# Classification Report:

A Classification report is used to measure the quality of predictions like how many predictions are True and how many are False. True Positives (true value is positive and prediction is also positive), False Positives (true value is negative and prediction is positive), True negatives (true value is negative and prediction is also negative) and False Negatives (true value is positive and prediction is also negative) are used for measurements. In report it give us precision, recall, and f1-score of both positives and negatives value.

### Precision:

It tells us the ratio of correct predictions from all predictions. E.g. to check precision of positives value we will divide value of true positives by sum of true positives and false positives. It will tell us ratio of how many positive value in predictions are correct from all positives values in prediction. Same for negative values we will divide true negative by sum of true negatives and false negatives.

### Recall:

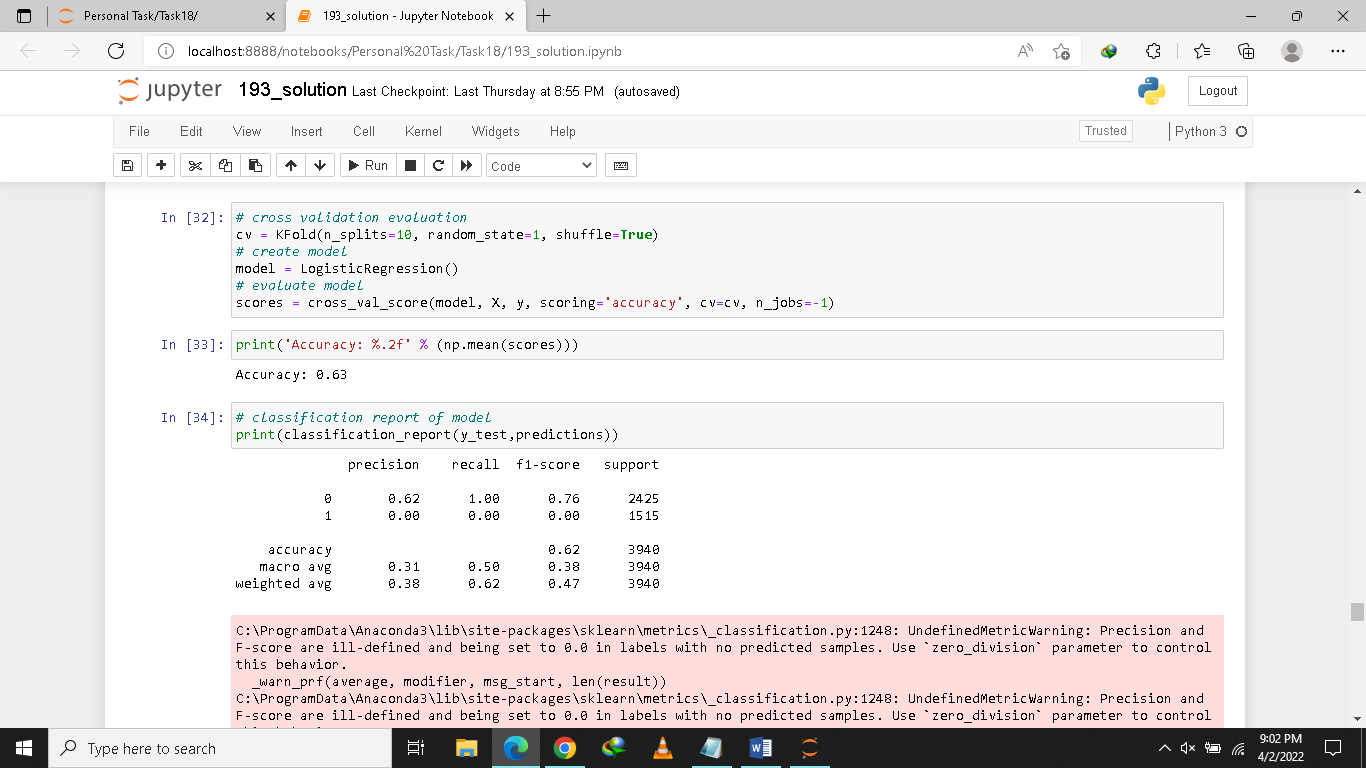
Recall is tell us the ratio of correct predictions from actual correct values. E.g. to check recall of positives value we will divide value of true positive by sum of true positives and false negatives. It will tell us ratio of how many positives values from actual positive values are predicted correctly. Same for negative values we will divide true negative by sum of true negatives and false positives.

