**Name: Shweta Lal**

**UCID: srl29**

**DATA MINING FINAL PROJECT**

**Dataset used:** CDC Diabetes Health Indicators

<https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators>

<https://github.com/uci-ml-repo/ucimlrepo>

**Import to Python:**

pip install ucimlrepo

from ucimlrepo import fetch\_ucirepo

# fetch dataset

cdc\_diabetes\_health\_indicators = fetch\_ucirepo(id=891)

# data (as pandas dataframes)

X = cdc\_diabetes\_health\_indicators.data.features

y = cdc\_diabetes\_health\_indicators.data.targets

# metadata

print(cdc\_diabetes\_health\_indicators.metadata)

# variable information

print(cdc\_diabetes\_health\_indicators.variables)

**Classification Algorithms Used**:

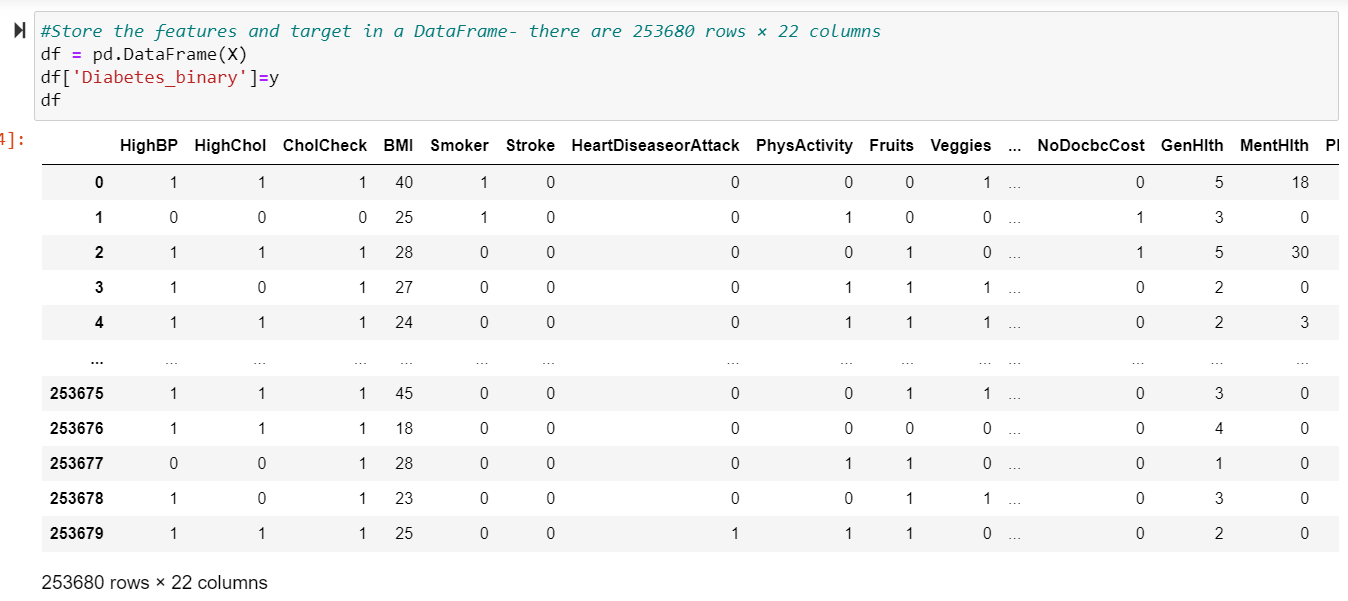
1. Random Forest
2. Support Vector Machine (SVM)
3. Long Short-Term Memory (LSTM)

