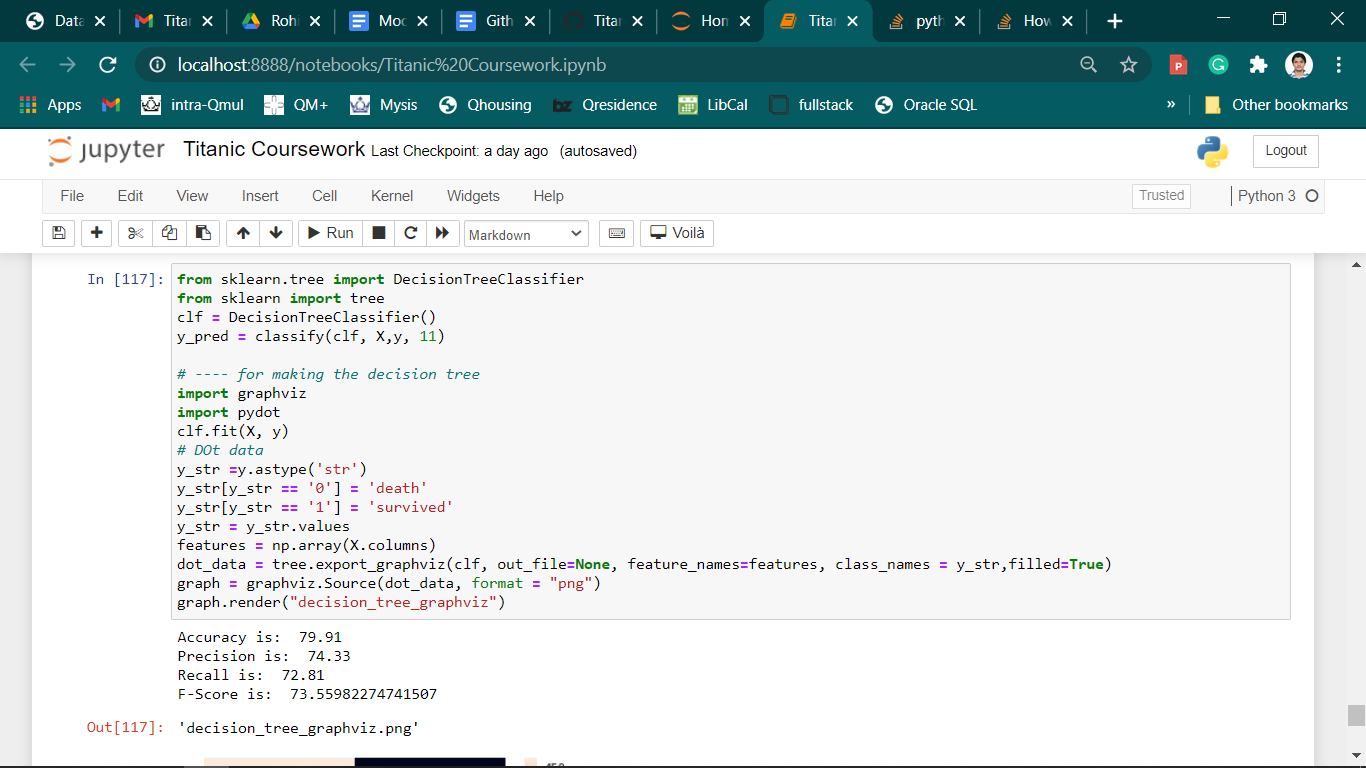
Confusion Matrix after Decision Tree modelling.

Here the number of predicted values for true negative (correctly predicted deaths) are 51.18 percentage of 891 data which is 456 and the number of true positive (correctly predicted survivals) are 248. The number of values which are predicted as deaths but were those who actually survived is 94 or 10.55 percentage of the whole dataset. From the above matrix, comparison metrics like accuracy, prediction, recall and F-Score were calculated as follows: -



Overall Comparison metrics value after Decision Tree modelling of the Titanic dataset

Decision tree is drawn get the real understanding of the model and the feature relevance. The code used for the making the decision tree is as follows: -



Code snippet (in continuation with previous code) to draw the Decision Tree.

The tree obtained after modelling is :-

Chart, timeline, scatter chart

Description automatically generated

Decision Tree obtained after training with Decision Tree using 80-20 data split.

The shades of red represent the cases where the predicted value is death and blue are the cases where the outcome is survived. The darker the shade, the more is the confidence or probability of that case.

**Random Forest**