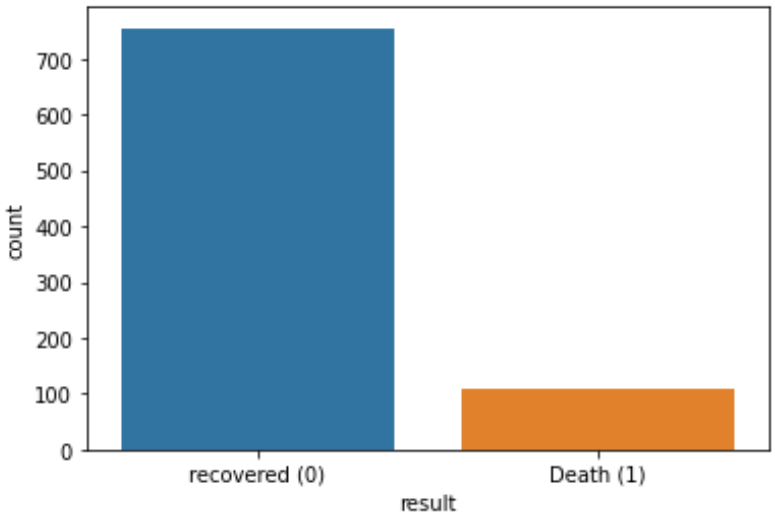
Name: Salma Emad El-mongi

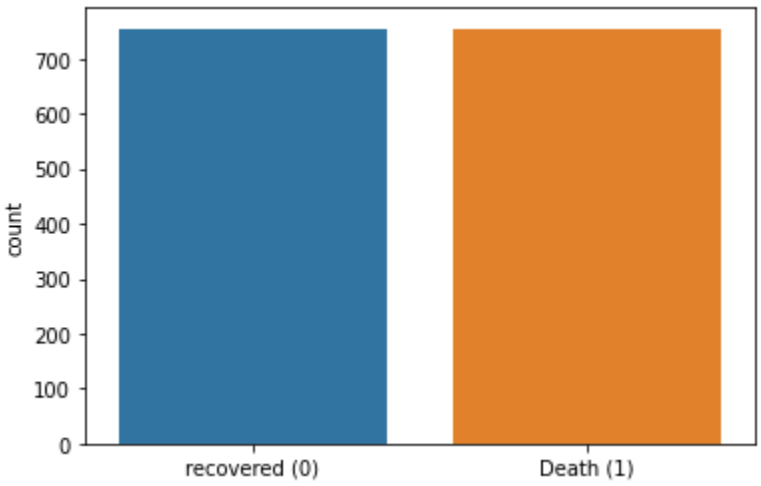
ID: 20398567

1) loading the data without Unnamed column 0

2) using the smote technique to balance the data

These plots show the data before and after using the smote technique





3) splitting the data

4) Optimization and evaluations:

1- K_ nearest Neighbour

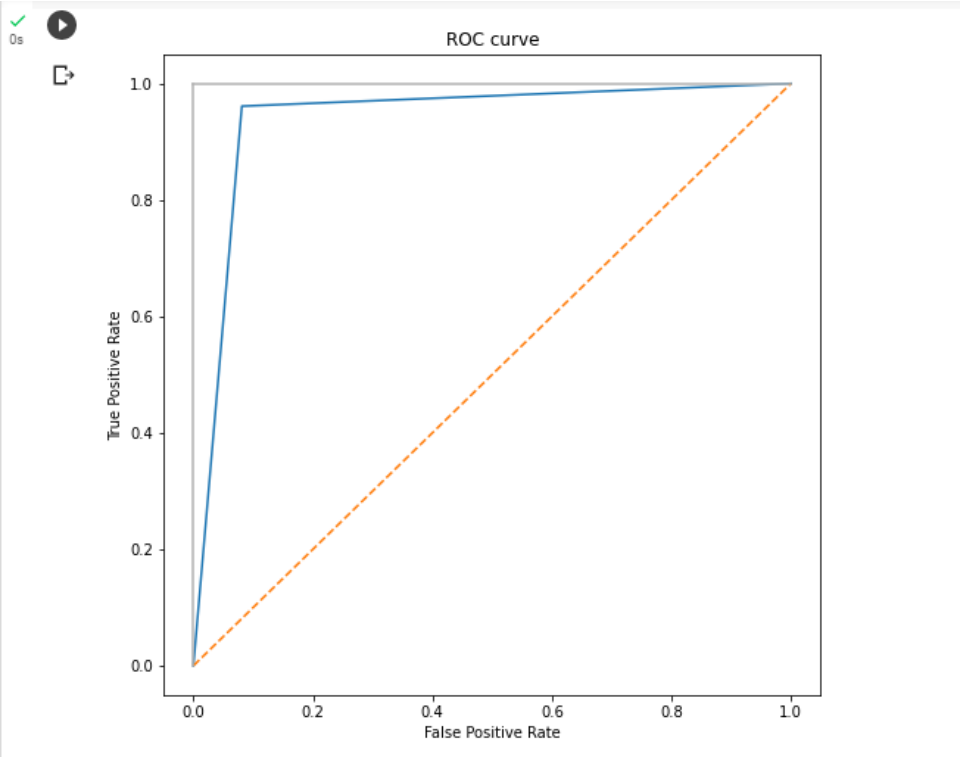
*Choose the best cross validation for best recall

