



**Data Glacier**

Your Deep Learning Partner

# Data Science: Healthcare Persistency of a drug (Group Project)

Dec 2021

# Team Members

Team				
Name	Email	Country	College/Company	Specialization
Bao Khanh Nguyen	<a href="mailto:nguyenkhanhbao8695@gmail.com">nguyenkhanhbao8695@gmail.com</a>	USA	American Energy Project	Data Science
Seyedeh Marzieh Hosseini	<a href="mailto:shosseini@uni-potsdam.de">shosseini@uni-potsdam.de</a>	Germany	University of Potsdam	Data Science
Guillermo Leija	<a href="mailto:leija.guillermo@gmail.com">leija.guillermo@gmail.com</a>	Mexico	IEBS Business School	Data Science
Zain Ul Haq	<a href="mailto:zainulhaq904@gmail.com">zainulhaq904@gmail.com</a>	Germany	University of Rostock	Data Science

# Agenda

- **Executive Summary**
- **Problem Statement**
- **Approach**
- **EDA**
- **Model Development**
- **Model Selection**
- **Model Evaluation**
- **Conclusion**

# Executive Summary

ABC company also one of pharmaceutical companies, wants to know how long a medicine will last in a patient's system (persistency of a drug). Based on prescription data, the ABC corporation needs to determine whether a patient is persistent or not. ABC pharma would manufacture medicines in that number based on the persistency count so that they could operate their firm effectively and avoid the risks of NTM infections.

## **ML Problem:**

With an objective to gather insights on the factors that are impacting the persistency, build a classification for the given dataset

**Target Variable:** Persistency\_Flag

# Problem Description

ABC is a pharmaceutical business that wants to know the persistency of a drug after a physician has prescribed it for a patient. This company has approached an analytics firm to automate the identifying procedure. This analytics firm has entrusted our team with the task of developing a solution to automate the persistence of a medicine for the client ABC.

# Business Understanding

One of the long-lasting business issues in the world of pharmaceutical companies is the persistency of drugs which can significantly affect the outcome of medical treatments. One of the important factors that is related to persistency is the adherence of the patient to the prescribed regimens, meaning if the patient is committed to the prescribed regimens or not. There is a lot of information about Non-Tuberculous Mycobacterial (NTM) infections. In fact, related studies show that around 50%-60% of the patients with different illnesses in US miss doses, take the wrong doses, or drop off treatment in the first year. Additionally, the illness, either chronic or acute can be related to the adherence and persistency of drugs.

# Projects steps

- Problem understanding
- Data Understanding
- Data Cleaning and Feature engineering
- Model Development
- Model Selection
- Model Evaluation
- Report the accuracy, precision and recall of both the class of target variable
- Report ROC-AUC as well
- Deploy the model
- Explain the challenges and model selection

# Methodologies

- Data was taken from github and analysed
- Problem understanding
- Data Understanding
- Data Cleaning and Feature engineering
- Model Development
- Model Selection
- Model Evaluation



# Data Cleaning

- Checking Missing Value/ NAN / Null Data
- Checking Outliers
- Data Wrangling , Transformation and Standardization

# Analyzing dependency of variable (Before Transformation)

Non-Persistent : 62.35 %  
Persistent : 37.65 %

The analysis showed more non persistence of drugs than persistence

## Missing Values

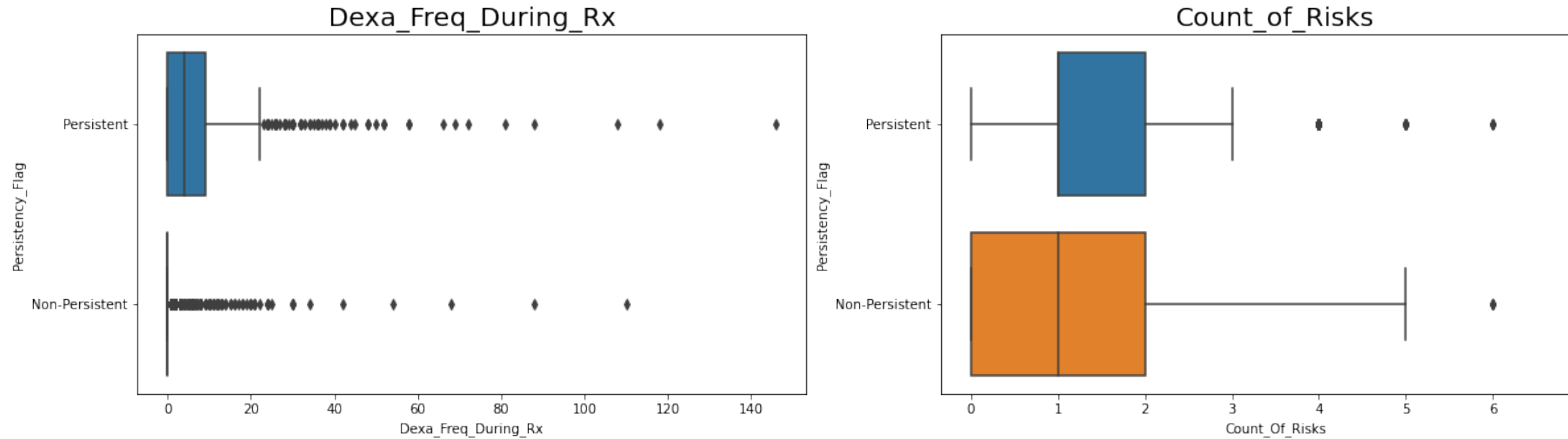
```
df.isnull().sum()
```

**No missing values were found**

```
for col in df.columns:  
    pct_missing = np.mean(df[col].isnull())  
    print('{} - {}'.format(col, pct_missing))
```

```
Ptid - 0.0  
Persistency_Flag - 0.0  
Gender - 0.0  
Race - 0.0  
Ethnicity - 0.0  
Region - 0.0  
Age_Bucket - 0.0  
Ntm_Speciality - 0.0  
Ntm_Specialist_Flag - 0.0  
Ntm_Speciality_Bucket - 0.0  
Gluko_Record_Prior_Ntm - 0.0  
Gluko_Record_During_Rx - 0.0  
Dexa_Freq_During_Rx - 0.0  
Dexa_During_Rx - 0.0  
Frag_Frac_Prior_Ntm - 0.0  
Frag_Frac_During_Rx - 0.0  
Risk_Segment_Prior_Ntm - 0.0  
Tscore_Bucket_Prior_Ntm - 0.0  
Risk_Segment_During_Rx - 0.0  
Tscore_Bucket_During_Rx - 0.0  
Change_T_Score - 0.0  
Change_Risk_Segment - 0.0  
Adherent_Flag - 0.0  
Idn_Indicator - 0.0  
Injectable_Experience_During_Rx - 0.0
```

# Checking Outliers



- both DEXA\_Freq\_During\_Rx and Count\_of\_Risks variables contains outliers
- Using IQR-score and Z-score to deal with them.

# Data Transformation to resolve outliers

- Our approach to deal with the skewness and outliers for these variables is using IQR Score. To remove outliers, this approach uses the IQR values calculated before. Anything outside of the range of  $(Q1 - 1.5 \text{ IQR})$  and  $(Q3 + 1.5 \text{ IQR})$  is considered an outlier and should be eliminated.
- In this project, for DEXA\_Freq\_During\_Rx and Count\_of\_Risks, we will remove any data outside of the range of  $(Q1 - 1.5 \text{ IQR})$  and  $(Q3 + 1.5 \text{ IQR})$  or two whiskers.
- For the Old Shape (3424,69) for both of Count of Risk" and "DEXA Freq During Rx" variable but after removed the outliers with this method, the new shape only (2964,69), which remove 460 data outside of the range of  $(Q1 - 1.5 \text{ IQR})$  and  $(Q3 + 1.5 \text{ IQR})$ .