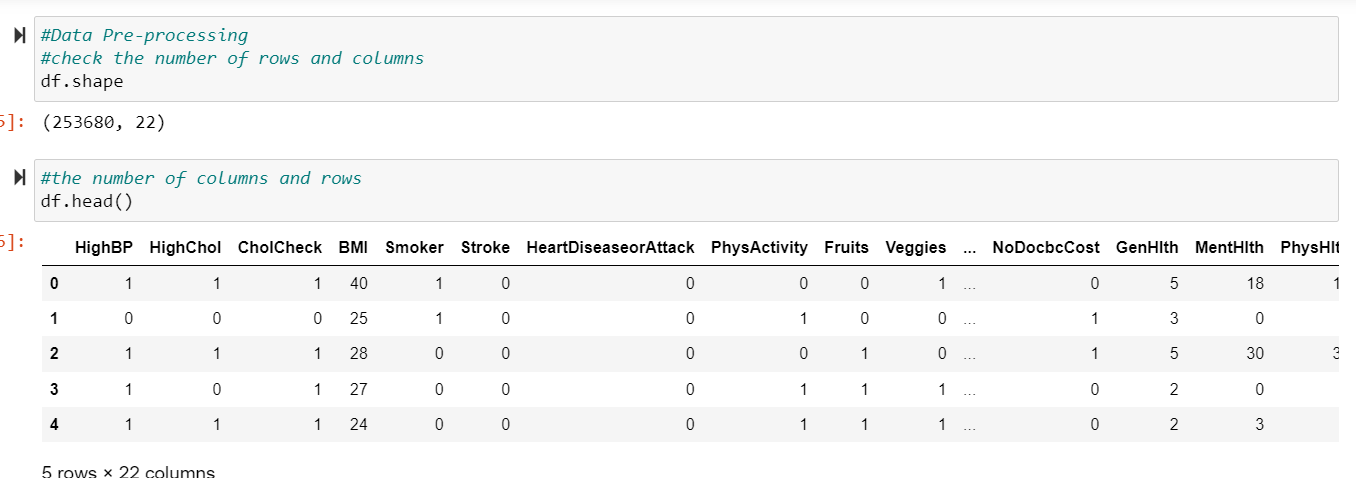
**Implementation of Binary Classification Algorithm using Python:**

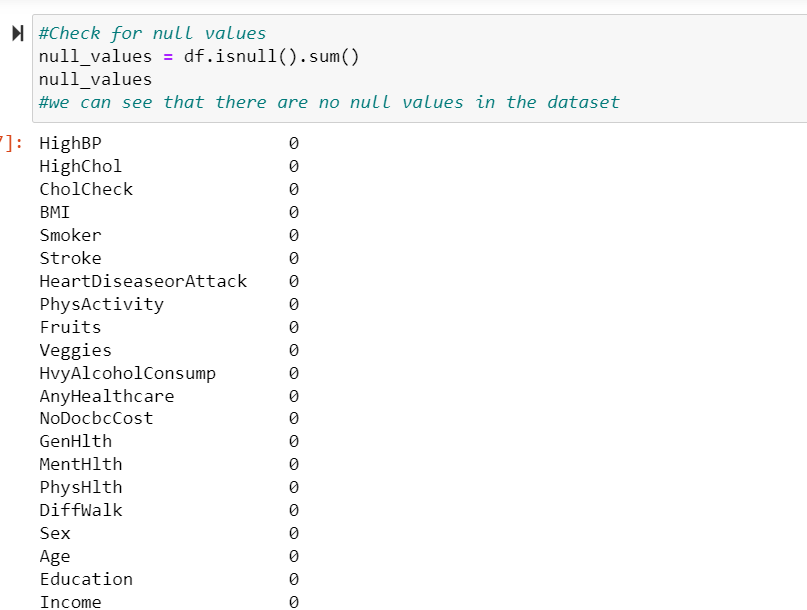
1. Loading the dataset:

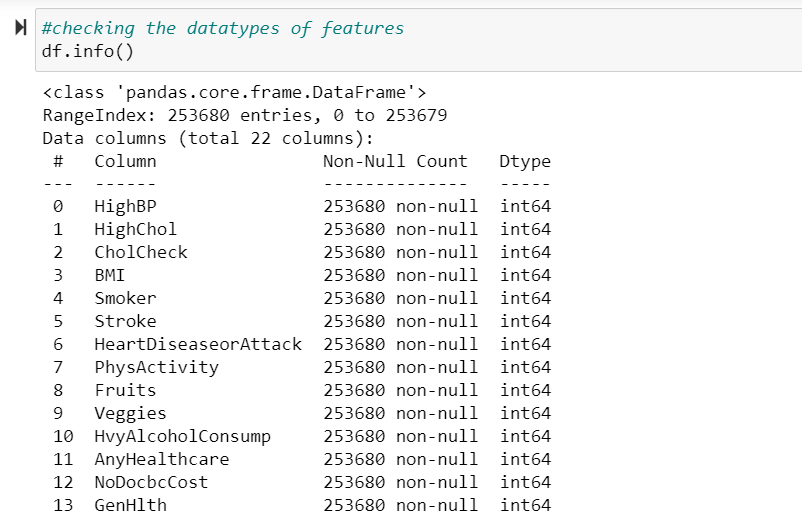


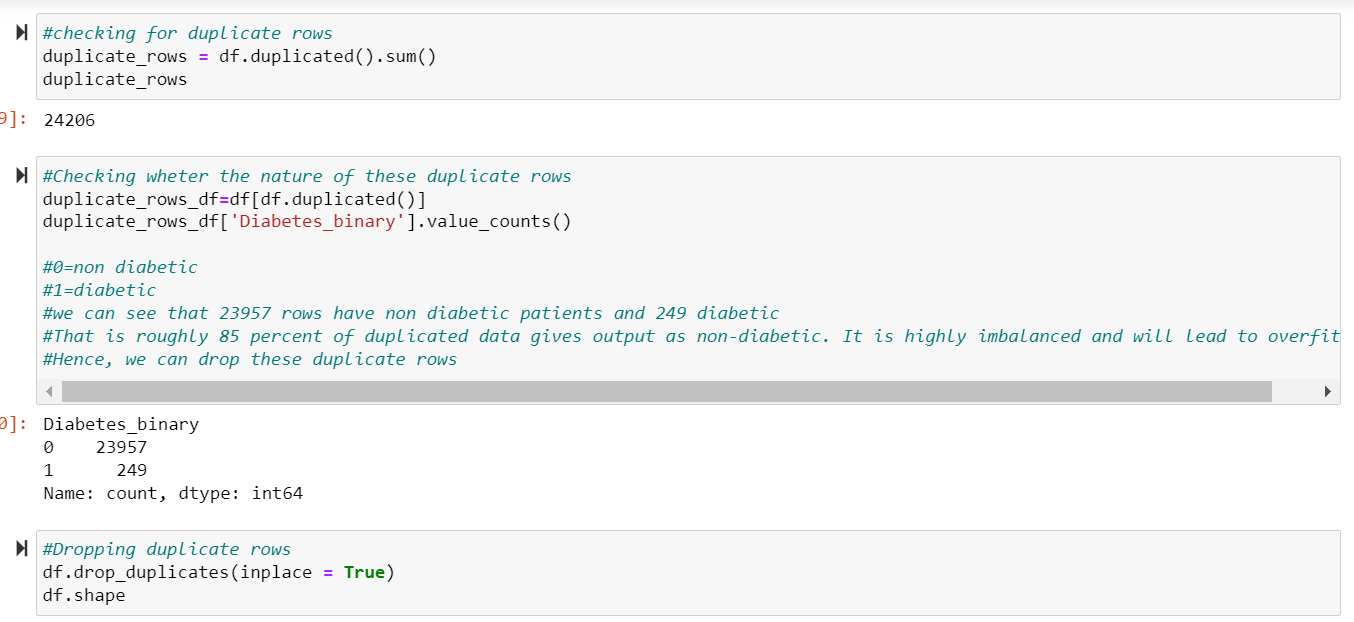
1. Data Pre-processing





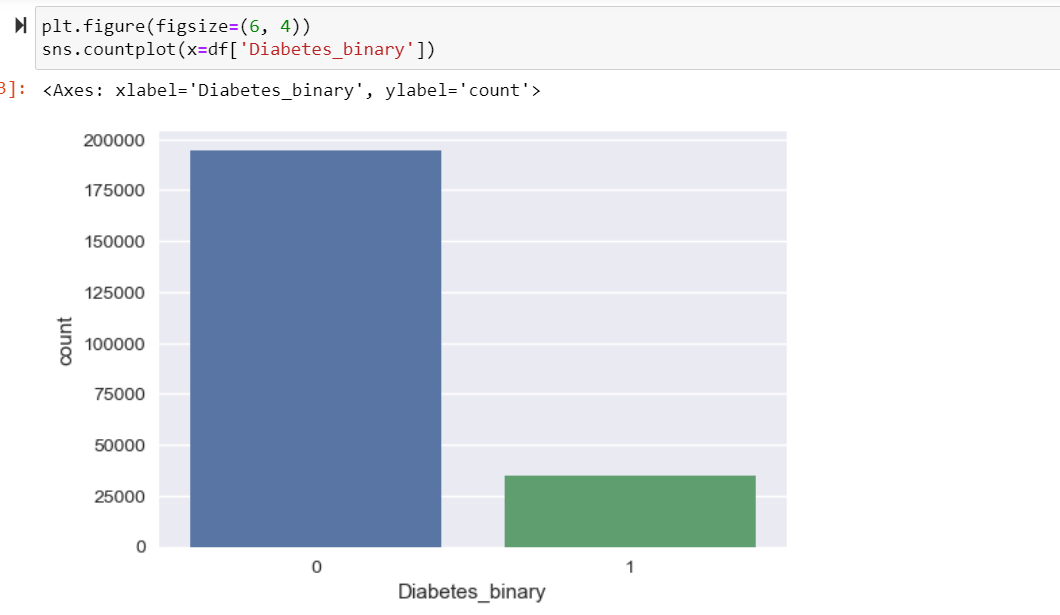






1. Exploratory Data Analysis







From above charts, following conclusions can be made:

* + Comparing distributions btwn non-diabetic(0) and diabetic(1), features like: 'HighBP', 'HighChol', 'CholCheck', 'Smoker', 'Stroke', 'HeartDiseaseorAttack', 'AnyHealthcare', 'DiffWalk' have large proportion of diabetic patients- helpful to determine if a patient has diabetes.
  + Features like: 'PhysActivity', 'Fruits', 'Veggies' have more non-diabetic patients- helpful in determing if a patient is non-diabetic