*Tuning K in KNN classifier

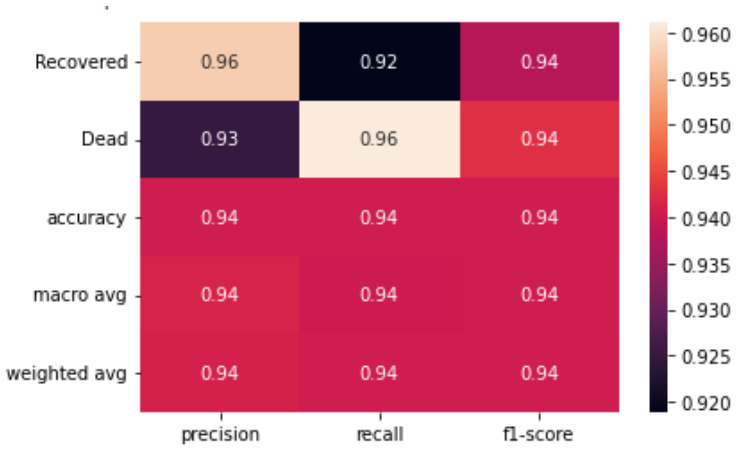
*Printing the best score and parameter to build a new model with optimal parameter

*Printing the recall score, f1 score, roc_auc_score , accuracy score, precision score