## LABEL Encoding for categorical variables

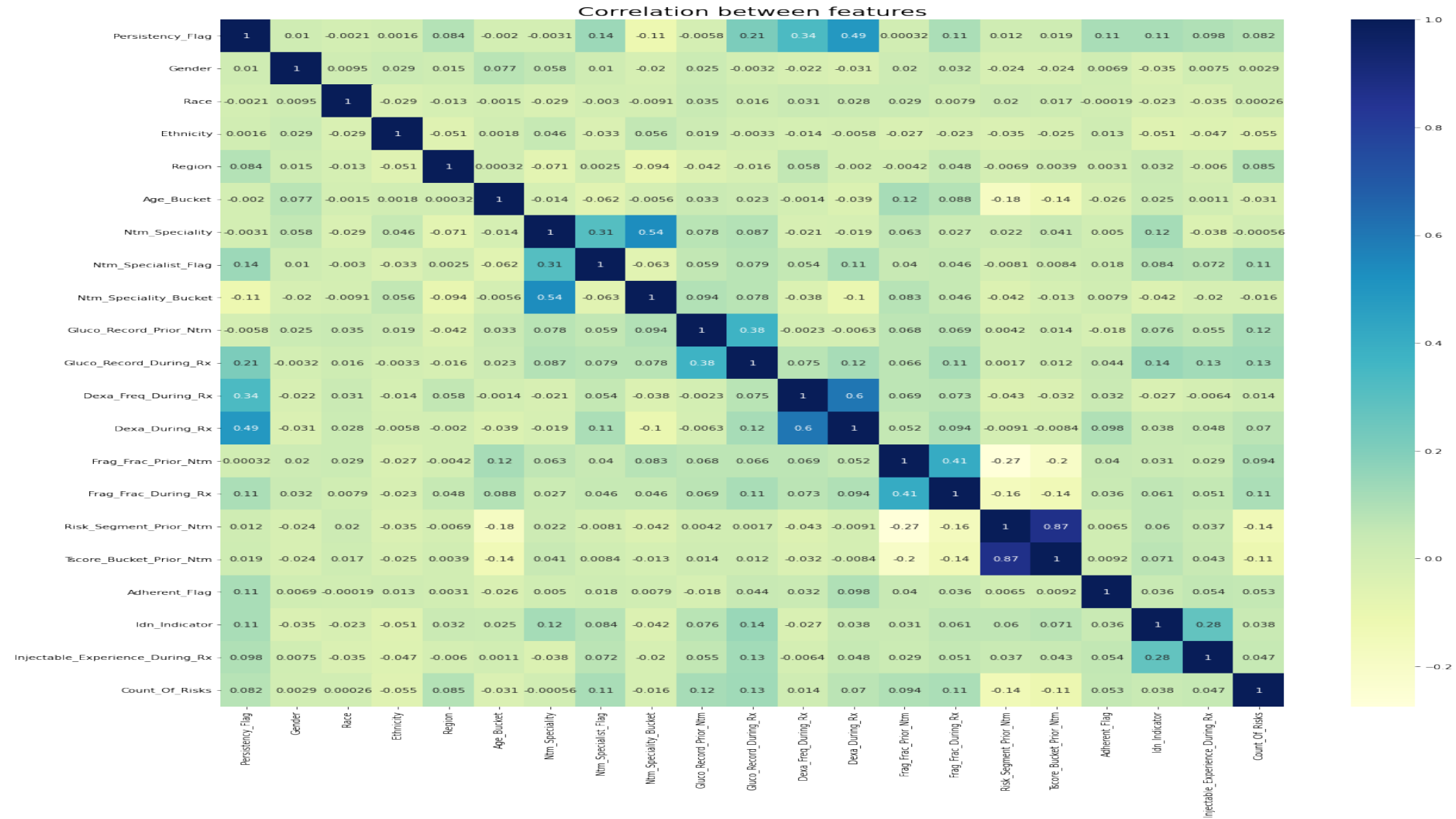
- We need to pre-process our categorical data from words to number to make it easier for the computer to understand. To do this we will use `LabelEncoder()` provided by `sklearn`. Basically, it will transform a categorical column from this (example to describe this approach):

Gender	Race
Male	Asian
Female	Other/Unknown
Male	Caucasian

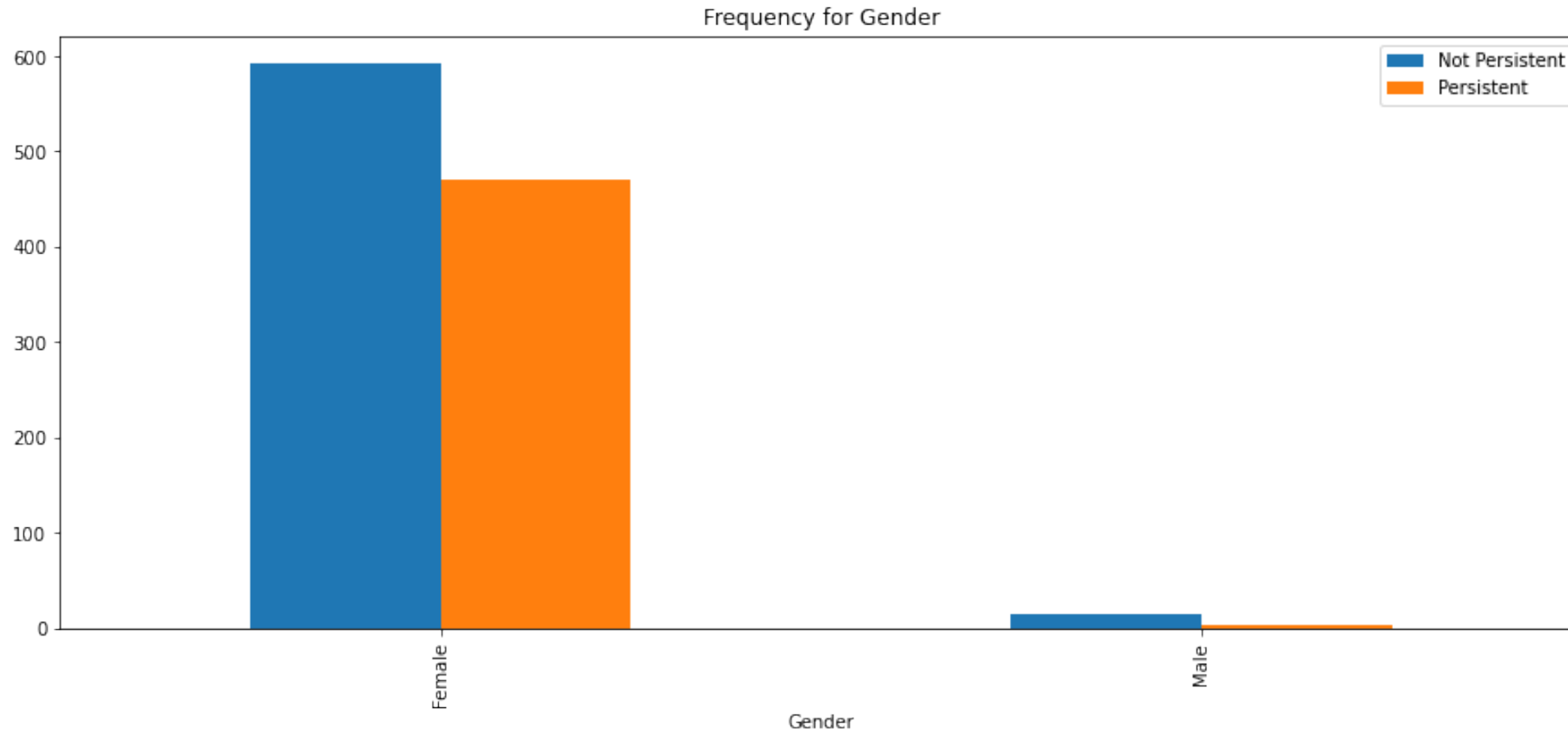
...into something like this... For example, Male will be 1 and Female will be 0