From above charts, following conclusions can be made:

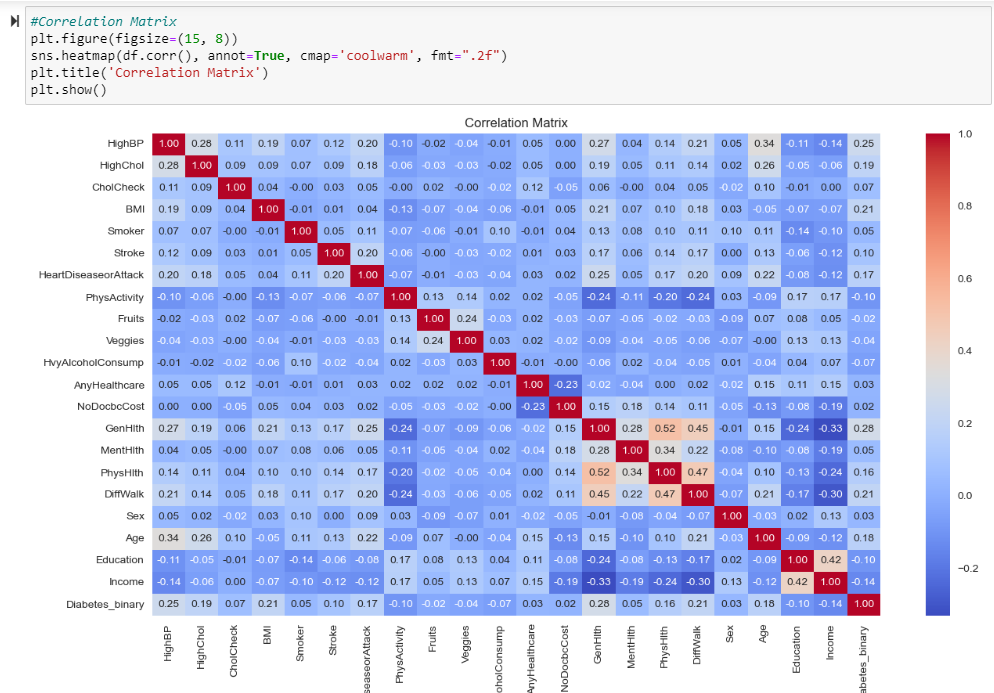
* + Diabetic patients have a high score GenHealth, which indicates poor general health (1=Excellent, ..,5=poor)
  + Age range(1-13): 1=18-24 yrs,..,9=60-64 yrs, 13= 80 yrs or older diabetic patients tender to be older
  + Education - middle range of people are diabetic, that means less educated
  + Income - as compared to non-diabetics, lower income range people are diabeticthese features are helpful to determine a diabetic patient