*plotting the ROC curve



*Visualizing the scores



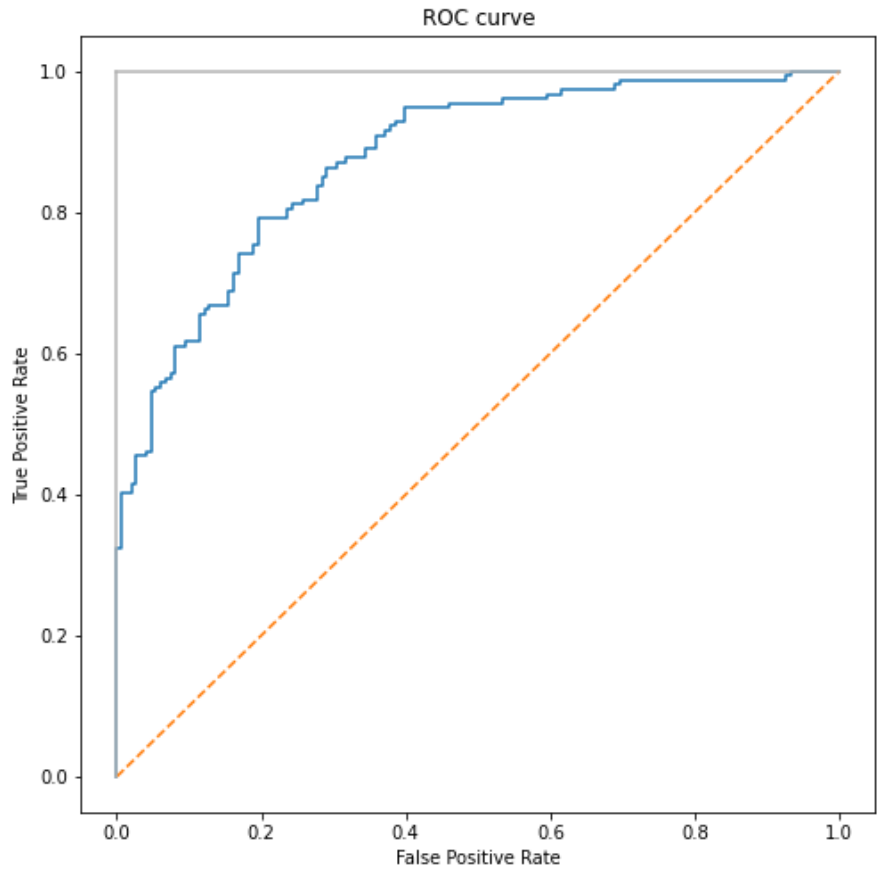
2- Naïve Bayes

*Replacing the negative values with zeros

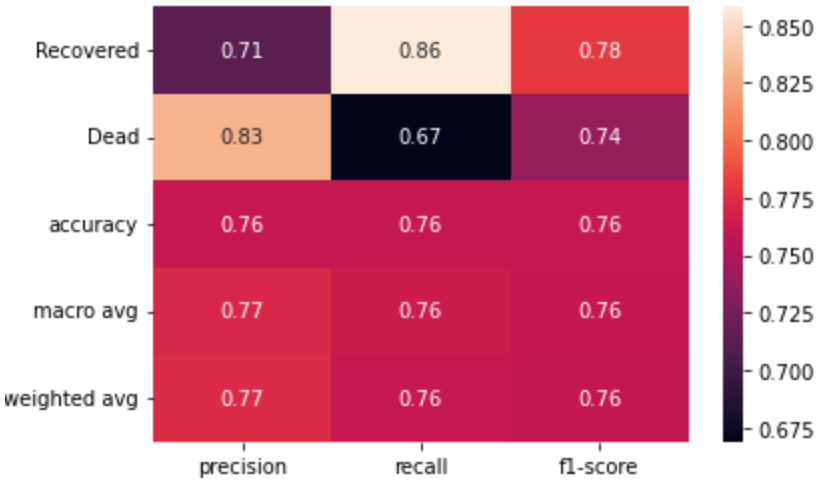
*Using MultinomialNB because it doesn't take negative values

*Printing the recall score, f1 score, roc_auc_score , accuracy score, precision score

*Plotting the ROC curve



*Visualizing the scores



3- Decision Tree

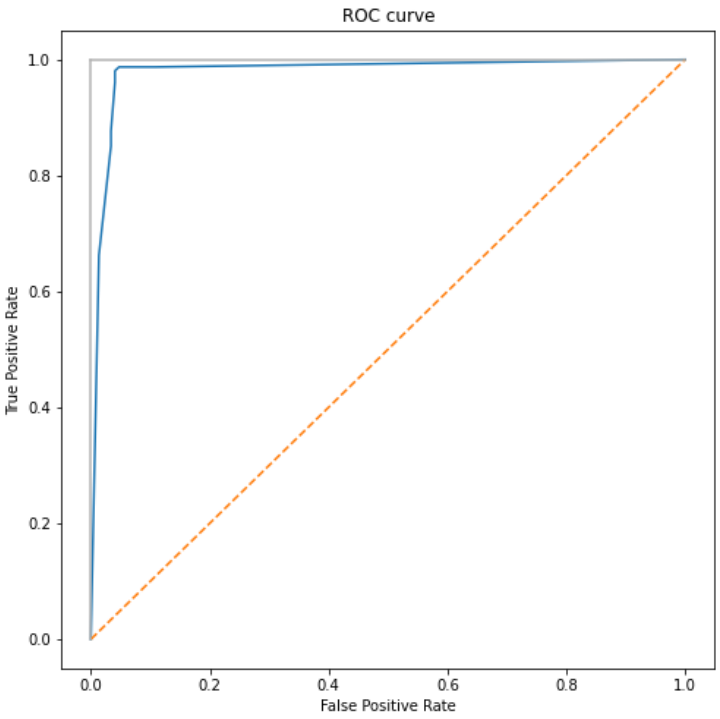
*Choose the best cross validation for best recall

* tuning parameters in Decision Tree Classifier

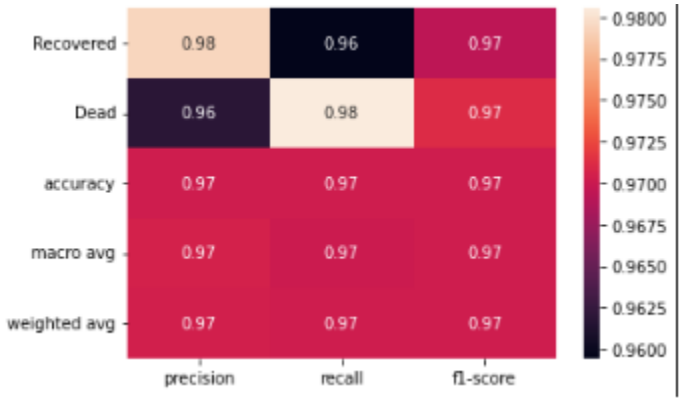
*Printing the best score and parameter to build new model with optimal parameters