Gender	Race
1	1
0	3
1	2

# EDA (1): Correlation after transformation

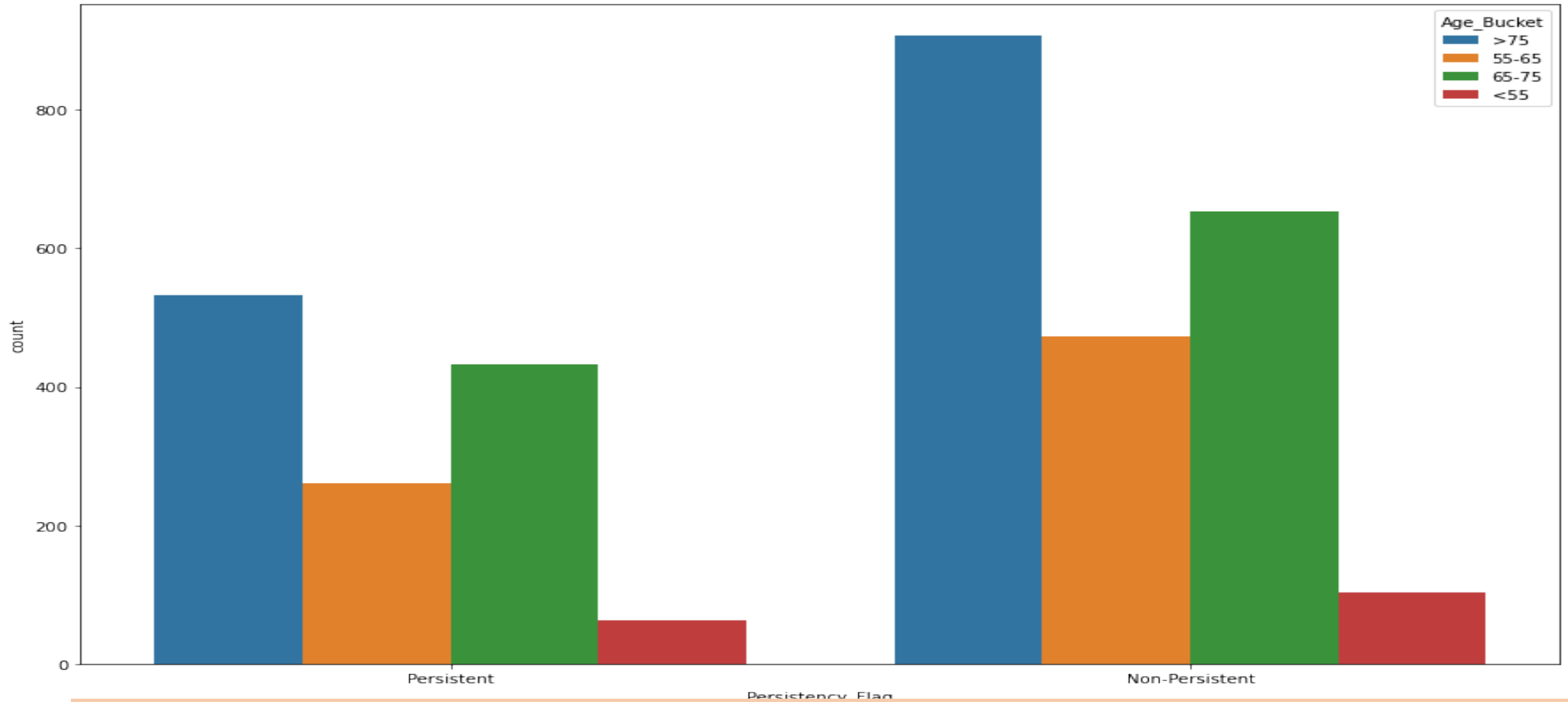


## EDA (2): Gender



**It shows that females are most persistent with drug compliance**

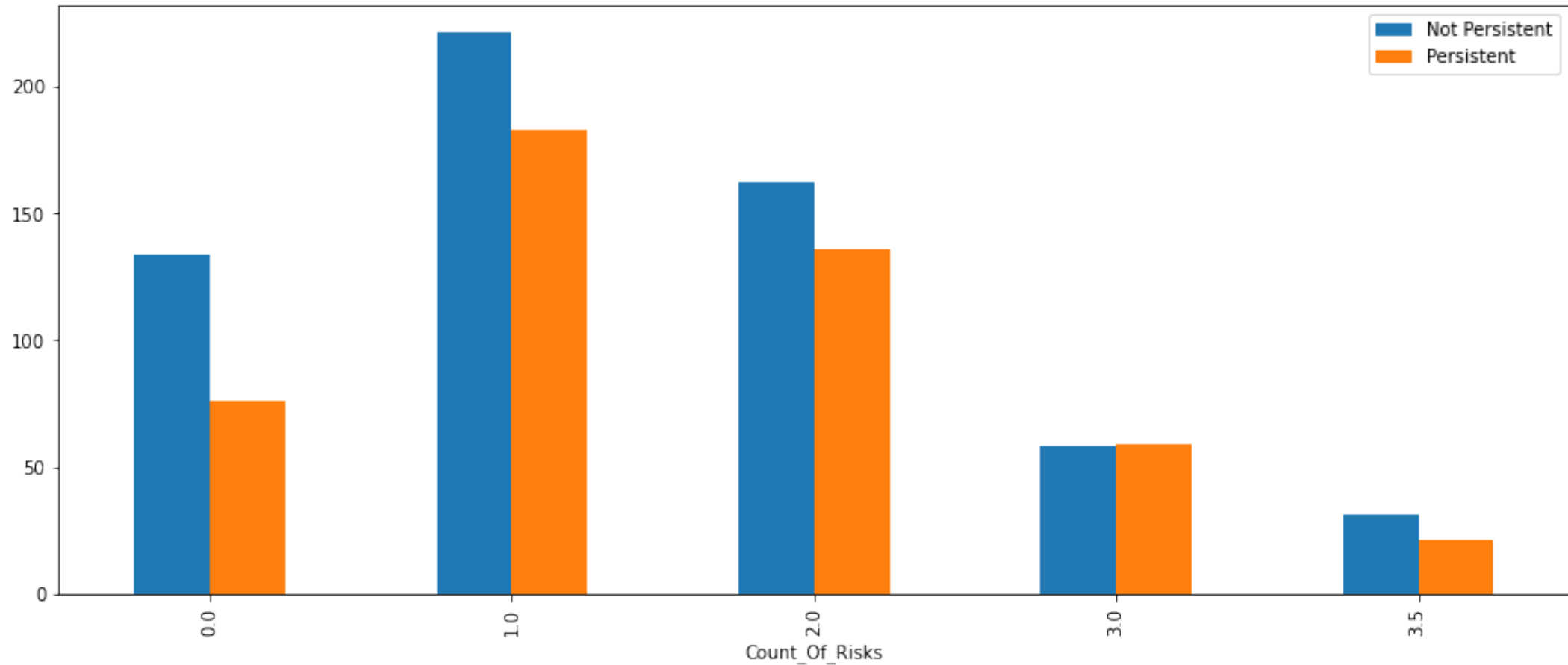
## EDA (3): Age



**Group of people under 55 years old tend to have higher persistent compared to other groups**

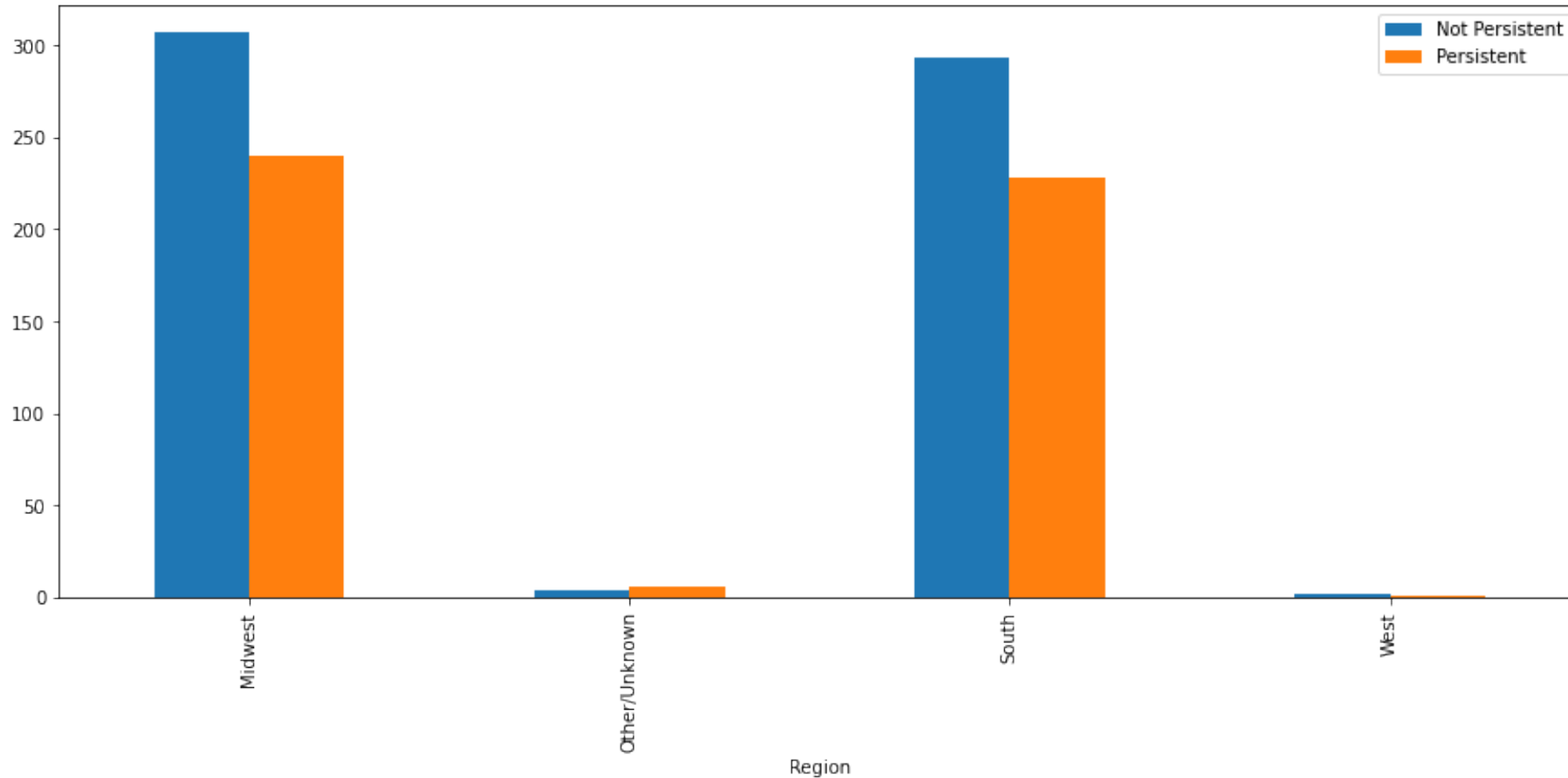


# EDA (4): Number of Risks and Persistent with drugs



**It shows that the number of risk arises when patients are not persistent with their drugs**

# EDA (5): Regions and Persistent with drugs



**South and Midwest Regions have both higher Persistent and Non-Persistent Drugs**

# Model Suggestion

- We will develop four different classification models:
  - Linear Models: Logistic Regression
  - Model for Ensemble: XGBoost Classifier
  - Model for Boosting: AdaBoost Classifier
  - Other models

# Model Creation-Logistic Regression

Accuracy : 0.8093385214007782

Precision : 0.7736389684813754

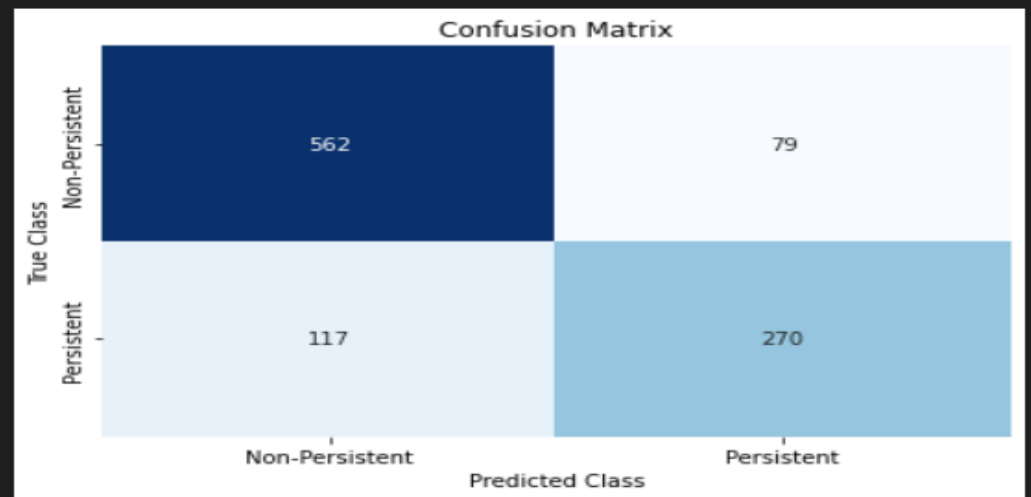
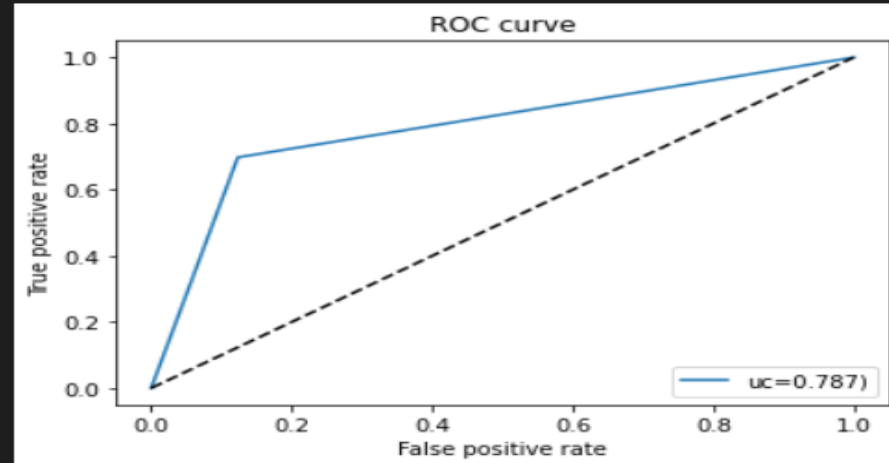
Recall : 0.6976744186046512