From above charts, following conclusions can be made:

* most of the box plots have outliers
* BMI- mostly between 20-40
* MentalHealth, PhysicalHealth- user select 0 when they do not have any issues with physical or mental health- hence the outliers are necessary
* MentalHealth- between 0-5
* PhysicalHealth- between 0-5

1. Finding Correlation between attributes



Observations from heatmap:

* There are no highly positively related features
* There are few which are moderately related, but eventually they are strongly associated with target variables, hence we can keep all these variables.

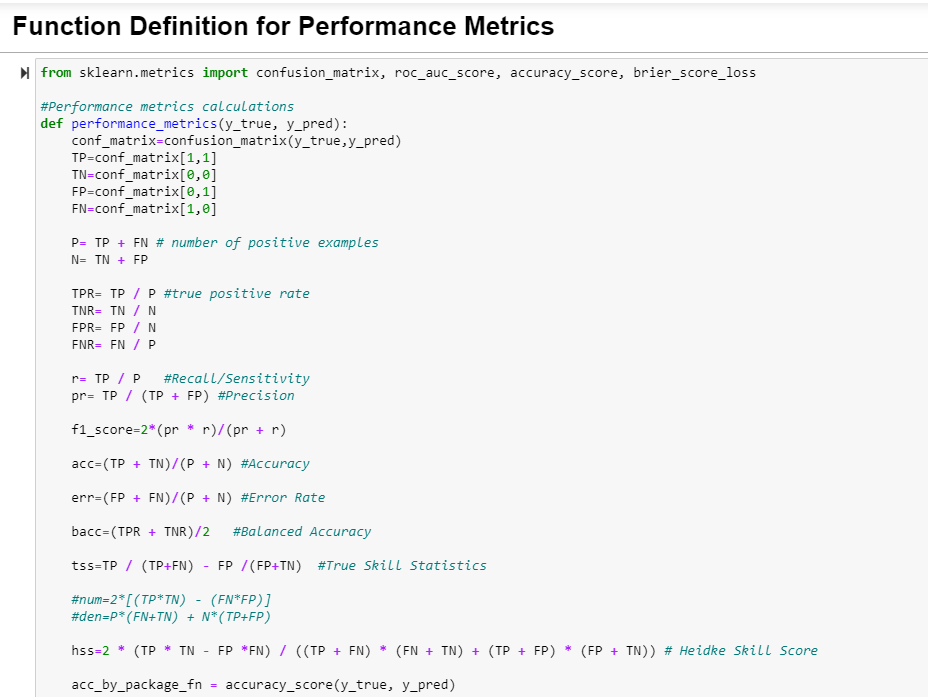
1. Splitting dataset into training and testing set