### F1-Score:

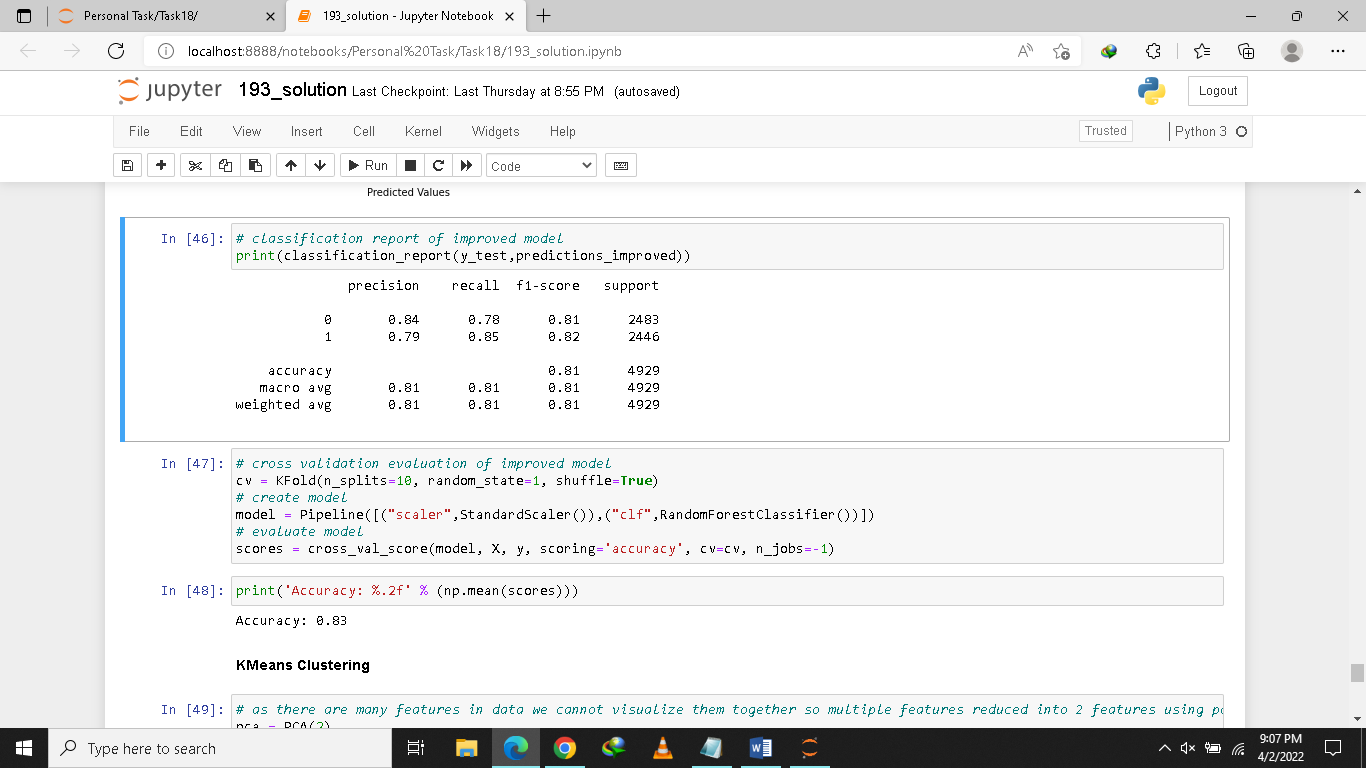
Sometimes by checking accuracy using simple accuracy formula (correct predictions divide by total predictions) gives us high accuracy even if model is not accurate. For example actual values are 0, 0, 0, 0, 1, 0, 0, 1, 0, 0 and model predicts 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 so when we check accuracy it will be 80% but in predictions model always predict 0 it will never predicts 1 so model is not accurate. The reason of this issue is class imbalance during training of model. Our project have same issue in logistic regression model as model is only predicting value 0 it never predicts 1. F1-score shows accuracy using recall and precision values. In same situation it will not give high accuracy. The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0.

F1 Score = 2\*(Recall \* Precision) / (Recall + Precision)

So basically, classification report helps in checking how much accurate model is.