F1 Score : 0.733695652173913

	precision	recall	f1-score	support
Non-Persistent	0.83	0.88	0.85	641
Persistent	0.77	0.70	0.73	387
accuracy			0.81	1028
macro avg	0.80	0.79	0.79	1028
weighted avg	0.81	0.81	0.81	1028

AUC : 0.7872147444037296

Logistic Regression Model shows the Accuracy, Recall, Precision ,f1 score and Support of Non-Persistent and Persistence of drugs.

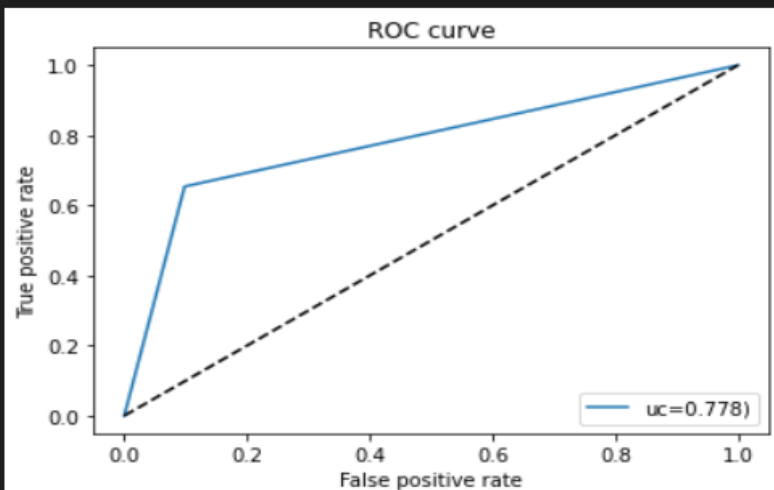


# Model Creation- Ridge Classifier

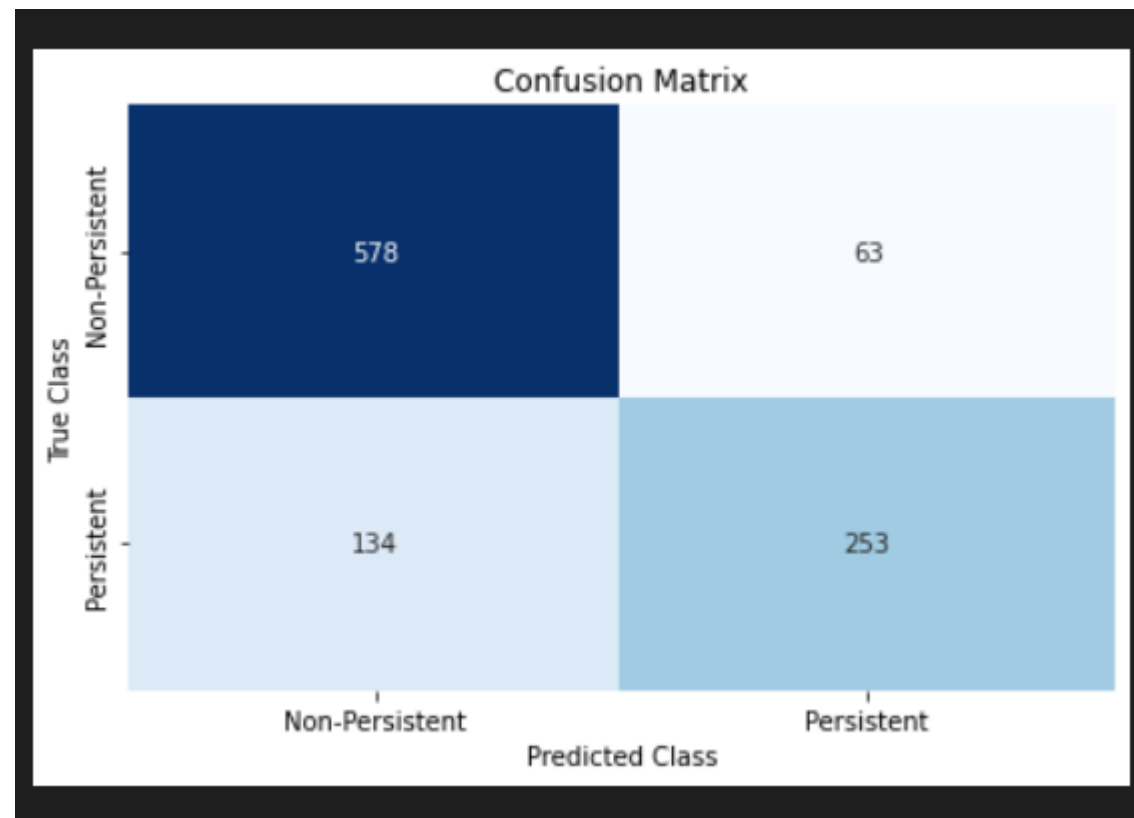
Accuracy : 0.8083657587548638  
Precision : 0.8006329113924051  
Recall : 0.6537467700258398  
F1 Score : 0.7197724039829304

	precision	recall	f1-score	support
Non-Persistent	0.81	0.90	0.85	641
Persistent	0.80	0.65	0.72	387
accuracy			0.81	1028
macro avg	0.81	0.78	0.79	1028
weighted avg	0.81	0.81	0.80	1028

AUC : 0.7777314193342927



Ridge Classifier Model shows the Accuracy, Recall, Precision, f1 score and Support of Non-Persistent and Persistence of drugs.