1. Hyperparameter Tuning

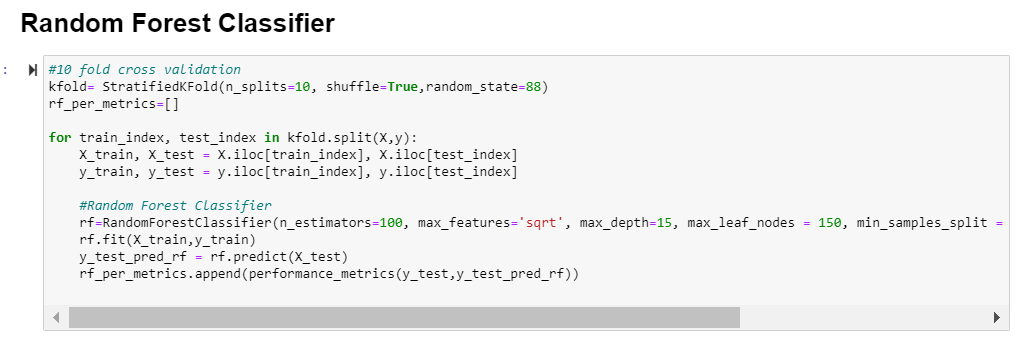


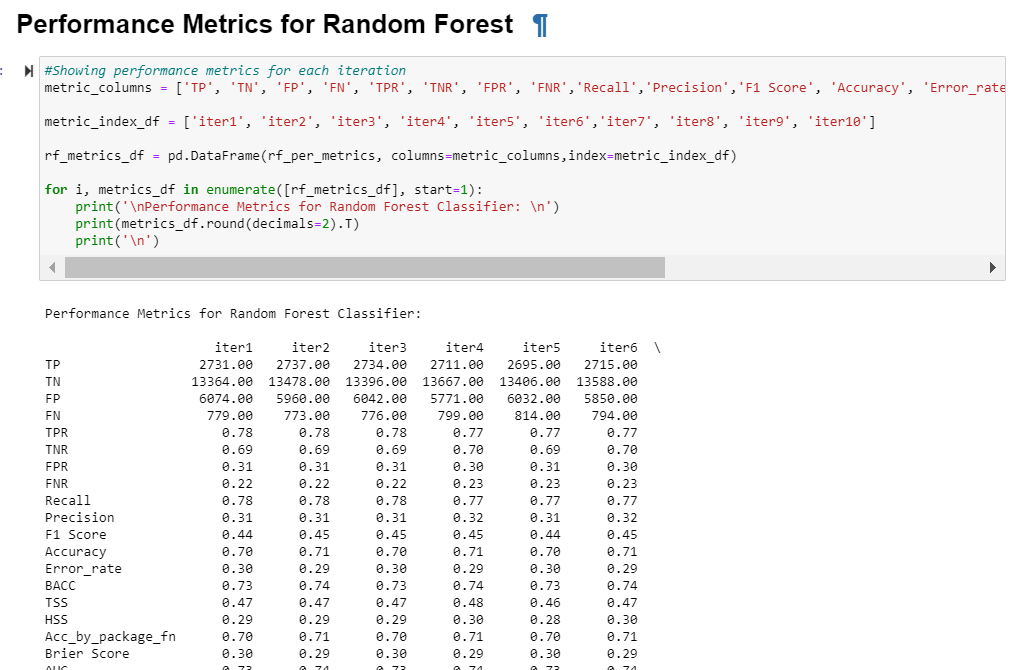
1. Calculating Performance Metrics:

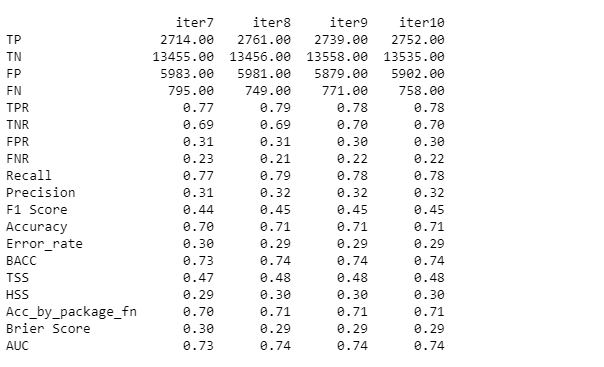


1. Random Forest Classifier

Implemented Random Forest using the 10- fold cross validation using Stratified function.