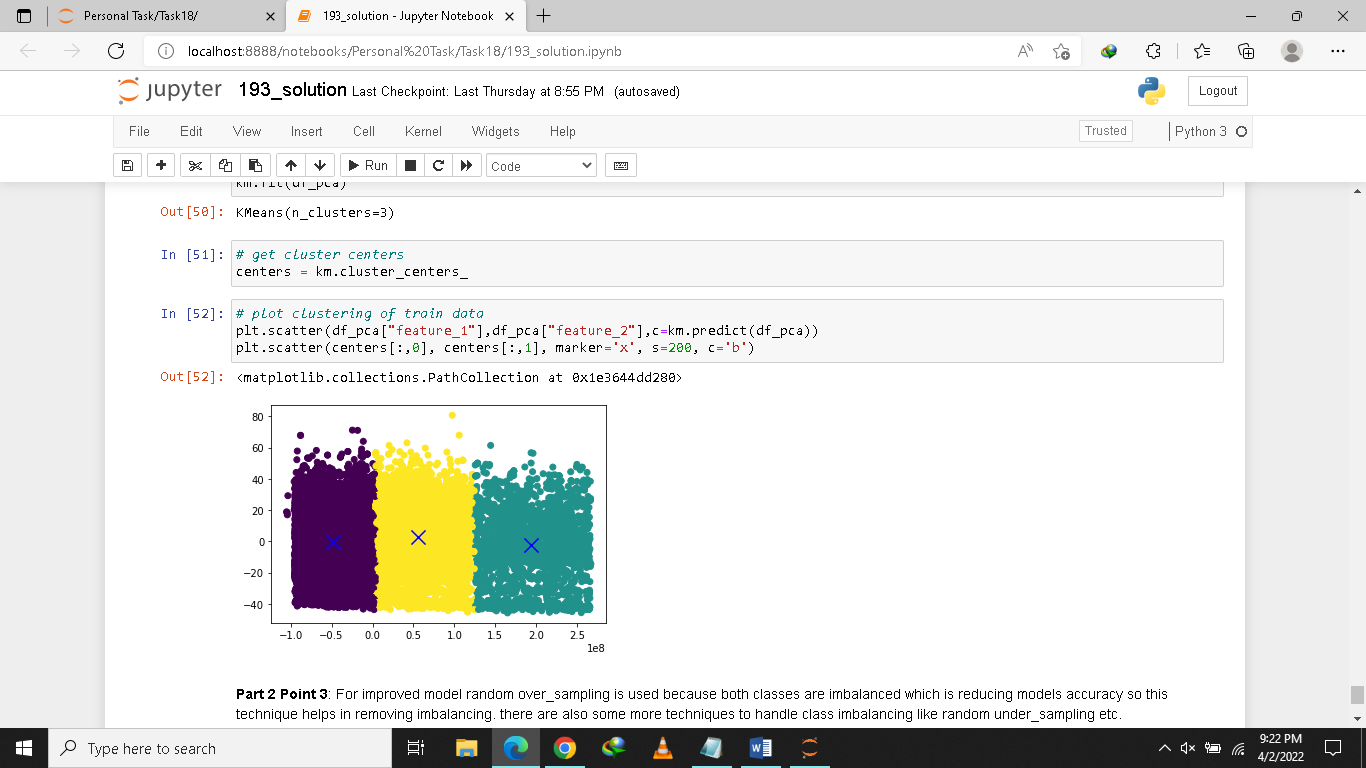
In our project accuracy of logistic regression model is 0.63. By help of classification report we can see precision and recall of 1 is 0.00 so its f1-score is also 0.00 means over model is not accurate as it only predicting 0.



After removing class imbalance both f1-scores are up to 80% means our model is 80% accurate which is much better than previous model.

# Clustering:

Clustering divides data points into number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters. In our project number of features in dataset are more than 2 as we cannot visualize more than 2 or 3 features at same time so we reduced number of features using PCA. PCA helps to reduce dimensions of data that still contain most of important information in data. It can reduce a little bit accuracy but also gives less dimensions data that are easier to explore and visualize and make analysis easier. So to visualize clustering plot we use PCA to visualize it easily.

