Name: Peida Han

**Adult Dataset:**

**Data description:**

For the Adult Income Dataset, based on the previous data processing and analysis, I decide to keep eleven attributes including age, fnlwgt, occupation\_num, workclass\_num, education\_num, marital\_num, race\_num, sex\_num, capital-gain, capital-loss, hours\_per\_week. Missing data and unknown data are dropped. However, I keep outliers since the outliers are not noise and the percentage of outliers is significant. Therefore, it is not appropriate to remove all the outliers.

**Data transformation:**

To fit in the models, I transfer categorical attribute values into numerical attribute values. For occupation\_num, I mark high-skilled occupation as integer 1 and low-skilled occupation as integer 0. The reason to split this way is that the occupations in the category have no significant differences from each other but can be evaluated by skills. For workclass, I choose to change private related work class as 0 and change nonprivate class as 1. Because the distinction between each work class is private and nonprivate. It also makes data analysis easier. Furthermore, I separated marital status in to married as 0 and unmarried as 1. From the data processing and visualization, too detailed categories don’t help the data analysis since some category only have a small portion of data. Therefore, a better way is to generalize all martial categories into married as integer 0 and unmarried as integer 1. Since sex only has two types in this dataset, I make this categorical data into 0 or 1 format in order to distinguish the difference. I also separate race from 0 to 4 that represents different races. Because from earlier data processing and visualization, race can be a factor that influences occupation, and each race has fairly large numbers of data. Therefore, I keep each race instead of generalizing them into larger categories. Finally, class label income only has two categories. Therefore, I make <=50K as 0 and >50K as 1.

The irrelevant attributes are Relationship, Education, and Native-country. Since relationship has a high correlation with marital status. Therefore, I only keep the marital status to represent the marriage status of the population. Education has a high correlation with education\_num. Therefore, I only have to keep education\_num since education\_num is an ordinal form that can represent the degree of education from 1 to 13. The native country attribute is irrelevant to determine the income. Therefore, I discard the native country attribute.

**Evaluation Approach:**

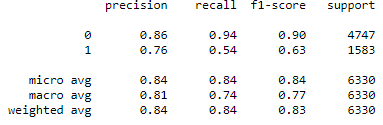
In this adult dataset, SVM, Logistic regression, Decision tree, and Naïve Bayes are the four models to analyze this dataset to find the best model.

By using the Single partition evaluation approach for the training set, I am able to divide the 70% of the total dataset into a training set and 30% of the total dataset into a test set. Furthermore, in the training set, I divide 30% of training data into a validation set and use the rest of the training set to train models. In this way, I can use my validation test set to compare the performance of my four models so that I can choose the best one to feed in the test set.

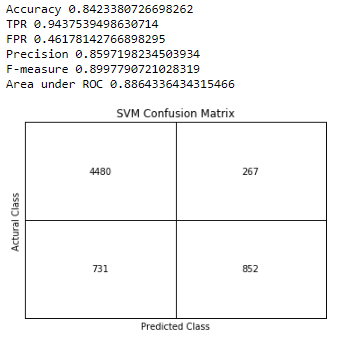
**SVM Classifier:**

After splitting the dataset into the training set, validation set, and test set explained above, I import SVM classifier from the sklearn package and fit the training data into SVM and use validation test set to determine the statistic information. For the SVM model, all the parameters are default except for the probability parameter. In order to calculate the ROC curve, I have to get the probability from the SVM model. Therefore, I set the probability parameter to true. The first graph below shows the classification report and the second graph contains the accuracy, True Positive, and False Positive rates, Precision rate, and F-measure of SVM model on Validation Dataset. The third graph shown below is the ROC of the SVM model for the Validation set.

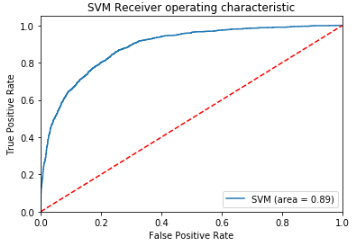
Classification Report:



Statistics Information and Confusion Matrix



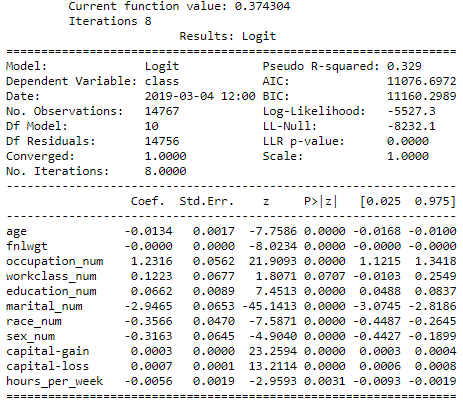
ROC Curve:



**Logistic Regression Classifier:**

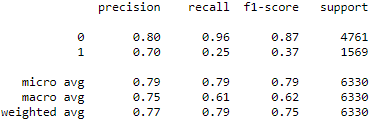
By using the same splitting approach as SVM, I feed my training set into the Logistic Regression model from the sklearn.linear\_model package and use validation test set to determine the statistic information. I also print out the summary with all the attributes information to estimate all attributes.

Summary Report:

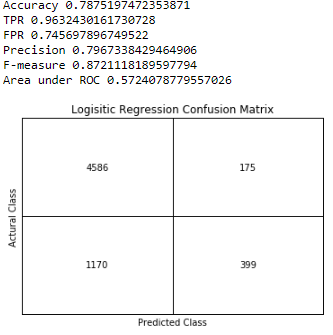


From the summary report, since there is no P value greater than 0.5, I can use all of the attributes to feed into the Logistic model. Because if the P value of an attribute is above 0.5, it could be no effect to the dataset. However, there is no P value of attributes is over 0.5. Then, I feed my training data into Logistic Regression model. The first graph shows the classification report and the second graph shown below contains the accuracy, True Positive, and False Positive rates, Precision rate, and F-measure of Logistic Regression model. The third graph shown below is the ROC of Logistic Regression model.

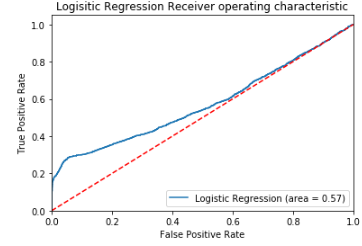
Classification Report:



Statistic Information and Confusion Matrix:



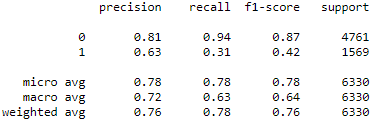
ROC Curve:



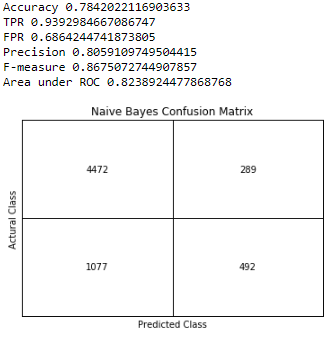
**Naïve Bayes Classifier:**

With the same splitting approach, I feed the training data into the Naïve Bayes model from sklearn.naive\_bayes package and use validation test set to determine the statistic information. The first graph below shows the classification report and the second graph contains the accuracy, True Positive, and False Positive rates, Precision rate, and F-measure of Naïve Bayes model. The third graph shown below is the ROC of the Naïve Bayes model.

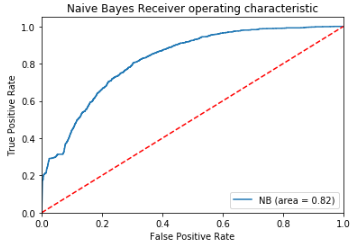
Classification Report:



Statistic Information and Confusion Matrix:



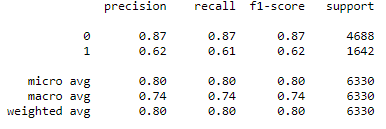
ROC Curve:



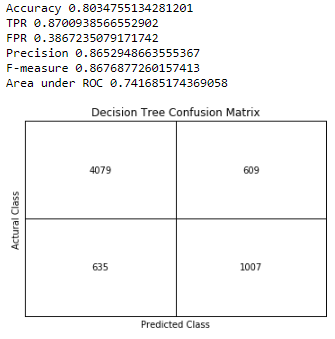
**Decision Tree Classifier:**

With the same splitting approach, I feed the training data into the Decision Tree model from the sklearn package and use the validation test set to determine the statistic information. The first graph below shows the classification report and the second graph contains the accuracy, True Positive, and False Positive rates, Precision rate, and F-measure of the Decision Tree model. The third graph shown below is the ROC of the Decision Tree model.