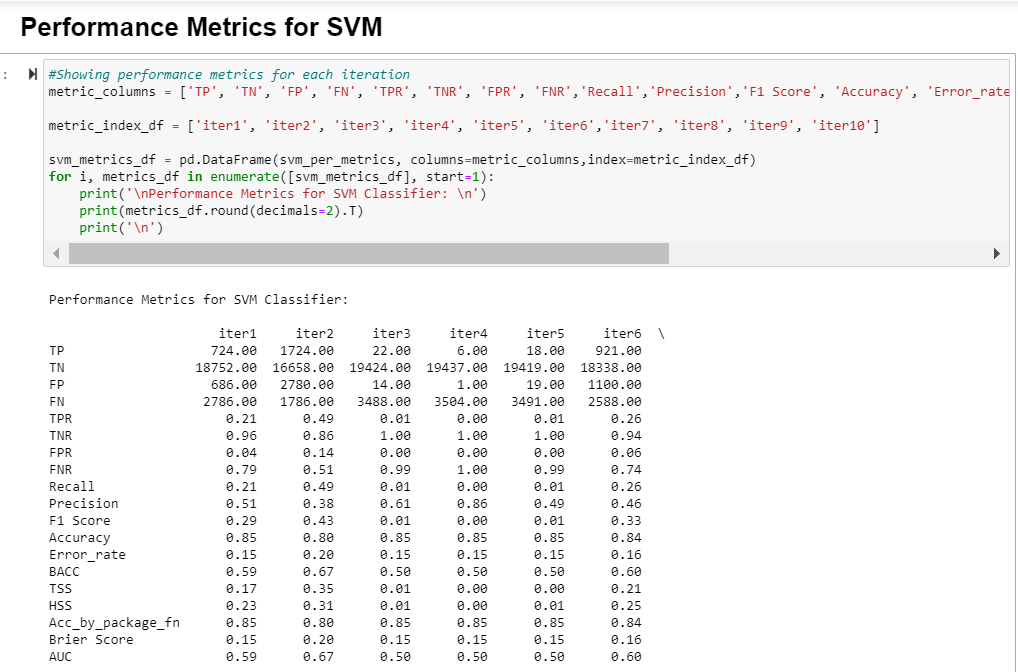


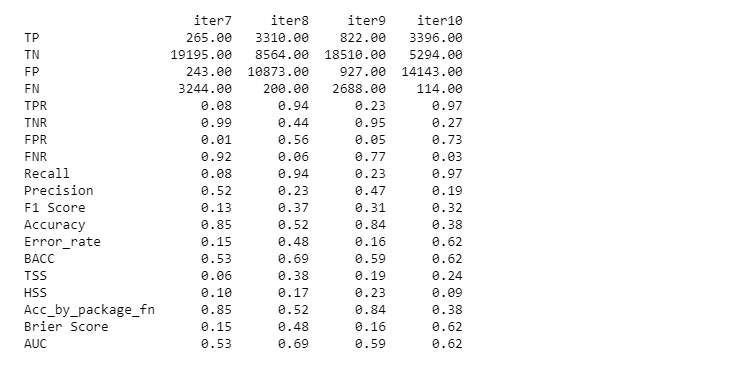


1. SVM:

Implemented SVM using the 10- fold cross validation using Stratified function.