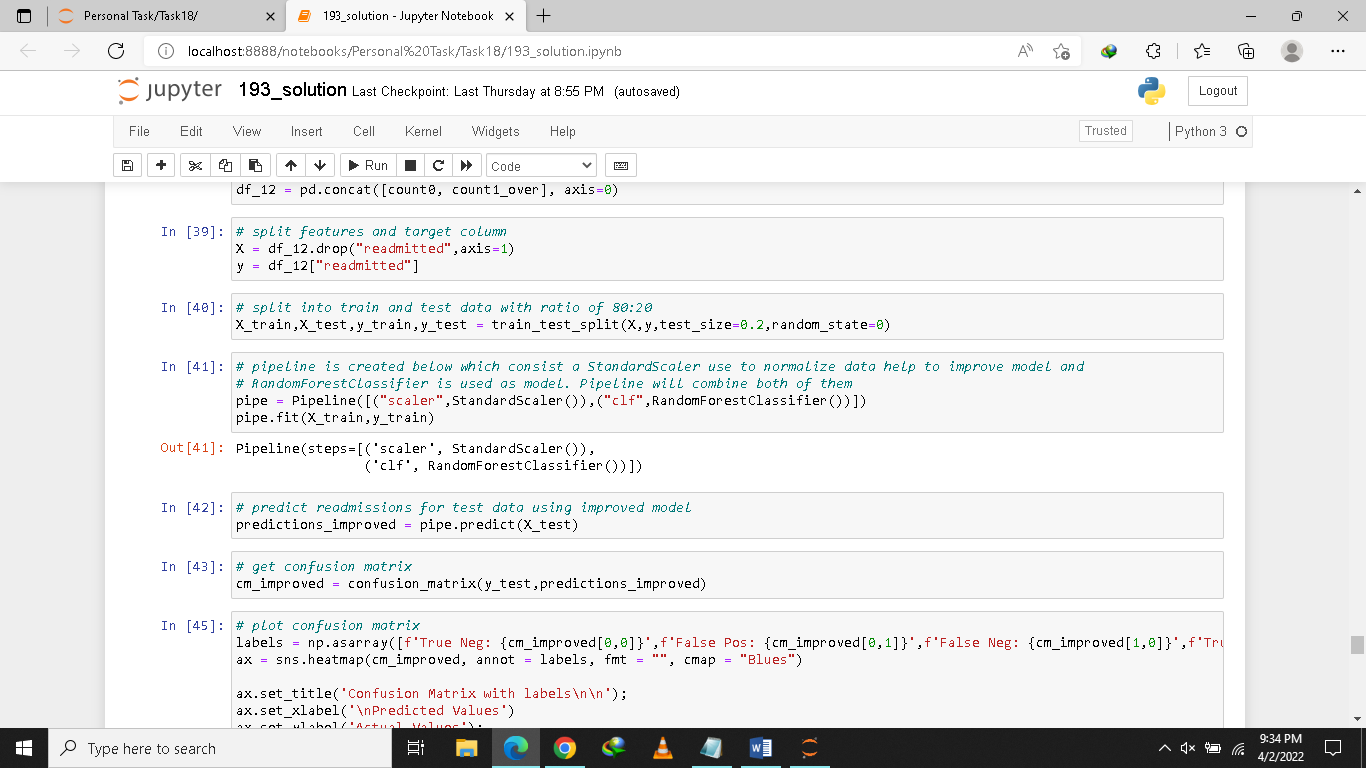


In plot above we visualize a scatter plot using 2 features reduced from data using PCA. And divided into 3 clusters or group. Points in each group have more similarity than points in other groups.

Another example of clustering is as we have data of faces of persons clustering can help us dividing males and females.

# Pipeline:

To get a better model we sometimes use many pre-processing techniques like normalization, transformation, removing or adding some features or missing values before training model. As we have to apply each technique individually on data before sending it for model training and creates many variables to store them. We can use pipeline instead which helps us to combine all steps we want to perform on data. Pipeline apply all steps on data in sequence and sends result of previous step to next step.



In our project we combined Standard Scaler which is use to normalize data with our classifier using pipeline instead of apply both steps individually on data. Pipeline will always normalize data first using Standard Scaler then use results of Standard Scaler to train classifier. When we predict value for test data first it will always normalize the test data then send result of Standard Scaler to classifier to predict result. In simple wording pipeline use to automate the whole pre-processing and model training steps.