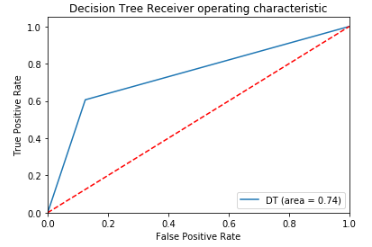
Classification Report:



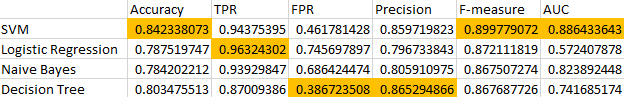
Statistic Information and Confusion Matrix:



ROC Curve:

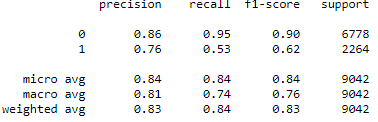


**The Best Model Evaluation:**

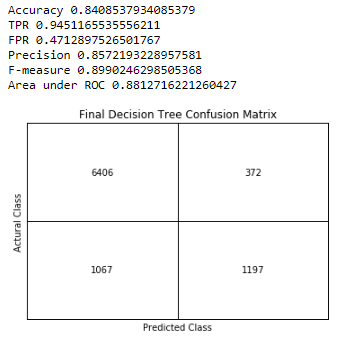


Based on the performance of the four model, accuracy, TPR, FPR, Precision, F-Measure, and the area under ROC of the model can be the factors to determine the best model fit in the dataset. In order to have the best performance, compared with the accuracy, TPR, FPR, Precision, F-Measure, and AUC, the best model is SVM among the four models. The SVM Model has the highest Accuracy, Highest F-measure, and Largest AUC. The Accuracy of the SVM model 84.2% can have the best prediction the class. Furthermore, the Highest Precision 90.0% can also give me the best possible correct prediction. The highest AUC 88.6% also shows that this model has better performance among the four models. Even though the True Positive Rate 94.4% and Precision 86% are not the highest, the SVM model still has a relatively better performance among the four models based on the best performance on Accuracy, F-Measure, and AUC. Then I use the SVM model as the best model. Then I feed the test set into SVM to check the performance on the test set. The performance of the SVM model with the test set has 84.1% accuracy with 88.1% AUC. The first graph shows the classification report and the second graph below contains the accuracy, True Positive, and False Positive rates, Precision rate, and F-measure of Naïve Bayes model. The third graph shown below is the ROC of the SVM model for the test set.

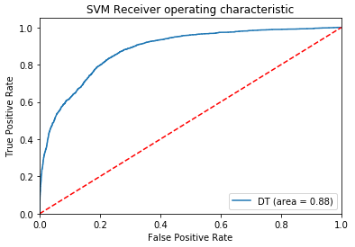
Classification Report:



Statistic Information and Confusion Matrix:



ROC Curve:



**Wine dataset:**

**Data description:**

The Wine Dataset doesn’t contain any categorical data and missing data. Therefore, there is no need to do any data transformation or data cleaning process. The only data transformation is to make class label into integer 0 and integer 1.

**Evaluation Approach:**

In this Wine dataset, I will try to find the best performance from four models including KNN, Logistic Regression, SVM, and Decision Tree.

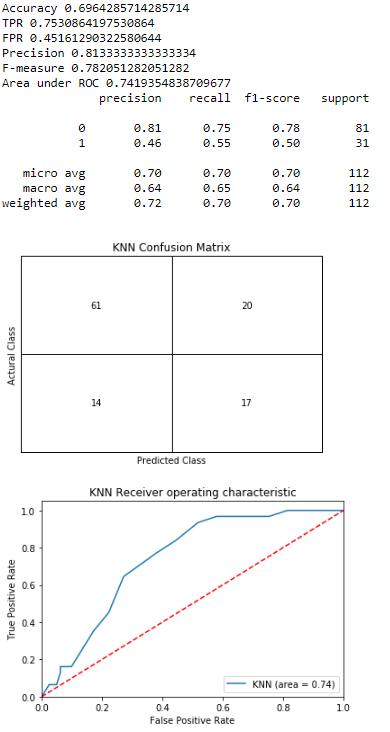
Since this dataset is relatively small, I implement the 10-Fold cross-validation approach to have a better performance. The reason to choose 10-Fold instead of 5-Fold is that I perform both cross-validation approach, 10-Fold Cross Validation gives me a better statistic information such as accuracy, AUC, TPR, and FPR. 10-Fold Cross Validation divides the training dataset into 10 pieces with 9 pieces of the training set and 1 piece of the validation set in each iteration. Therefore, after ten iterations, I can have better performance by averaging the results. Therefore, I have to iterate the model ten times and then average the result to get the final performance of KNN model.

**KNN Classifier:**

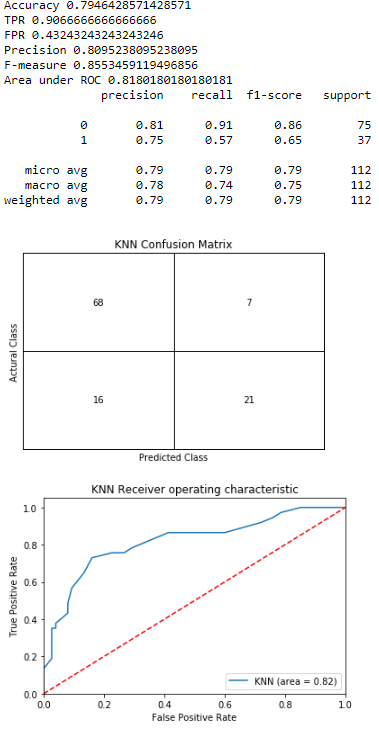
I use KNN classifier for this dataset. The choice for the K parameter in the KNN model is based on the result of the square root of the total number of the data, which is the usual way to determine the K value. And before feeding the data into the model, I need to standardize the whole dataset so that every attribute value is in the same range and avoid the situation that one attribute has extremely high value and affect the weight of other attributes.

Here are the 10 iterations statistic information and ROC.

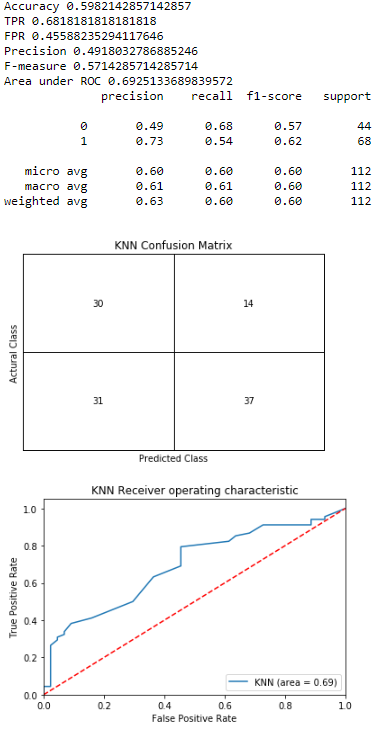
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the KNN model for the First Iteration:



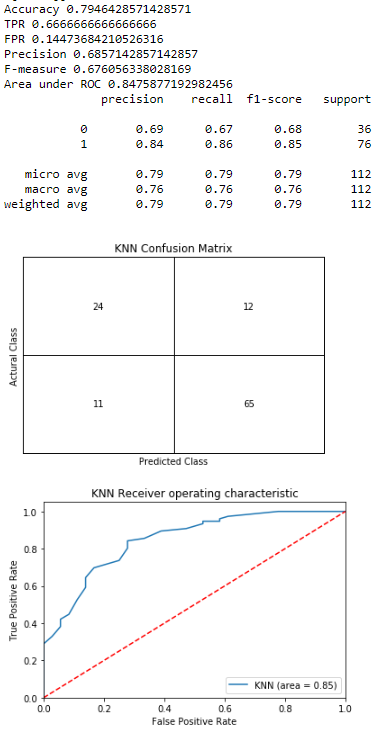
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the KNN model for the Second Iteration:



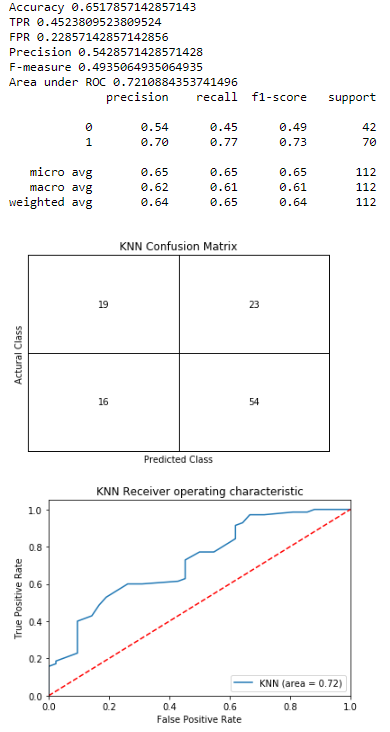
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the KNN model for the Third Iteration:



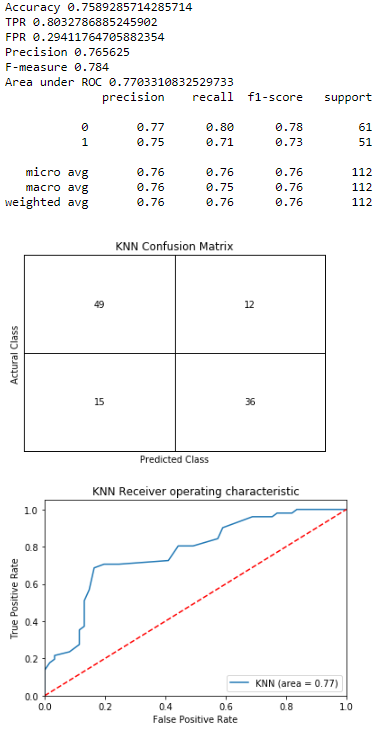
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the KNN model for the Forth Iteration:



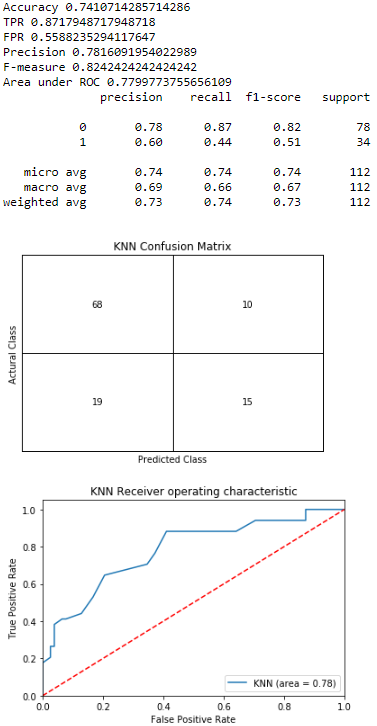
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the KNN model for the Fifth Iteration:



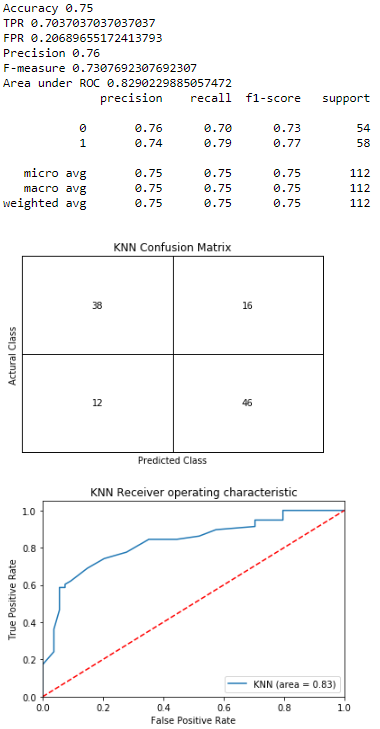
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the KNN model for the Sixth Iteration:



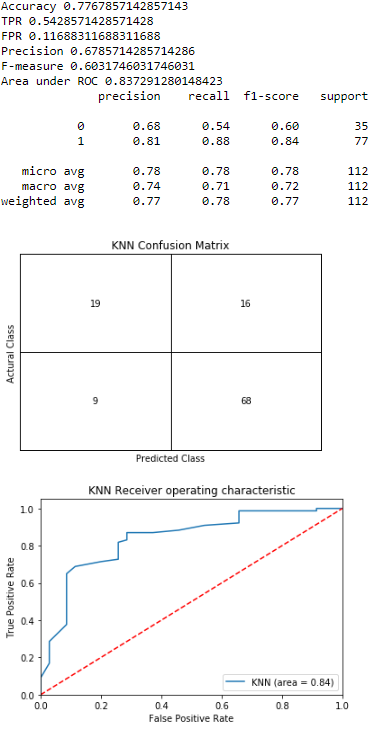
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the KNN model for the Seventh Iteration:



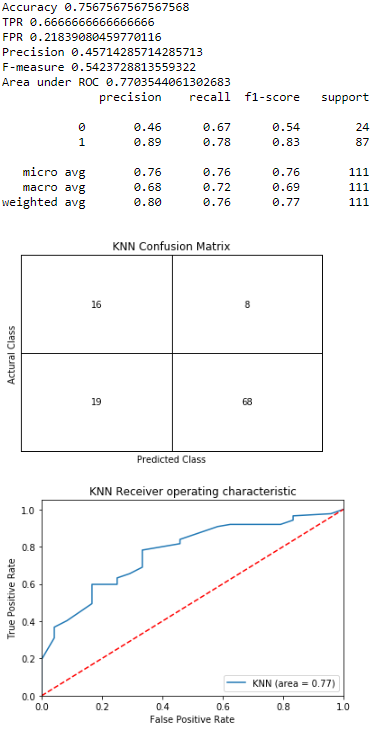
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the KNN model for the Eighth Iteration:



Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the KNN model for the Ninth Iteration:

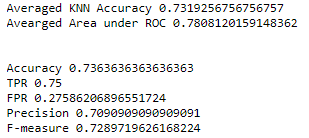
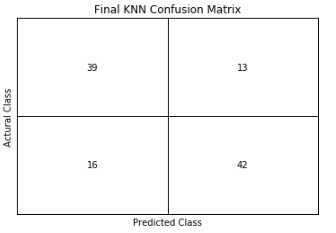


Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the KNN model for the Tenth Iteration:



After ten iterations, I have the statistic information including confusion matrix with the accuracy, True Positive and False Positive rates, Precision, and F-measure. In addition, I also plot the ROC curve and compute the area under the ROC curve. Then, I sum all the data in each iteration. I have a new confusion matrix that is the average of the ten confusion matrices. Then I recalculate the statistic information.

Statistic Information and Confusion Matrix:

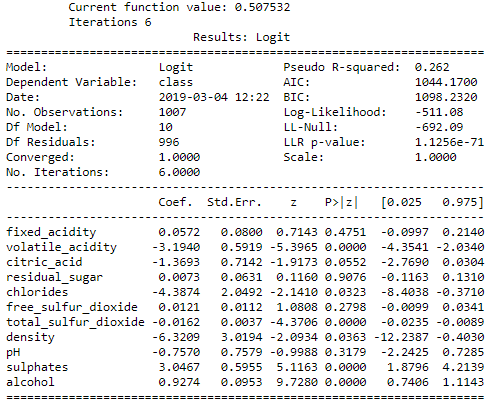
 

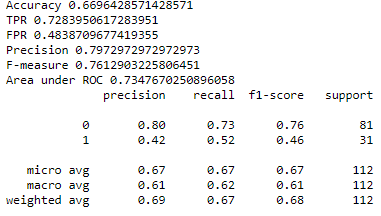
**Logistic Regression Classifier:**

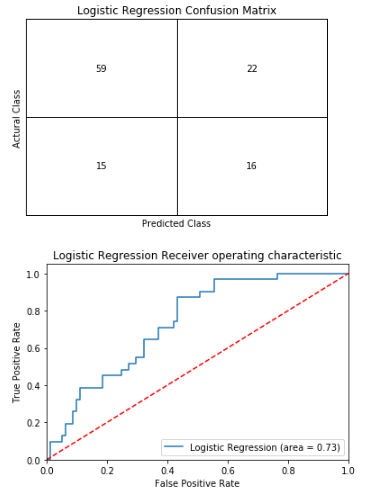
The second model I choose is the Logistic Regression model. I feed my training set into the Logistic Regression model from the sklearn.linear\_model package and use validation test set to determine the statistic information. I also print out the summary with all the attributes information to estimate all attributes for each iteration and the averaged statistic information, confusion matrix, and ROC curve.

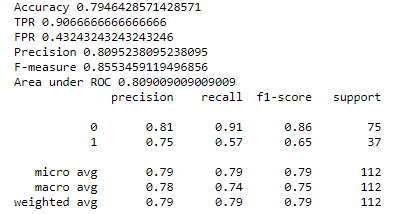
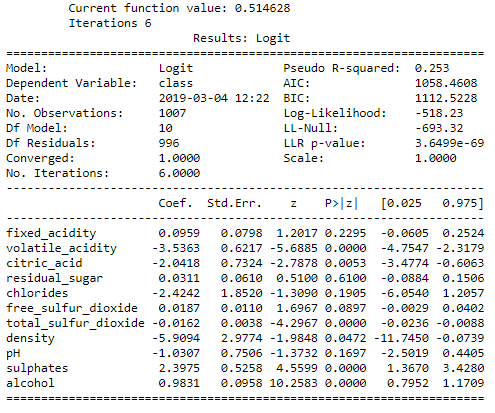
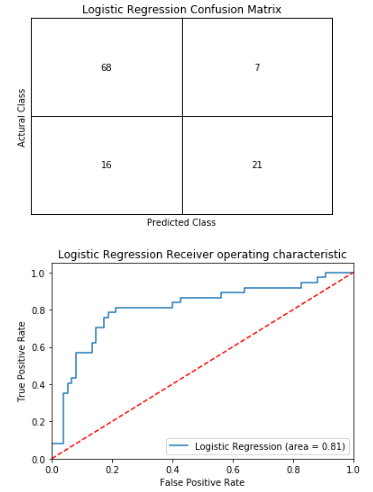
Here are the 10 iterations statistic information and ROC.

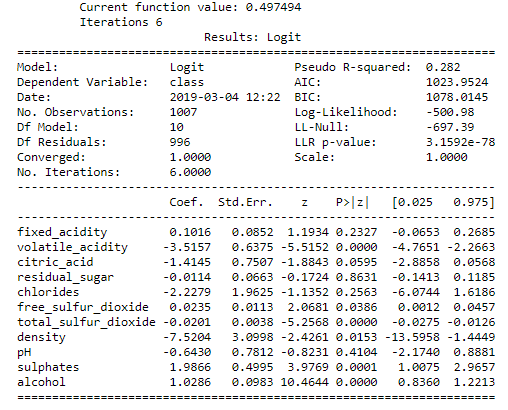
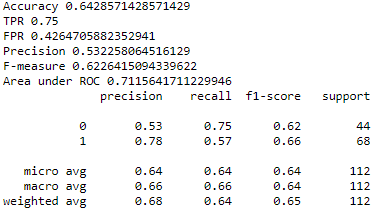
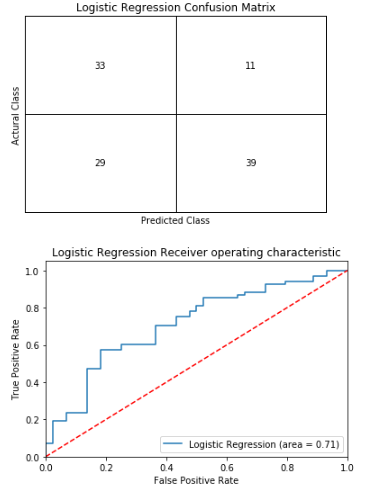
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the Logistic Regression model for the First Iteration:

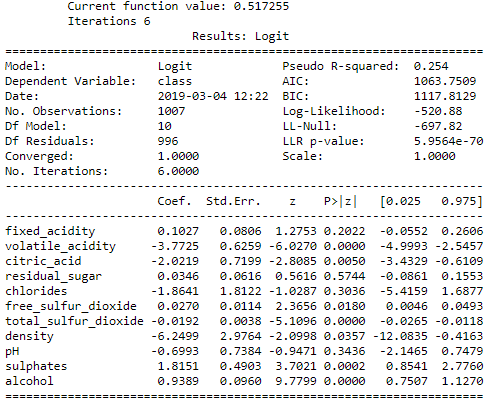
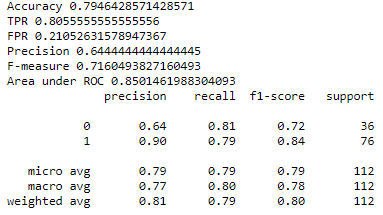
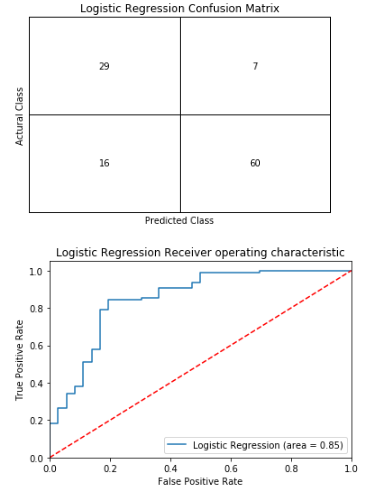


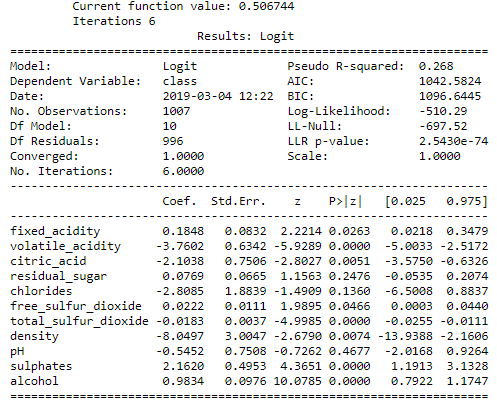
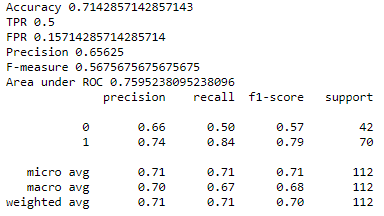
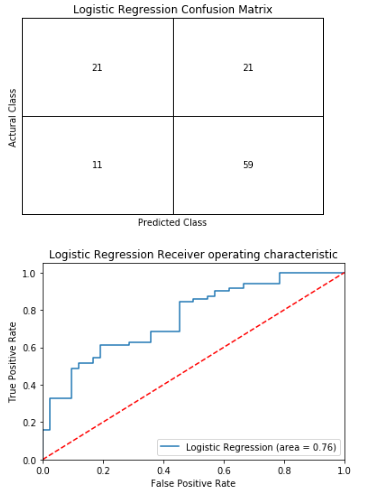


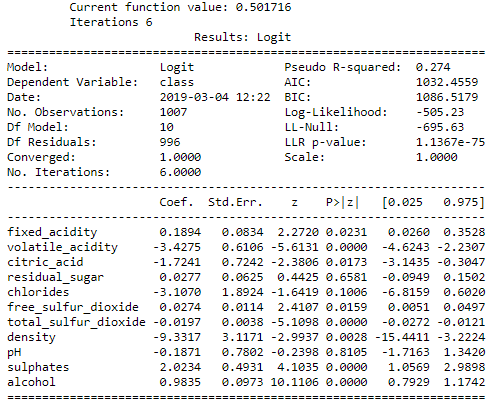
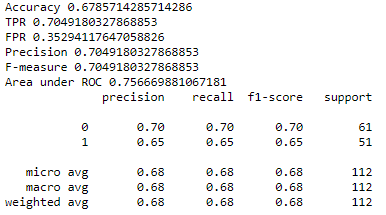
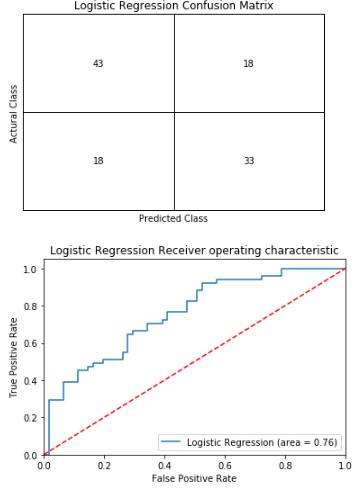


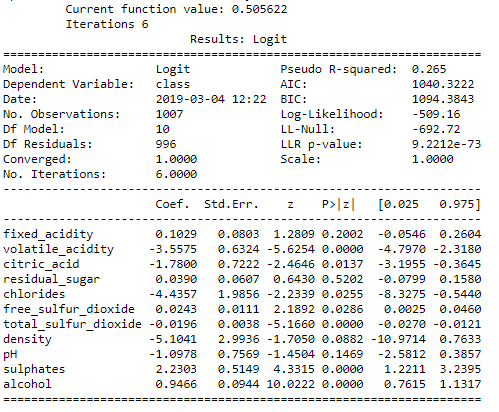
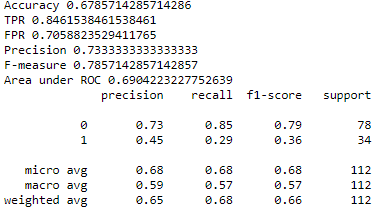
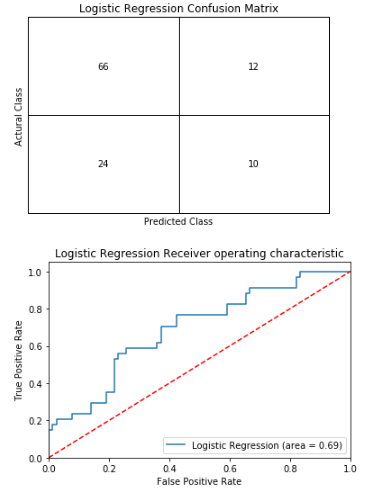
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the Logistic Regression model for the Second Iteration: 

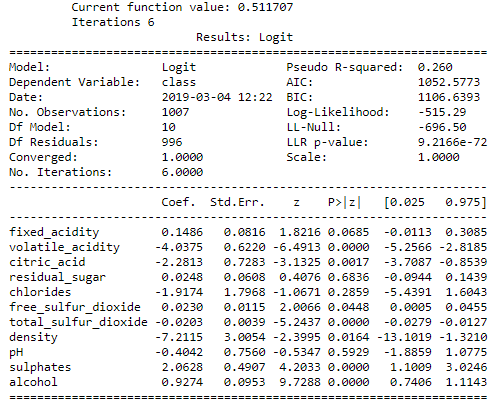
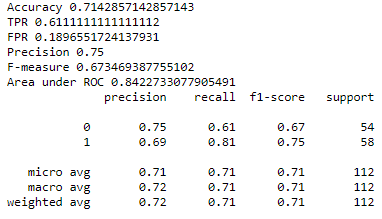
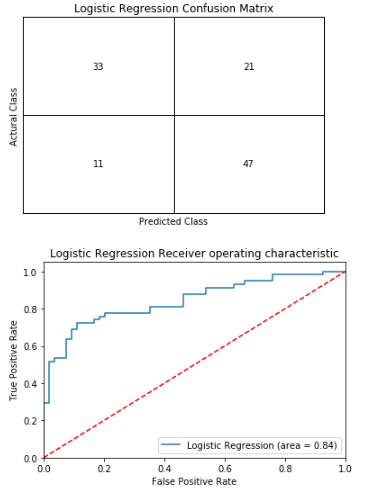
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the Logistic Regression model for the Third Iteration:  

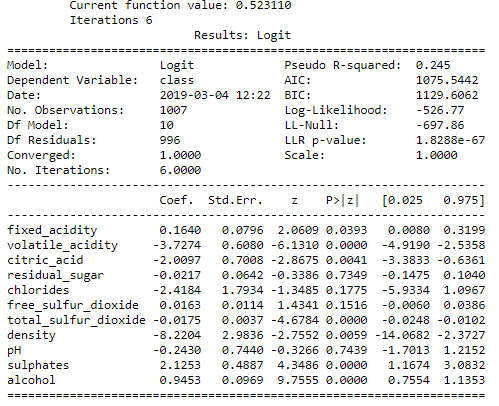
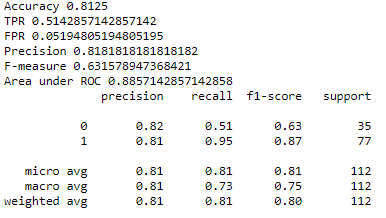
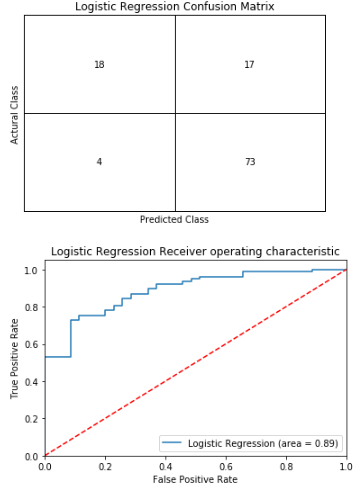
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the Logistic Regression model for the Forth Iteration:  