*Printing the recall score, f1 score, roc_auc_score , accuracy score, precision score

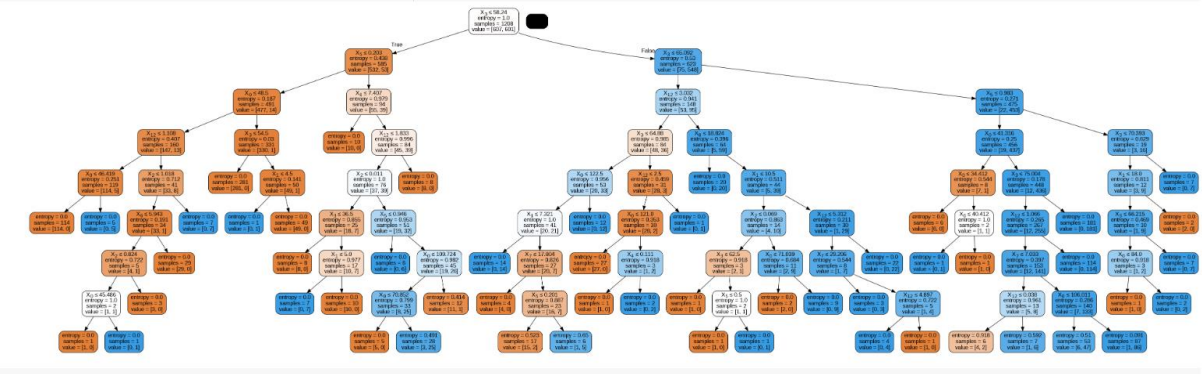
*Plotting the ROC curve



*Visualizing the scores



*Visualizing the Decision Tree



4- Support Vector Machine

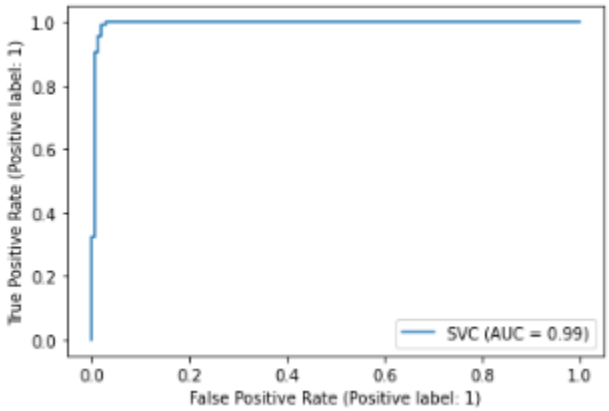
*Choose the best cross validation for best recall

* tuning parameters in SVM Classifier

*Printing the best score and parameter to build new model with optimal parameters

*Printing the recall score, f1 score, roc_auc_score , accuracy score, precision score

*Plotting the ROC curve



*Visualizing the scores

5- Logistic Regression

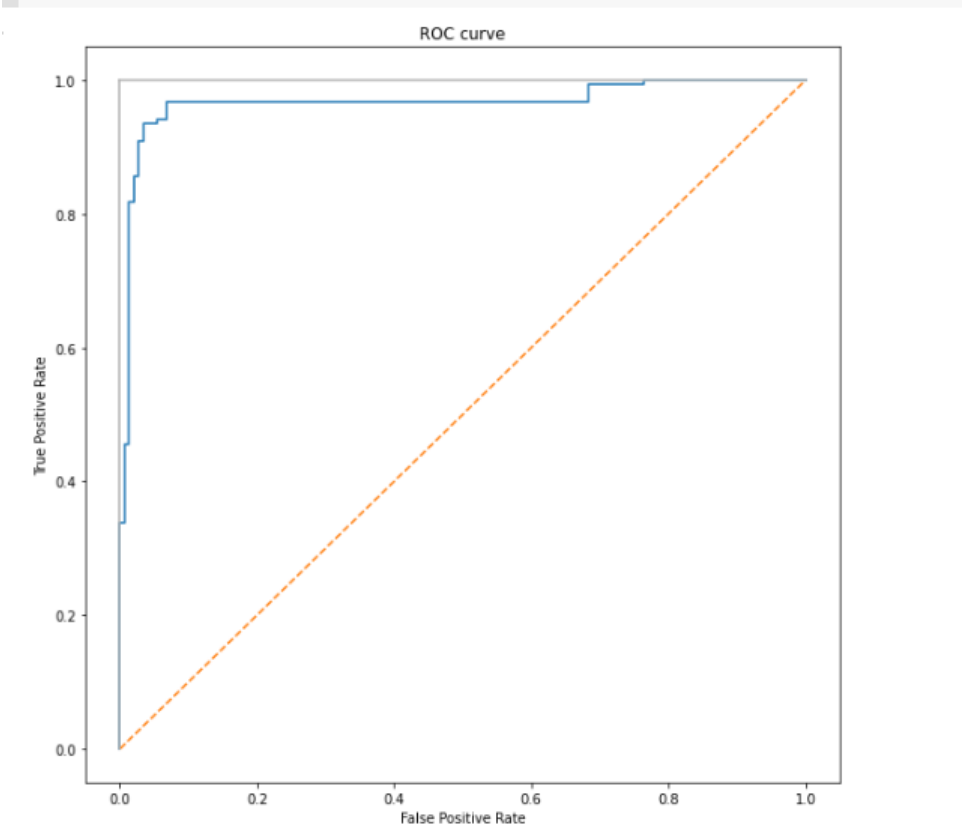
*Choose the best cross validation for best recall

* Tuning parameters in Logistic Regression Classifier

*Printing the best score and parameter to build new model with optimal parameters

*Printing the recall score, f1 score, roc_auc_score , accuracy score, precision score

*Plotting the ROC curve



*Visualizing the scores

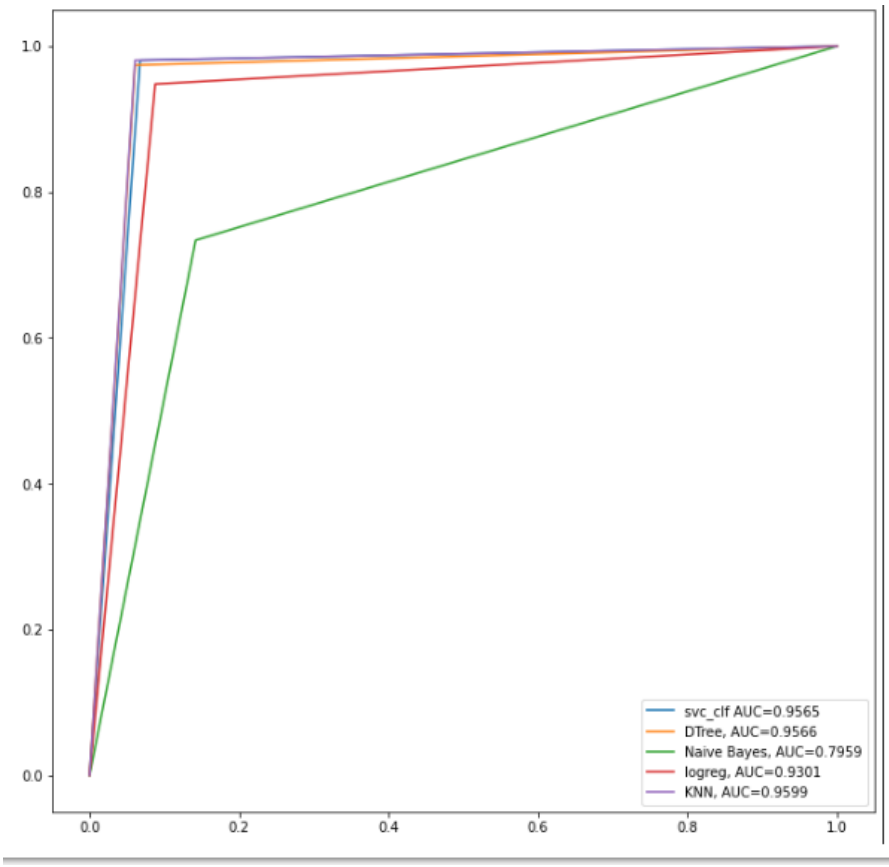
Comparison between each model

*The highest accuracy: KNN

*The lowest accuracy: Naive Bayes

	Classifier	Accuracy	precision	Recall	F1 score	ROC/AUC
0	KNN	0.960265	0.943750	0.980519	0.961783	0.961783
1	Naive bayes	0.794702	0.843284	0.733766	0.784722	0.784722
2	Decision Tree	0.956954	0.943396	0.974026	0.958466	0.958466
3	Logistic regression	0.956954	0.943396	0.974026	0.958466	0.958466
4	SVM	0.956954	0.943396	0.974026	0.958466	0.958466

Plot showing the area under curve


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
svc_clf AUC=0.9565
DTree, AUC=0.9566
Naive Bayes, AUC=0.7959
logreg, AUC=0.9301
KNN, AUC=0.9599