# Model Creation-SDG Classifier

Accuracy : 0.7957198443579766

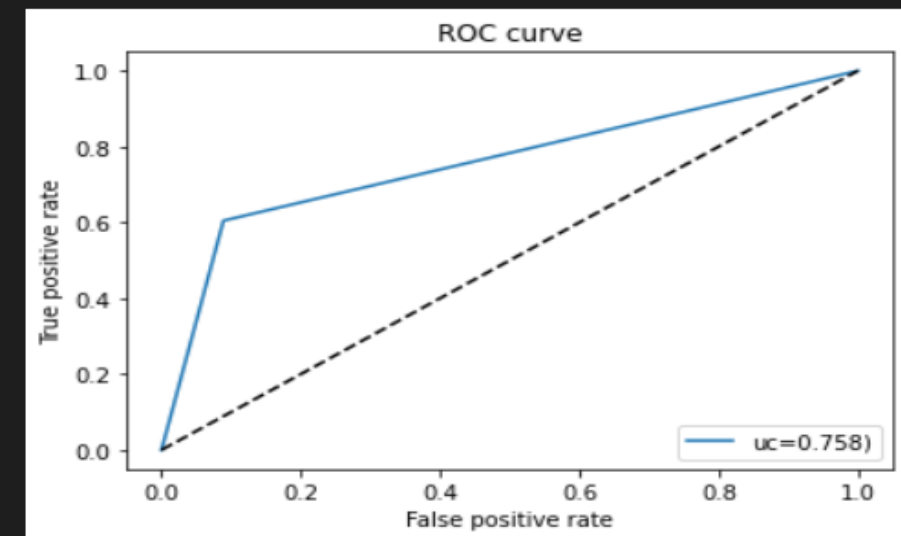
Precision : 0.8041237113402062

Recall : 0.6046511627906976

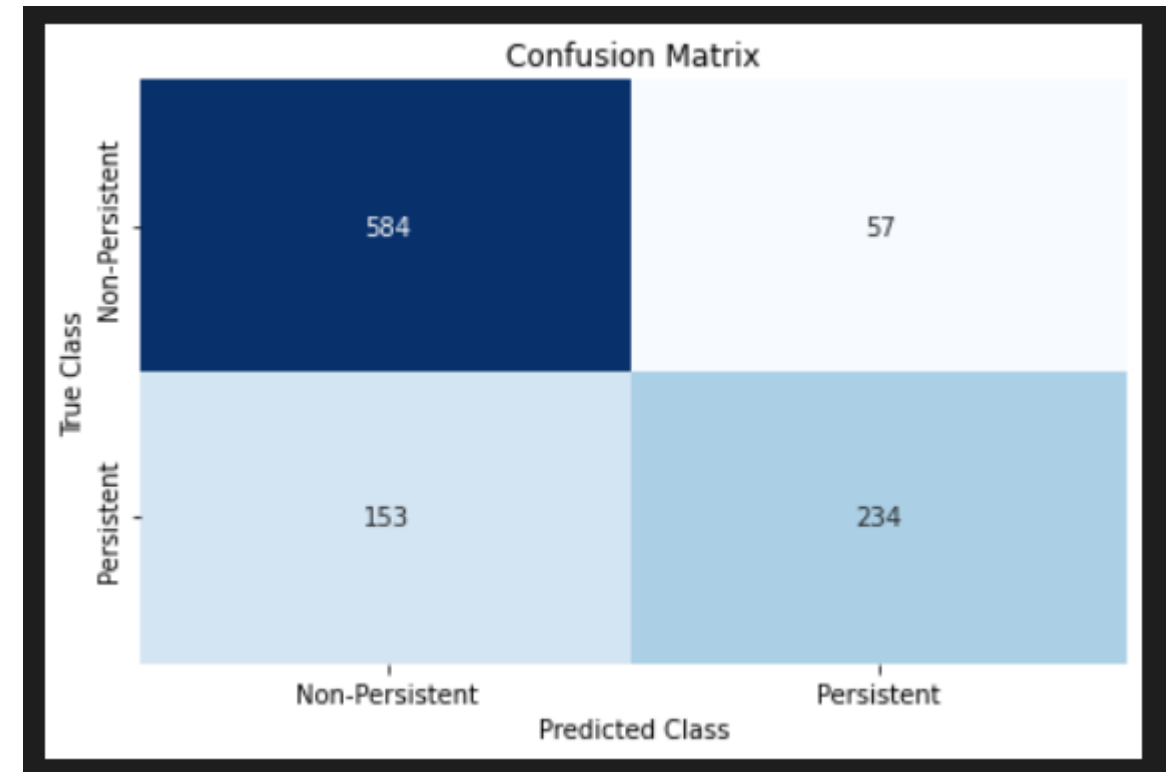
F1 Score : 0.6902654867256637

	precision	recall	f1-score	support
Non-Persistent	0.79	0.91	0.85	641
Persistent	0.80	0.60	0.69	387
accuracy			0.80	1028
macro avg	0.80	0.76	0.77	1028
weighted avg	0.80	0.80	0.79	1028

AUC : 0.75786380292421



SDG Classifier Model shows the Accuracy, Recall, Precision ,f1 score and Support of Non-Persistent and Persistence of drugs.

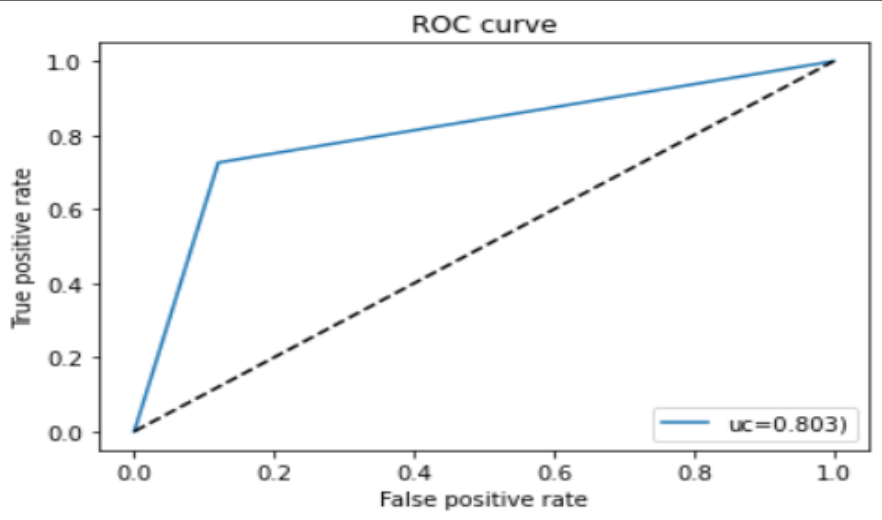


# Random Forest Classifier

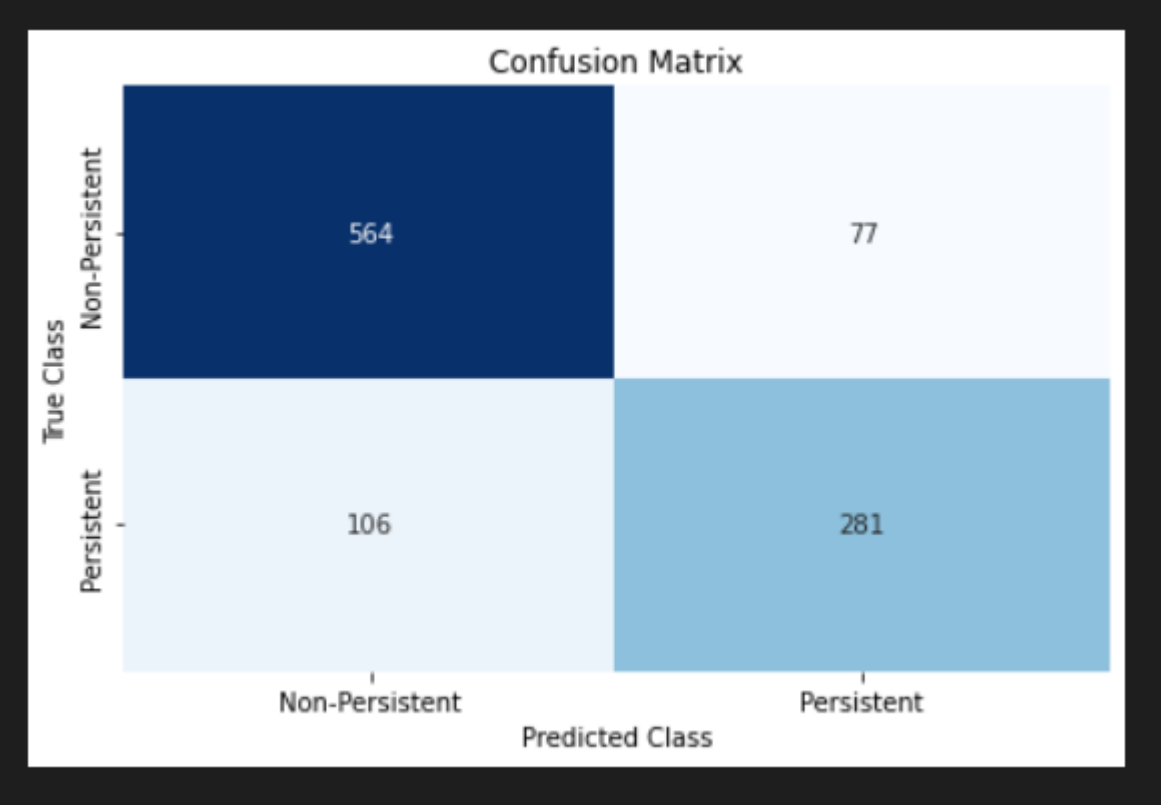
Accuracy : 0.8219844357976653  
Precision : 0.7849162011173184  
Recall : 0.7260981912144703  
F1 Score : 0.7543624161073826

	precision	recall	f1-score	support
Non-Persistent	0.84	0.88	0.86	641
Persistent	0.78	0.73	0.75	387
accuracy			0.82	1028
macro avg	0.81	0.80	0.81	1028
weighted avg	0.82	0.82	0.82	1028

AUC : 0.8029866931111354



Random Forest Classifier Model shows the Accuracy, Recall, Precision ,f1 score and Support of Non-Persistent and Persistence of drugs.



# Ada Boost Classifier

Accuracy : 0.8171206225680934

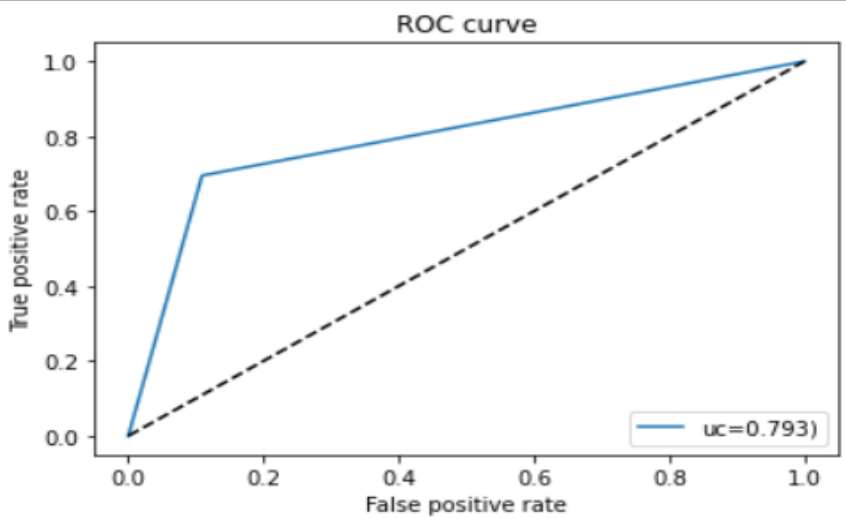
Precision : 0.7935103244837758

Recall : 0.6950904392764858

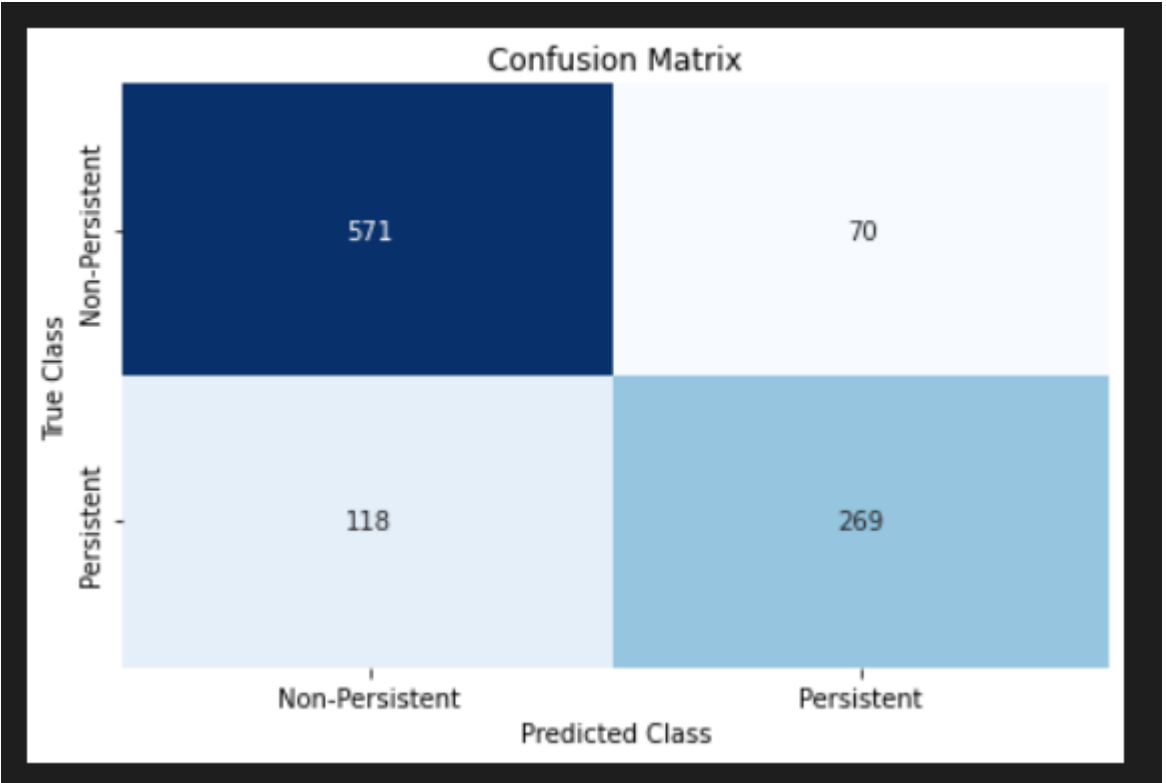
F1 Score : 0.7410468319559228

	precision	recall	f1-score	support
Non-Persistent	0.83	0.89	0.86	641
Persistent	0.79	0.70	0.74	387
accuracy			0.82	1028
macro avg	0.81	0.79	0.80	1028
weighted avg	0.82	0.82	0.81	1028

AUC : 0.7929430355508794



Ada boost classifier shows the Accuracy, Recall, Precision ,f1 score and Support of Non-Persistent and Persistence of drugs.