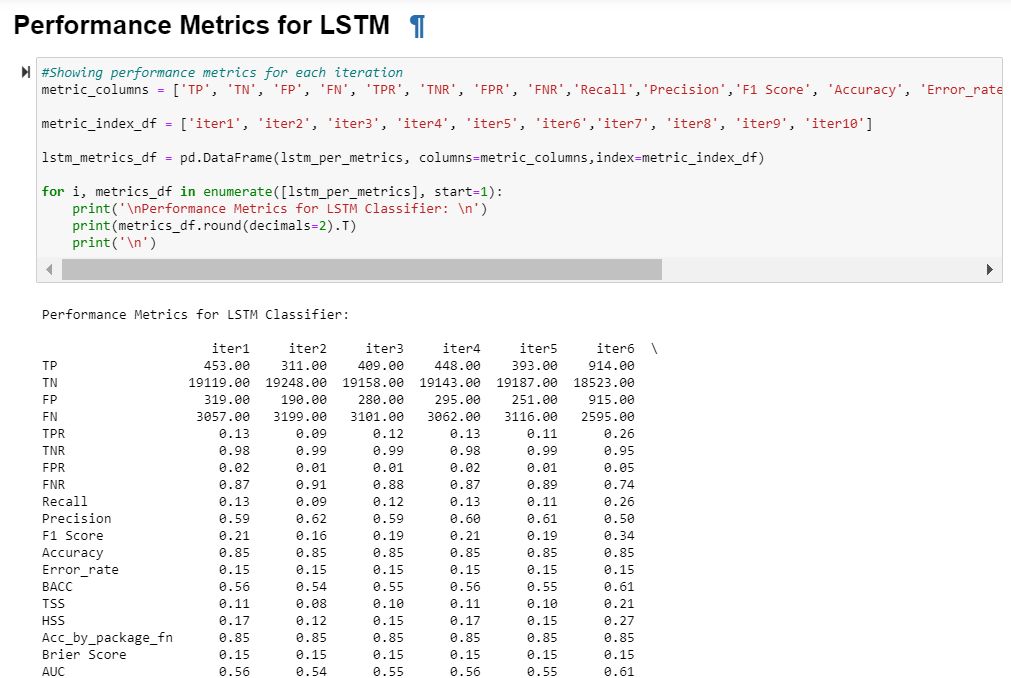


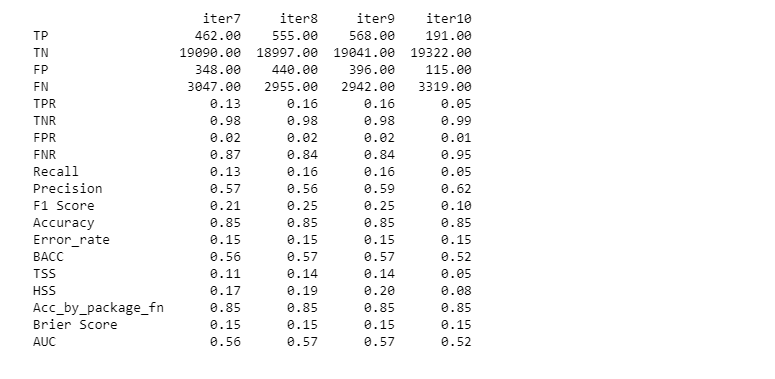


1. LSTM:

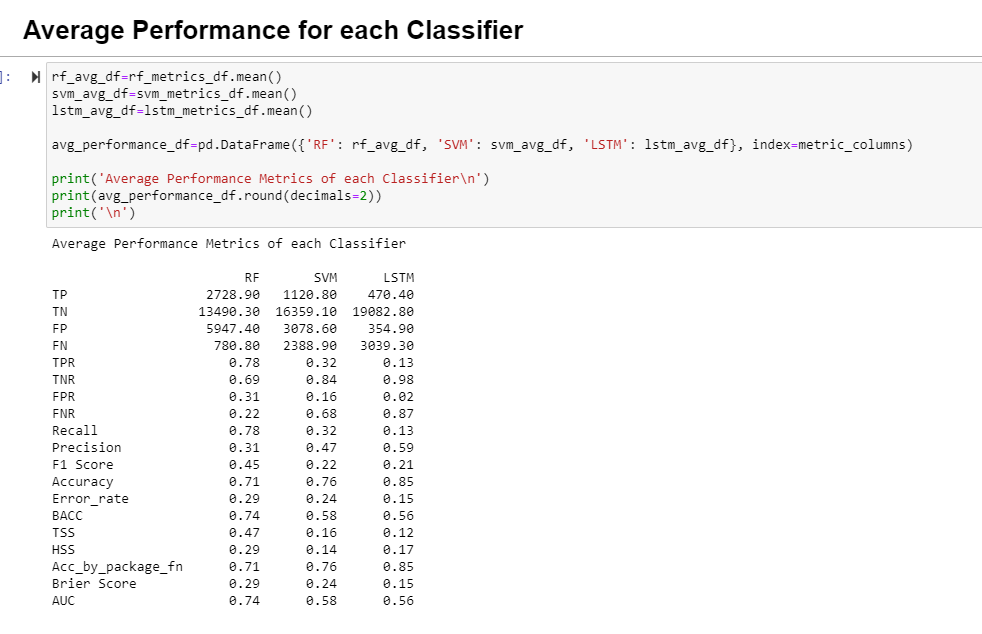
Implemented LSTM using the 10- fold cross validation using Stratified function.







1. Compare average of all classifiers



1. Plotting ROC Curve



