**Report: Data Mining Assignment 3**

**Roll Number:20100004 Name: Dyass Khalid**

**Task:**

The main task of this assignment is to classify the evaluation data set using different models implemented in sklearn.

**Preprocessing:**

In this task first all the categorical variables are converted into encodings using the label encoder so that they can be fed to the defined models. No missing value is present in the data so further pre-processing is not required. The labels are separated from the data set then so that models are trained.

**Data Split:**

The training and testing data are split using sklearn train\_test\_split function with size of testing data to be 20% of the entire data with random state being kept as 2.

**Naïve Bayes Classifier:**

Gaussian Naïve bayes model is used for this task.

**Accuracy:**

Accuracy on training data is: 0.6316931982633864

Accuracy on testing data is: 0.615606936416185

**Confusion Matrix:**

[[5 0 38 29]

[6 0 5 7]

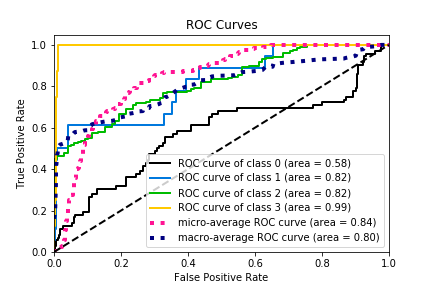
[5 0 200 43]

[0 0 0 8]]

**F1 Score:**

F1 Score:0.2741803418470264

**ROC Curve:**

The following ROC curve is obtained:

**Decision Tree:**

Decision tree is used for this task using the same data split for training and testing data.

**Accuracy:**

Accuracy on training data: 1.0

Accuracy on testing data: 0.976878612716763

**Confusion Matrix:**

[[ 67 0 5 0]

[ 0 18 0 0]

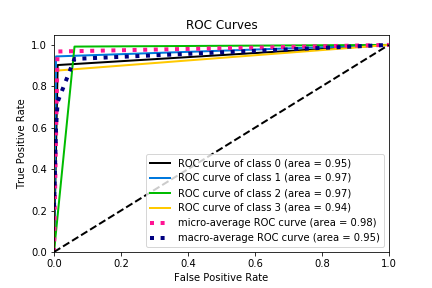
[ 2 0 246 0]

[ 0 1 0 7]]

**F1 Score:**

F1 Score:0.9606582150307901

**ROC Curve:**



**K means Clustering:**

K means clustering is an unsupervised learning algorithm in which labels are not provided. The model itself learn to classify the data. The number of clusters are provided while initializing the classifier. Here we take k = 4

**Accuracy:**

Accuracy on training data: -5143.5084397033015

Accuracy on testing data: -1293.9315055139346

**Confusion Matrix:**

[[23 15 13 21]

[ 3 5 8 2]

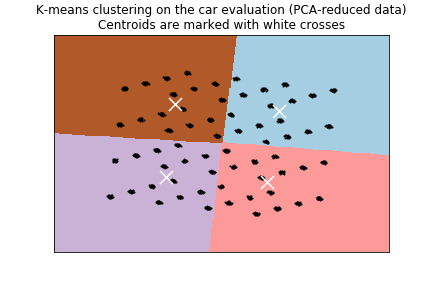
[70 60 48 70]

[ 0 1 3 4]]

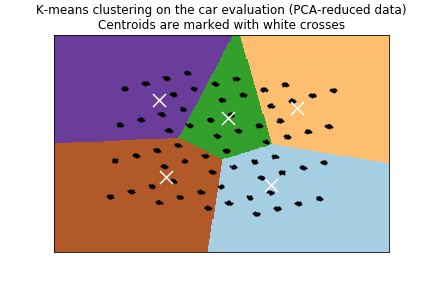
**F1 Score:**

F1 Score:0.18775252525252523

**Cluster Diagram for k =4:**



**Cluster Diagram for k =5:**



**Cluster Diagram for k =6:**