# Stacking Classifier

Accuracy : 0.8103112840466926

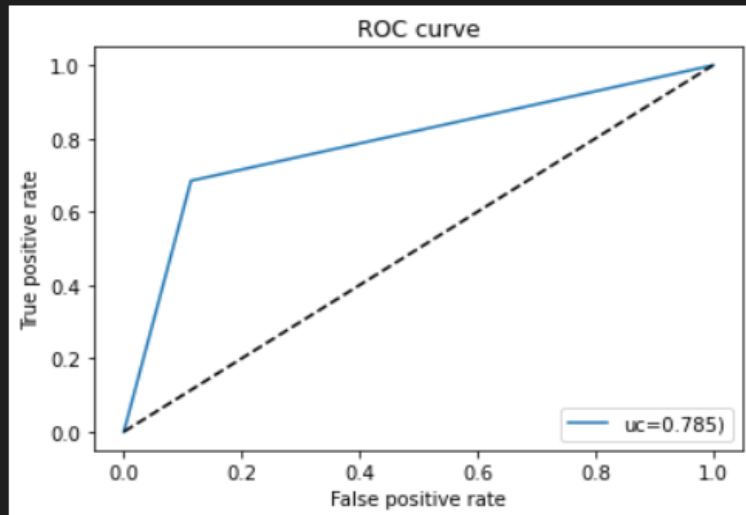
Precision : 0.7840236686390533

Recall : 0.6847545219638242

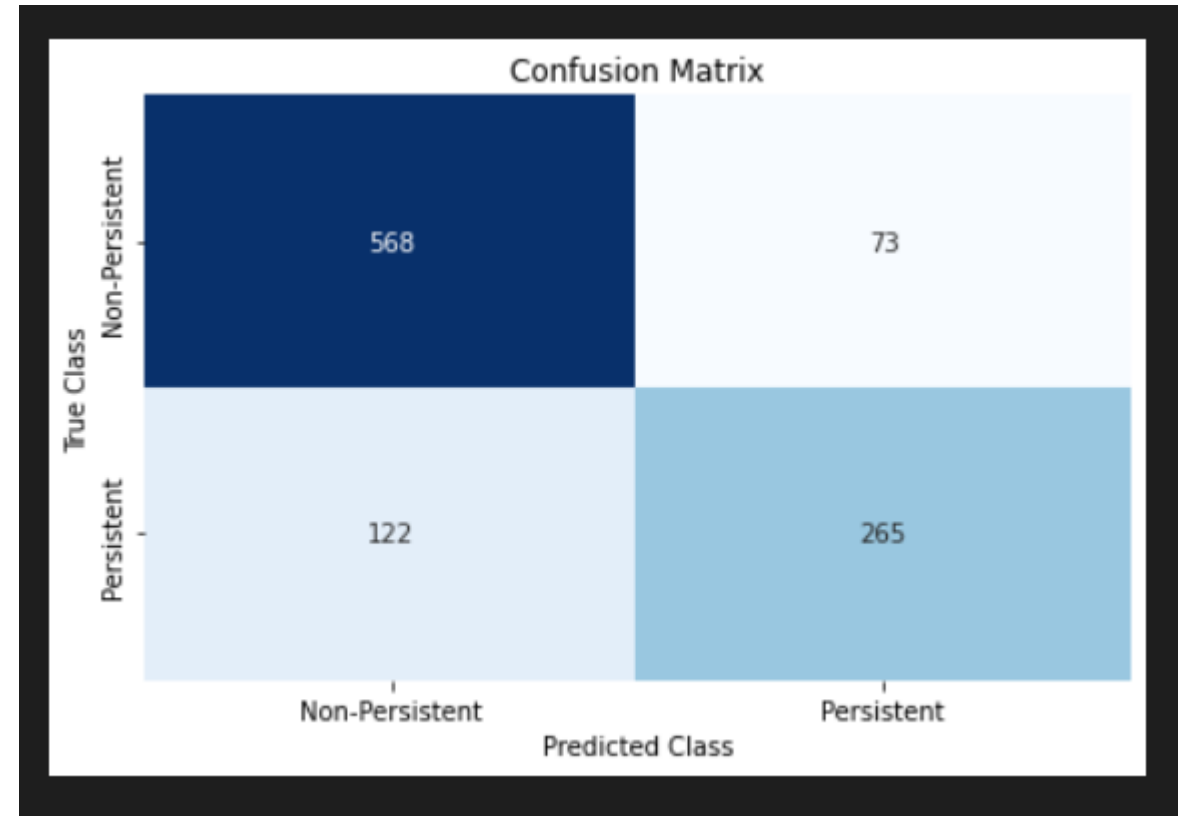
F1 Score : 0.7310344827586208

	precision	recall	f1-score	support
Non-Persistent	0.82	0.89	0.85	641
Persistent	0.78	0.68	0.73	387
accuracy			0.81	1028
macro avg	0.80	0.79	0.79	1028
weighted avg	0.81	0.81	0.81	1028

AUC : 0.7854349832908045



Stacking Classifier Model shows the Accuracy, Recall, Precision, f1 score and Support of Non-Persistent and Persistence of drugs.

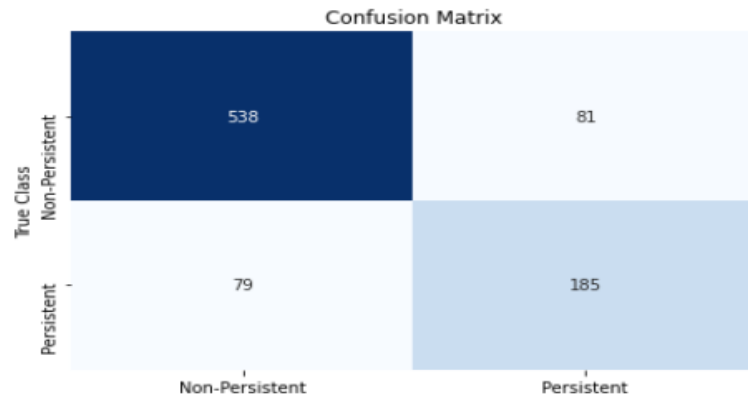
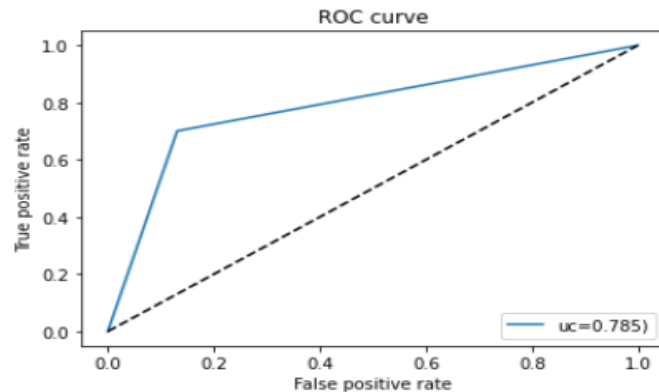


# XG Boost Classifier

Accuracy : 0.8187995469988675  
Precision : 0.6954887218045113  
Recall : 0.7007575757575758  
F1 Score : 0.6981132075471698


	precision	recall	f1-score	support
Non-Persistent	0.87	0.87	0.87	619
Persistent	0.70	0.70	0.70	264
accuracy			0.82	883
macro avg	0.78	0.78	0.78	883
weighted avg	0.82	0.82	0.82	883

AUC : 0.7849506780241836



XG Boost Classifier Model shows the Accuracy, Recall, Precision, f1 score and Support of Non-Persistent and Persistence of drugs.

# Conclusion

- Approximately all the classifiers have same result, but three of them are the bests :
- Logistic Classifier (Linear) with Accuracy 81%, Recall 70%, 73% F1-Score, and 78% AUC
- AdaBoost Classifier (Ensemble/Boosting) with Accuracy 81%, Recall 69%, 74% F1-Score, and 79% AUC
- XGBoost Classifier (Ensemble/Boosting) with Accuracy 82%, Recall 73%, 75% F1-Score, and 80% AUC  Higher accuracy