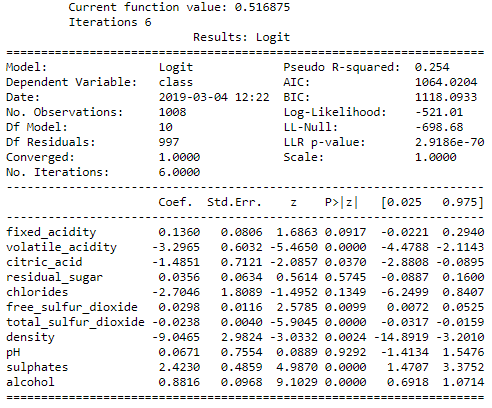
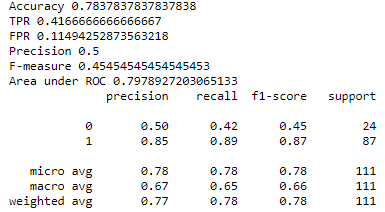
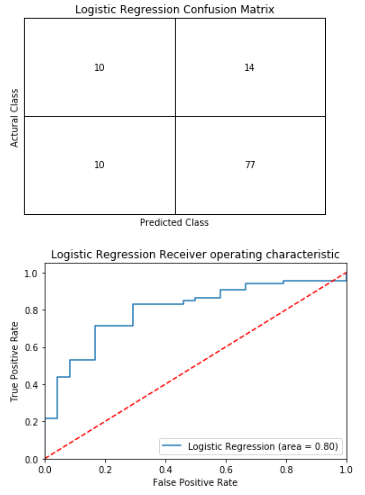
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the Logistic Regression model for the Fifth Iteration:  

Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the Logistic Regression model for the Sixth Iteration:  

Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the Logistic Regression model for the Seventh Iteration:  

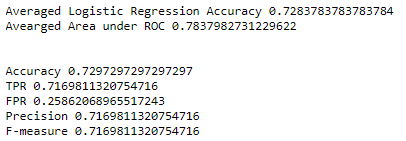
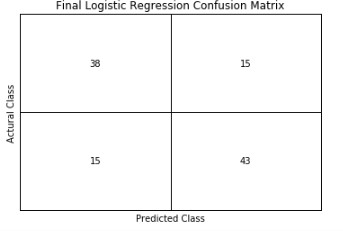
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the Logistic Regression model for the Eighth Iteration:  

Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the Logistic Regression model for the Ninth Iteration:   

Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the Logistic Regression model for the Tenth Iteration:  

After ten iterations, I have the statistic information including confusion matrix with the accuracy, True Positive and False Positive rates, Precision, and F-measure. In addition, I also plot the ROC curve and compute the area under the ROC curve. Then, I sum all the data in each iteration. I have a new confusion matrix that is the average of the ten confusion matrices. Then I recalculate the statistic information.

Statistic Information and Confusion Matrix:

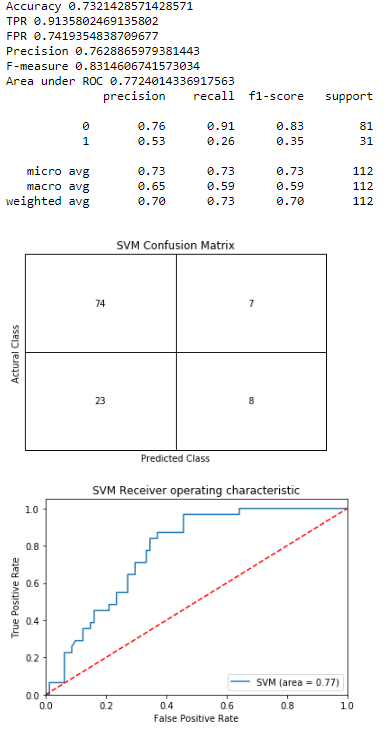
 

**SVM Classifier:**

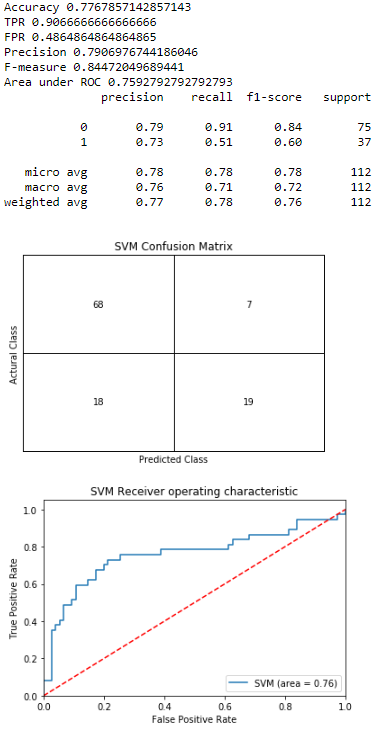
With the 10-Fold Cross Validation evaluation approach, I import SVM classifier from the sklearn package and fit the training data into SVM and use validation test set to determine the statistic information. For the SVM model, all the parameters are default except for the probability parameter. In order to calculate the ROC curve, I have to get the probability from the SVM model. Therefore, I set the probability parameter to true. The first graph below shows the classification report and the second graph contains the accuracy, True Positive, and False Positive rates, Precision rate, and F-measure of SVM model on Validation Dataset. The third graph shown below is the ROC of the SVM model for the Validation set.

Here are the 10 iterations statistic information and ROC.

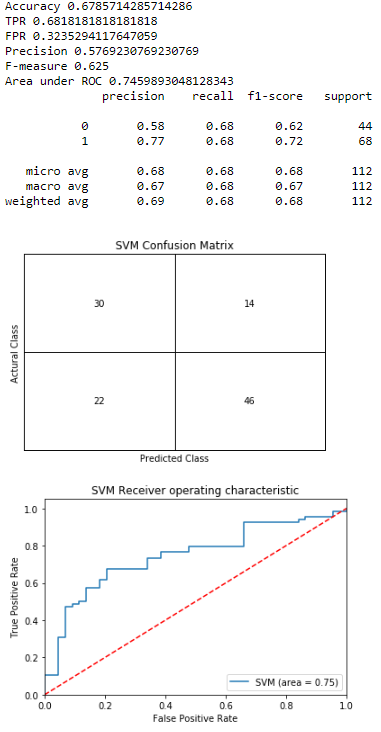
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the SVM model for the First Iteration:



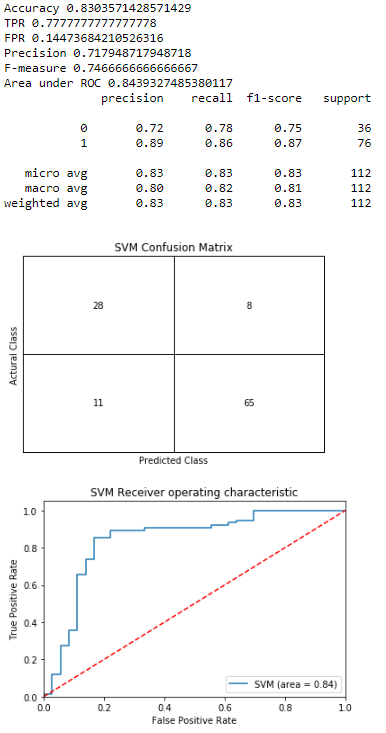
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the SVM model for the Second Iteration:



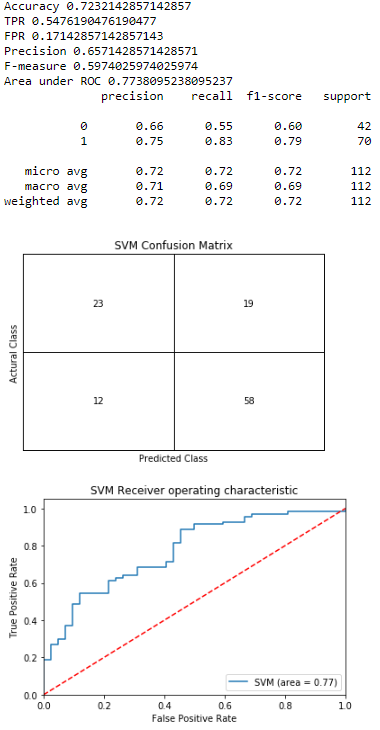
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the SVM model for the Third Iteration:



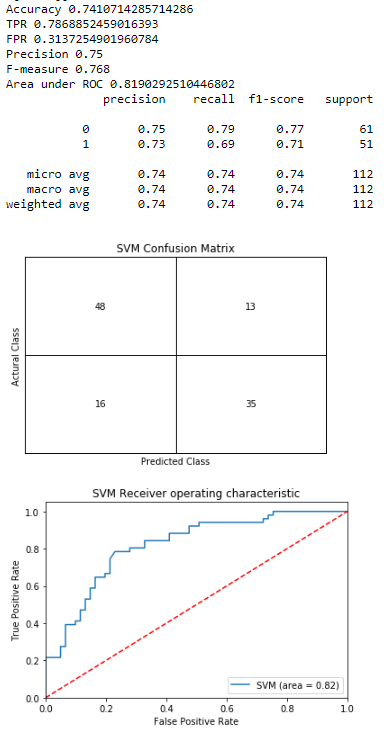
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the SVM model for the Forth Iteration:



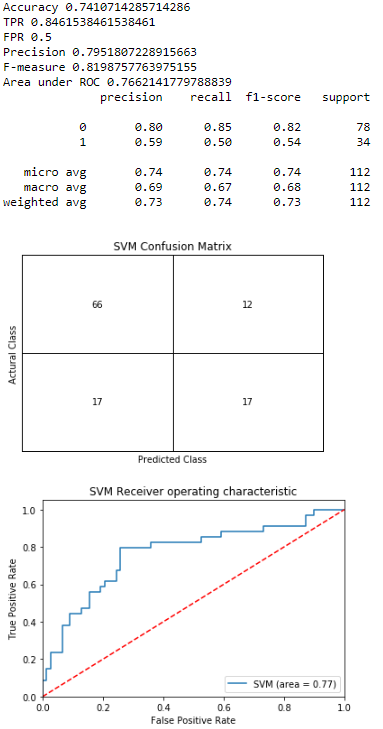
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the SVM model for the Fifth Iteration:



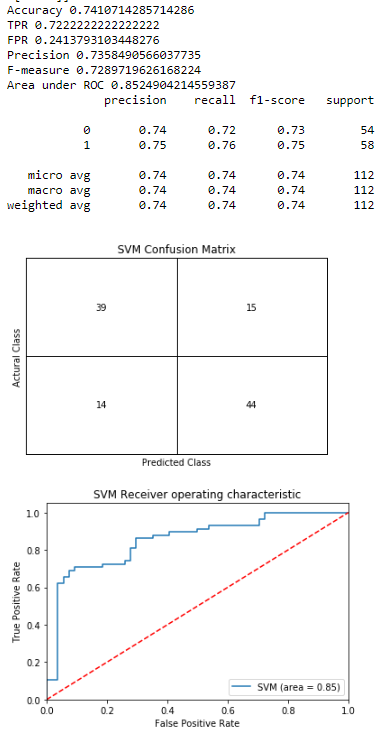
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the SVM model for the Sixth Iteration:



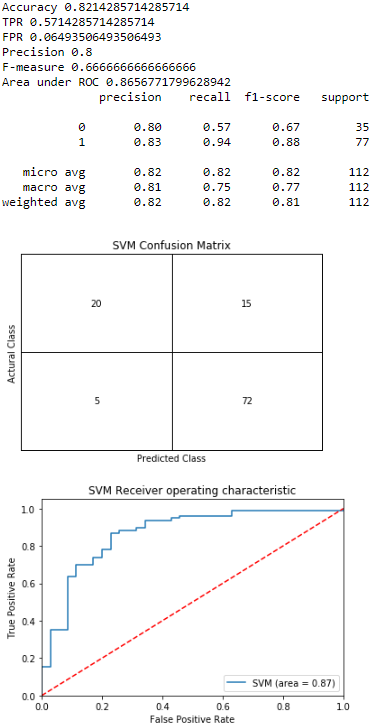
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the SVM model for the Seventh Iteration:



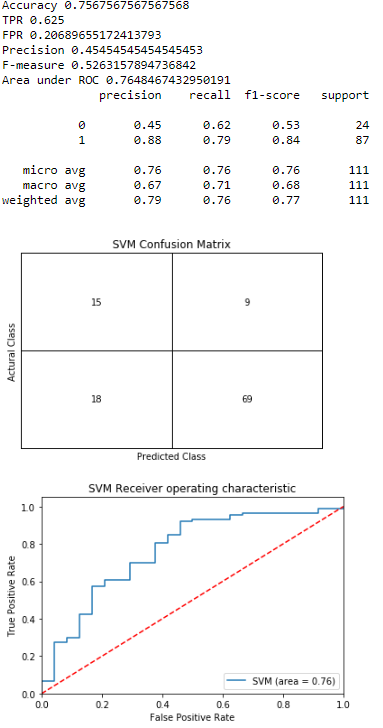
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the SVM model for the Eighth Iteration:



Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the SVM model for the Ninth Iteration:

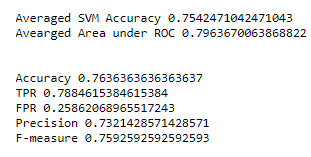
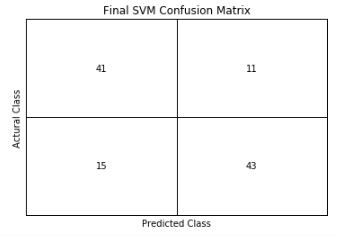


Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the SVM model for the Tenth Iteration:



After ten iterations, I have the statistic information including confusion matrix with the accuracy, True Positive and False Positive rates, Precision, and F-measure. In addition, I also plot the ROC curve and compute the area under the ROC curve. Then, I sum all the data in each iteration. I have a new confusion matrix that is the average of the ten confusion matrices. Then I recalculate the statistic information.

Statistic Information and Confusion Matrix:

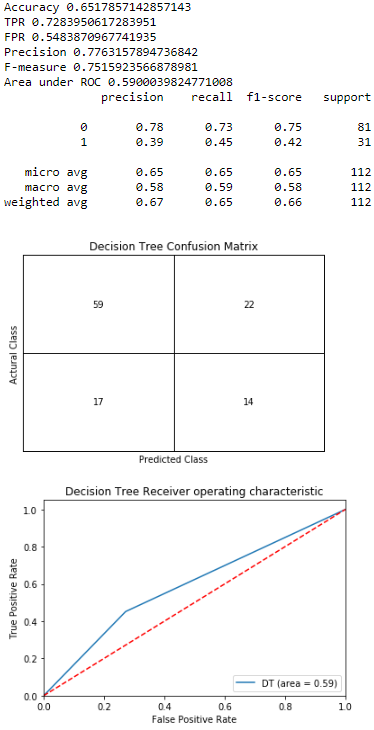
 

**Decision Tree Classifier:**

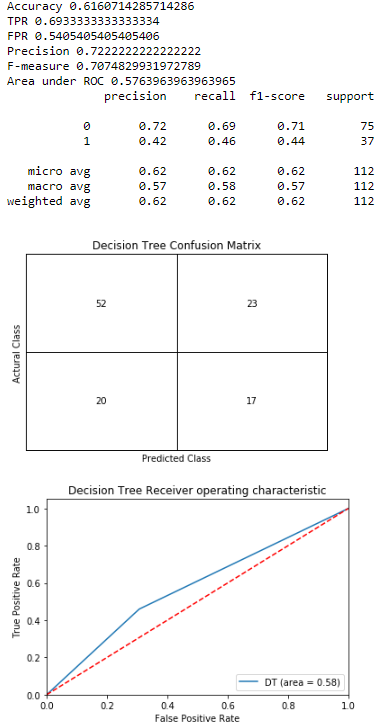
Based on the 10-Fold Cross Validation evaluation approach, I feed the training data into the Decision Tree model from the sklearn package and use the validation test set to determine the statistic information. The first graph below shows the classification report and the second graph contains the accuracy, True Positive, and False Positive rates, Precision rate, and F-measure of the Decision Tree model. The third graph shown below is the ROC of the Decision Tree model.

Here are the 10 iterations statistic information and ROC.

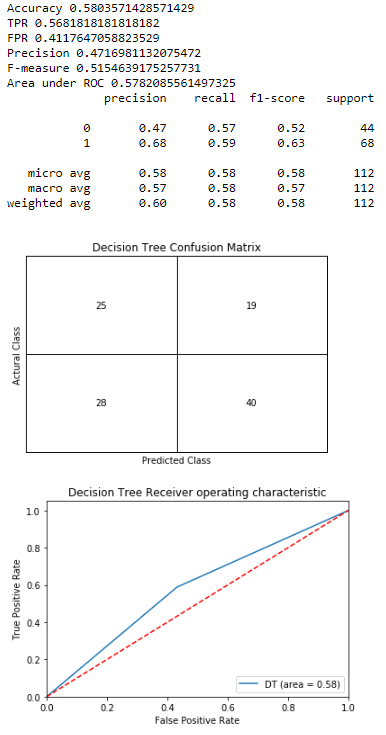
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the Decision Tree model for the First Iteration:



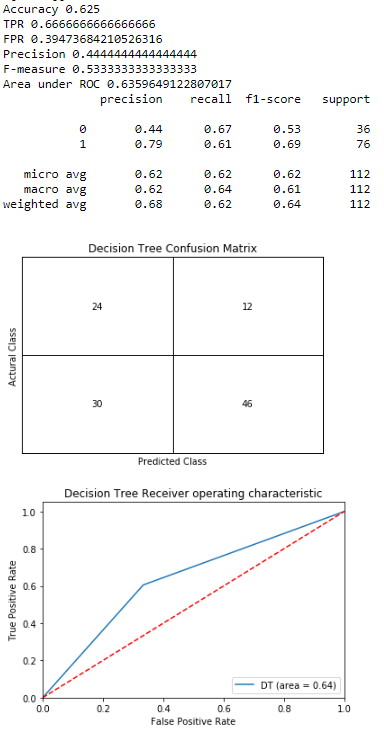
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the Decision Tree model for the Second Iteration:



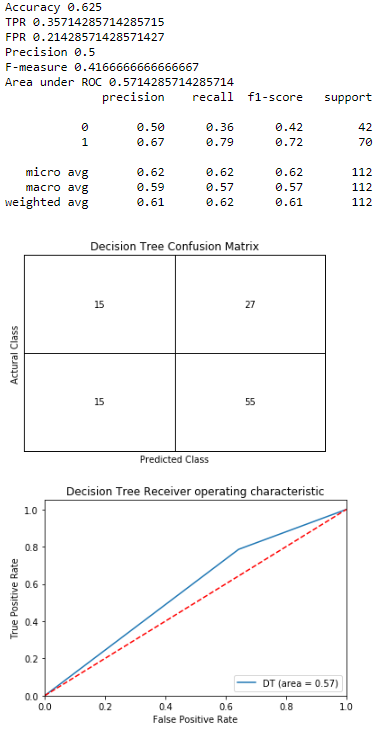
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the Decision Tree model for the Third Iteration:



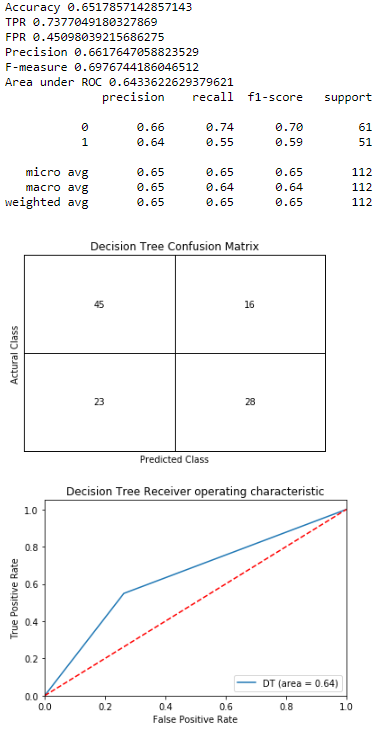
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the Decision Tree model for the Forth Iteration:



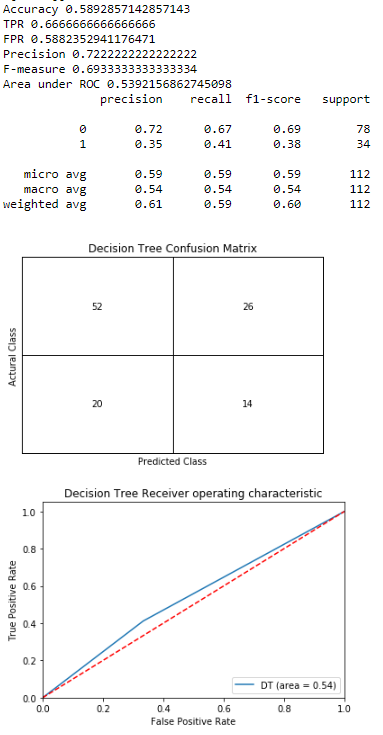
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the Decision Tree model for the Fifth Iteration:



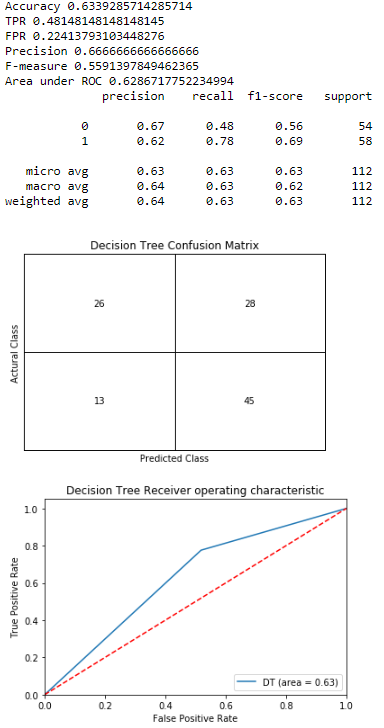
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the Decision Tree model for the Sixth Iteration:



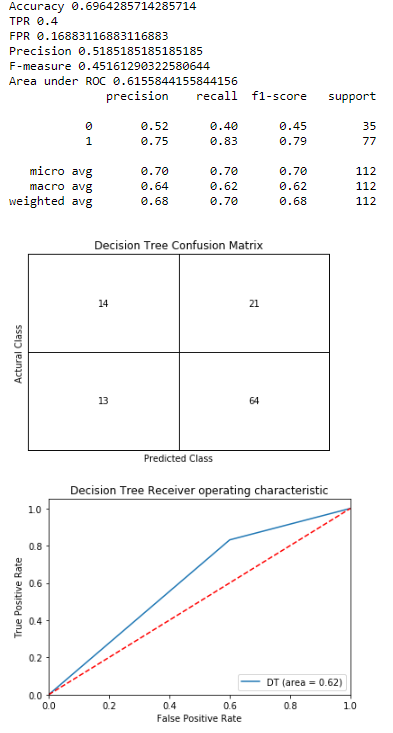
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the Decision Tree model for the Seventh Iteration:



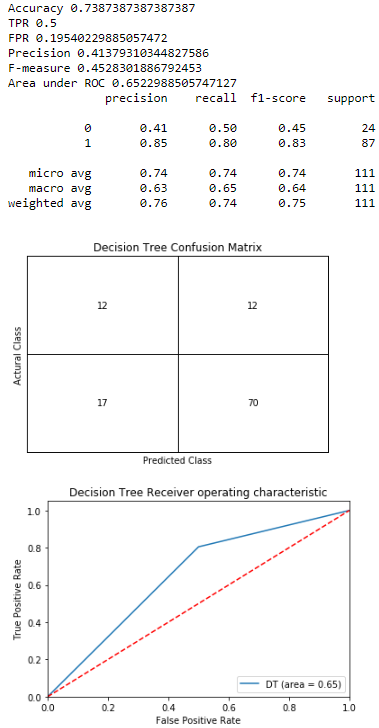
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the Decision Tree model for the Eighth Iteration:



Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the Decision Tree model for the Ninth Iteration:

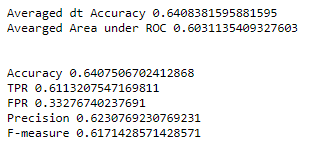
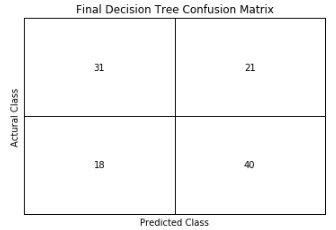


Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the Decision Tree model for the Tenth Iteration:

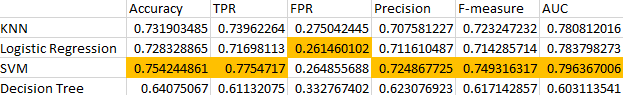


After ten iterations, I have the statistic information including confusion matrix with the accuracy, True Positive and False Positive rates, Precision, and F-measure. In addition, I also plot the ROC curve and compute the area under the ROC curve. Then, I sum all the data in each iteration. I have a new confusion matrix that is the sum of the ten confusion matrices. Then I recalculate the statistic information.

Statistic Information and Confusion Matrix:

**Final Model:**

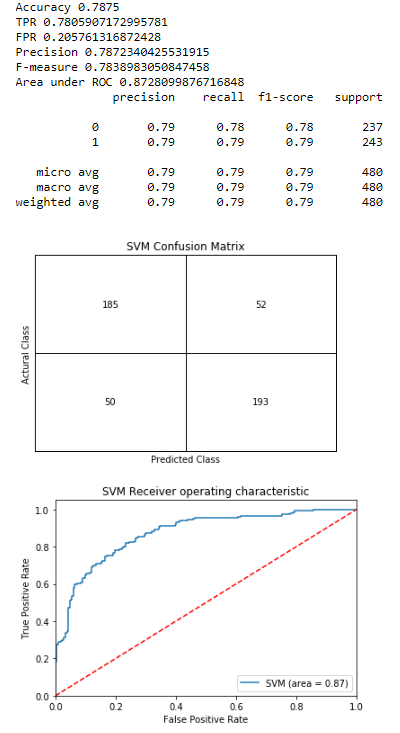


Based on the performance of the four model, accuracy, TPR, FPR, Precision, F-Measure, and the area under ROC of the model can be the factors to determine the best model fit in the dataset. In order to have the best performance, compared with the accuracy, TPR, FPR, Precision, F-Measure, and AUC, the best model is SVM among the four models. The performance of SVM Model is the best among the four models expect for FPR. The Accuracy of the SVM model 75.4% can have the best prediction for the class label and less error. Furthermore, the Highest True Positive Rate 77.5% can give me the highest possible correct prediction for the True class. The Highest Precision 72.5% can also give me the best possible prediction precisely. The highest F-measure 74.9% and AUC 79.6% also shows that this model has better performance among the four models. Even though the False Positive Rate is not the lowest, there is only less than 1% difference between the other three models, which is not totally fine. Therefore, the SVM model still has a relatively better performance among the four models based on the best performance on Accuracy, TPR, Precision, F-Measure, and AUC. Then I use the SVM model as the best model. Then I feed the test set into SVM to check the performance on the test set.

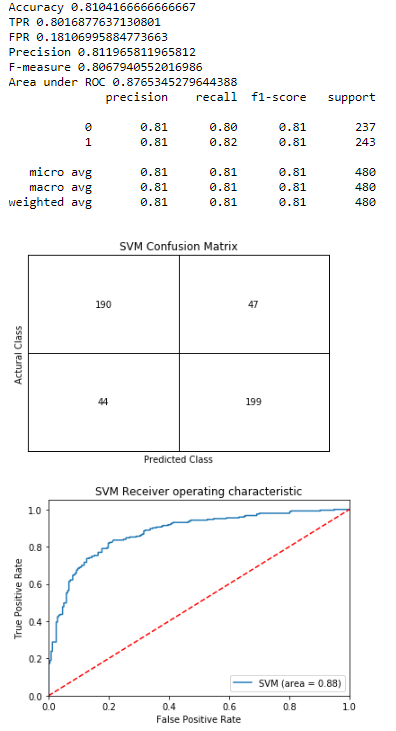
The performance of SVM model on the test set has 79.1% accuracy with 87.3% AUC. The first graph shows the classification report and the second graph below contains the accuracy, True Positive, and False Positive rates, Precision rate, and F-measure of Naïve Bayes model. The third graph shown below is the ROC of the SVM model for the test set.

Generalization Performance for SVM:

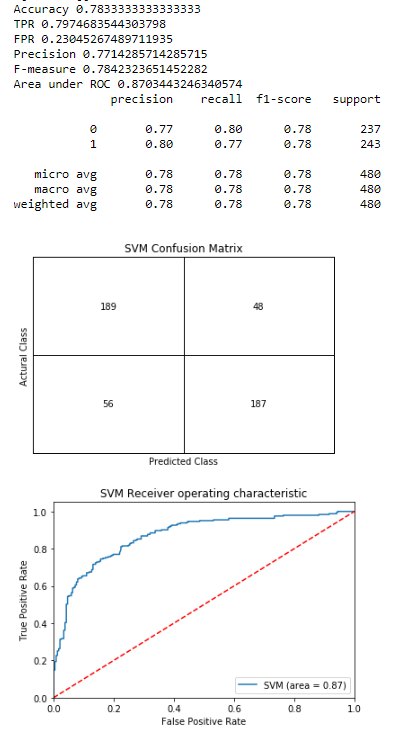
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the SVM model for the First Iteration:



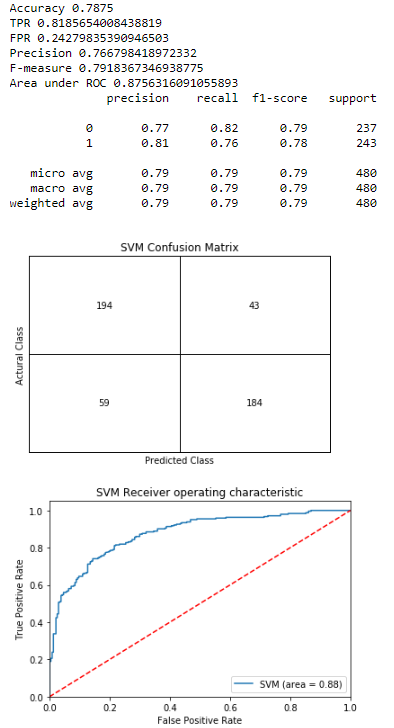
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the SVM model for the Second Iteration:



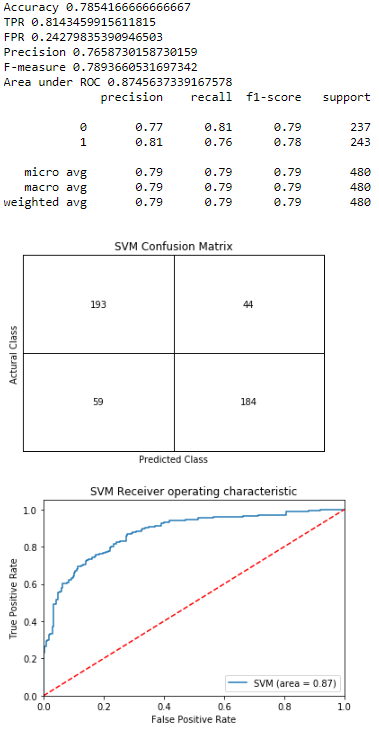
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the SVM model for the Third Iteration:



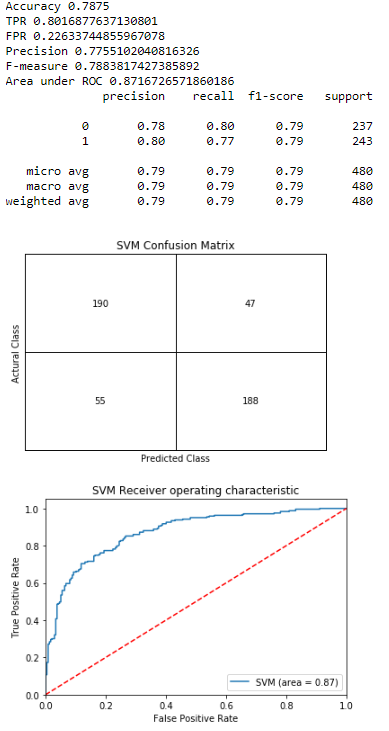
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the SVM model for the Forth Iteration:



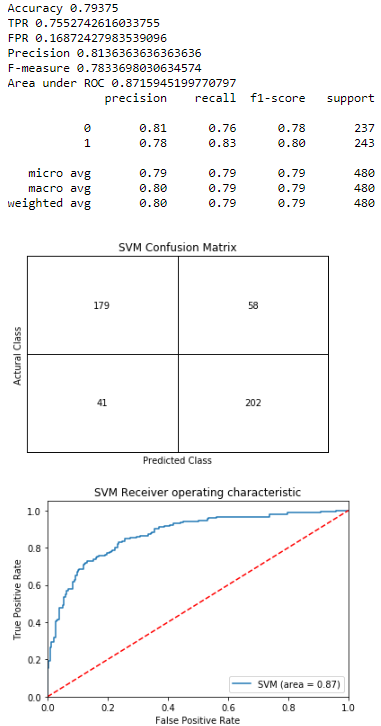
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the SVM model for the Fifth Iteration:



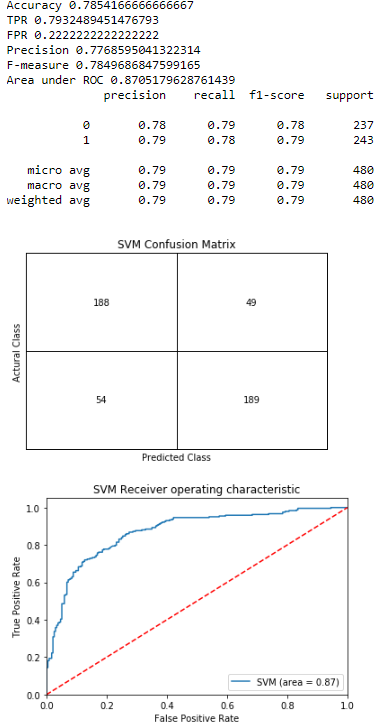
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the SVM model for the Sixth Iteration:



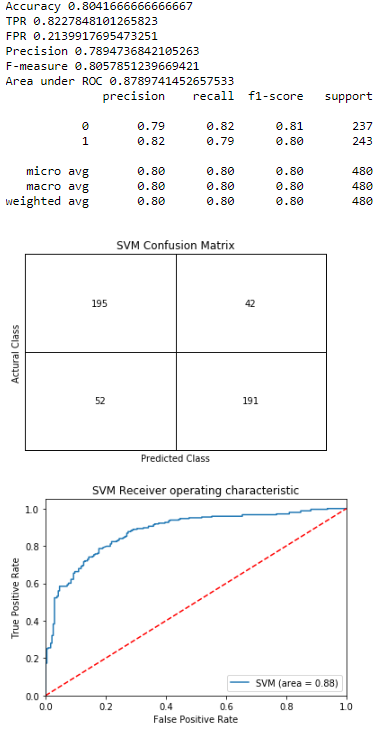
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the SVM model for the Seventh Iteration:



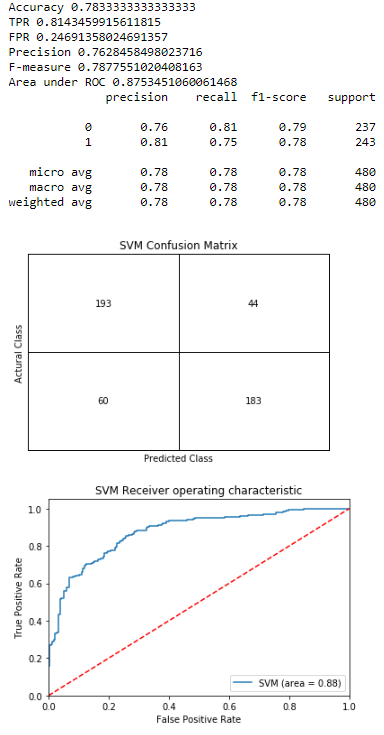
Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the SVM model for the Eighth Iteration:



Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the SVM model for the Ninth Iteration:



Statistic Information, Classification Report, Confusion Matrix and ROC Curve of the SVM model for the Tenth Iteration:



After ten iterations, I have the statistic information including confusion matrix with the accuracy, True Positive and False Positive rates, Precision, and F-measure. In addition, I also plot the ROC curve and compute the area under the ROC curve. Then, I sum all the data in each iteration. I have a new confusion matrix that is the sum of the ten confusion matrices. Then I recalculate the statistic information.

Statistic Information and Confusion Matrix:

