Diabetes Prediction

CSE454

Data Mining Project

What is Diabetes?

Diabetes is a serious and growing health issue worldwide. It is a chronic condition that occurs when the body is unable to properly use and store glucose (a type of sugar). This can lead to a host of complications, including heart disease, kidney failure, and amputations. Early detection and management of diabetes is crucial for preventing these complications. In this project, we aim to predict diabetes in patients using a dataset of demographic and medical information.

What's our project's aim?

Project Overview

The Goal of Project

The goal of this project is to use machine learning to predict diabetes in patients. The dataset used for this project includes information on patient demographics (age, gender, etc.) and medical history (blood pressure, BMI, etc.). We will be using this information to train a model that can accurately predict whether a patient has diabetes or not.

Stages of Project

- Data Preprocessing and Cleaning
- Data Analysing
- Visualization
- Model Training
- Results

Data Preprocessing

Saving CSV into DataFrame

To work on data efficiently using the quite helpful libraries such as pandas, numpy, seaborn and matplotlib.

Dealing with Missing Values

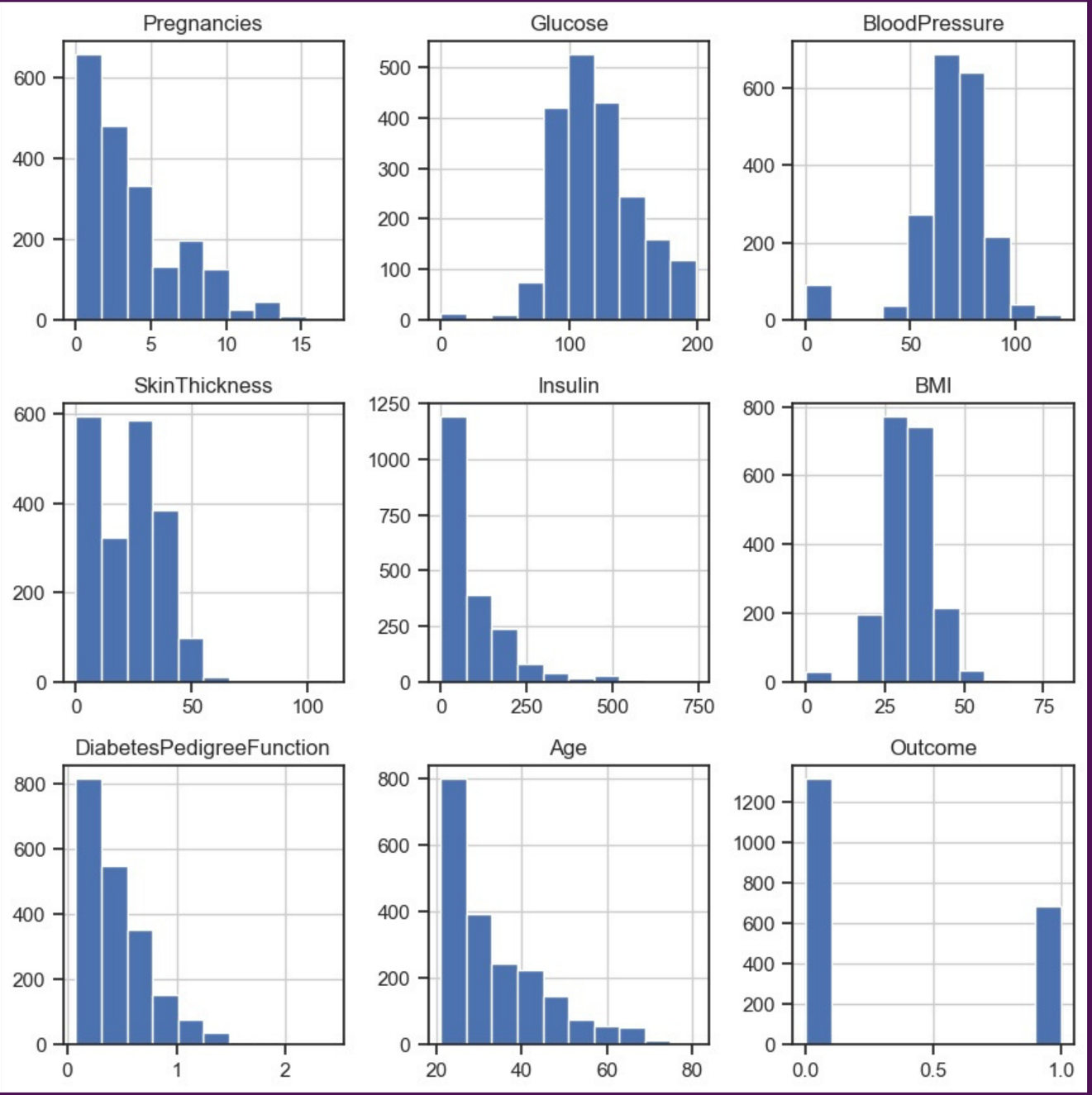
Thankfully, it observed that there are not a null-value in any row.

Data Analysing

Understanding data better

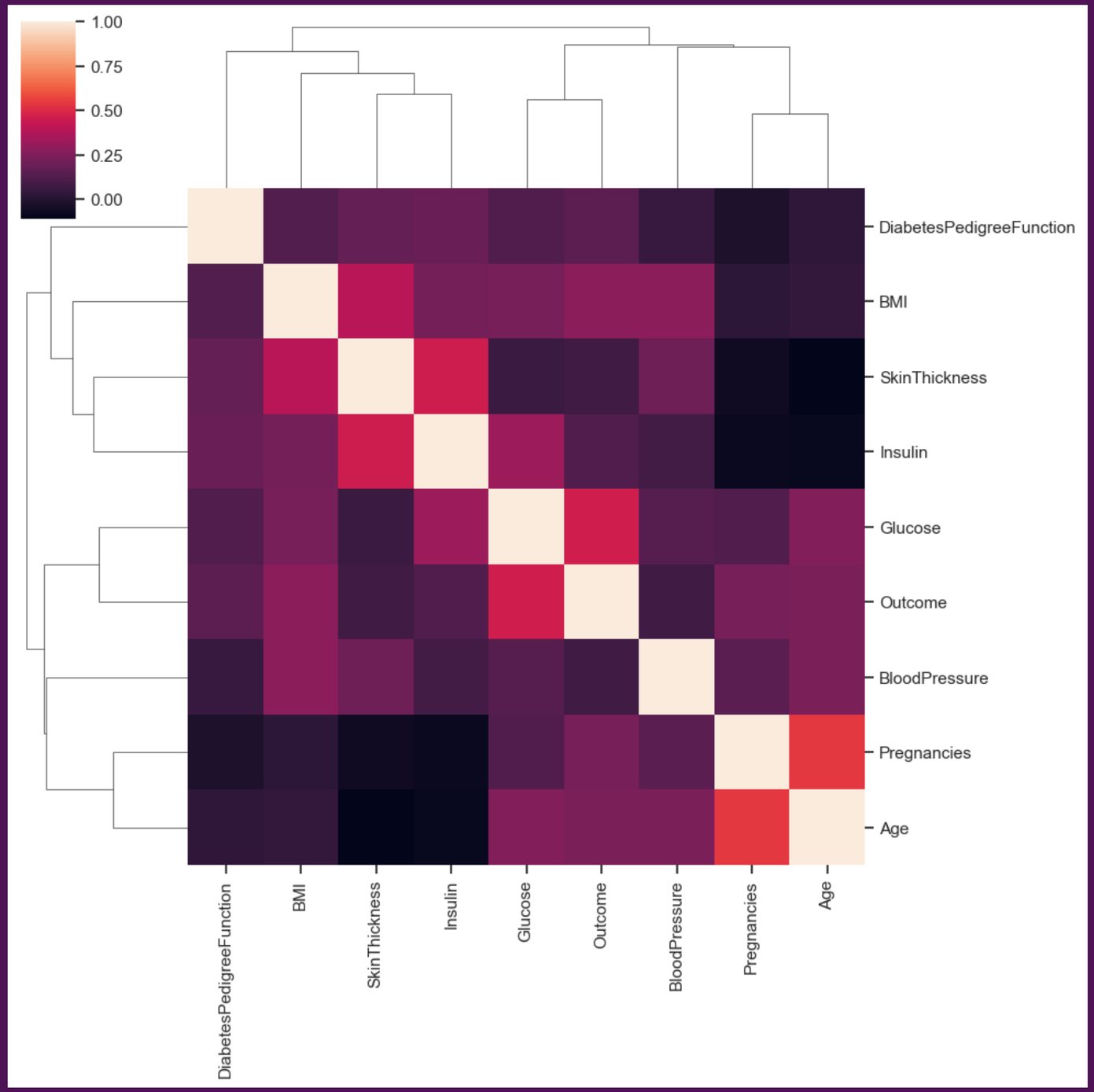
Histogram Tables

Using histogram tables to examine the distribution of data.

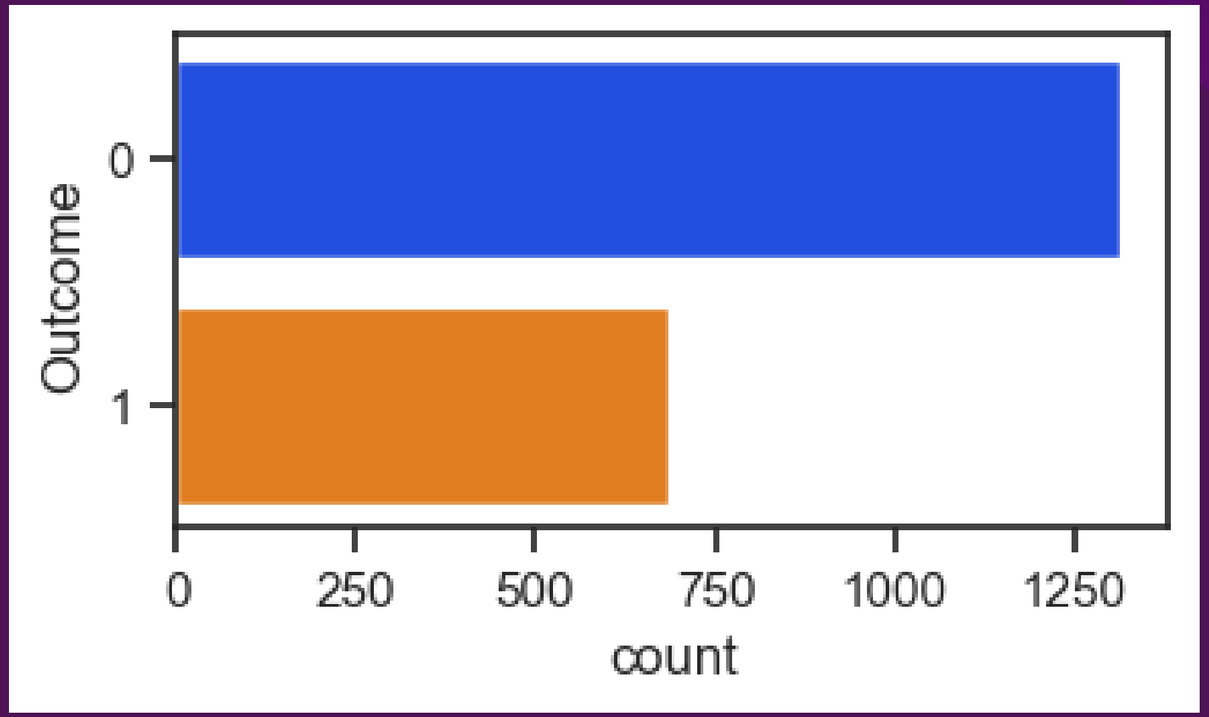

Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age
Outcome

Clustermapping

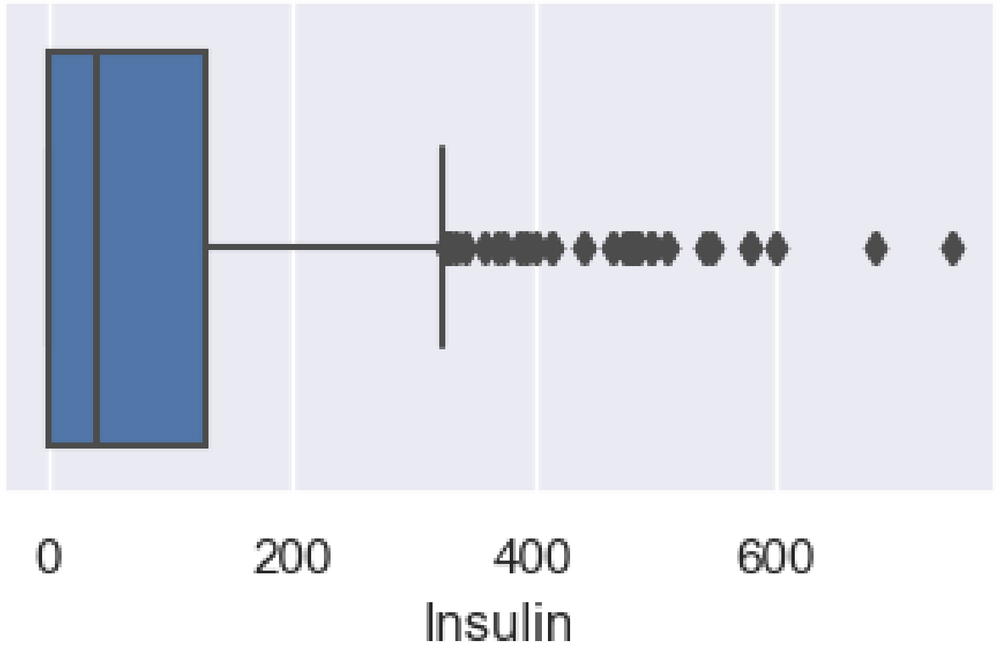
To see correlations

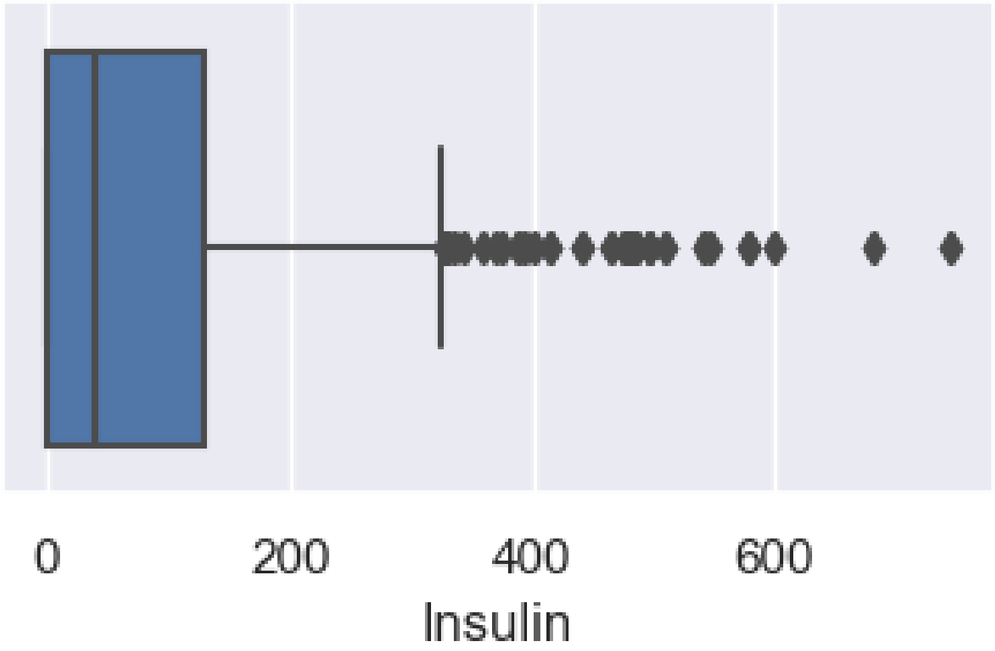

1.00
0.75
0.50
0.25
0.00
DiabetesPedigreeFunction
BMI
SkinThickness
Insulin
Glucose
Outcome
BloodPressure
Pregnancies
Age

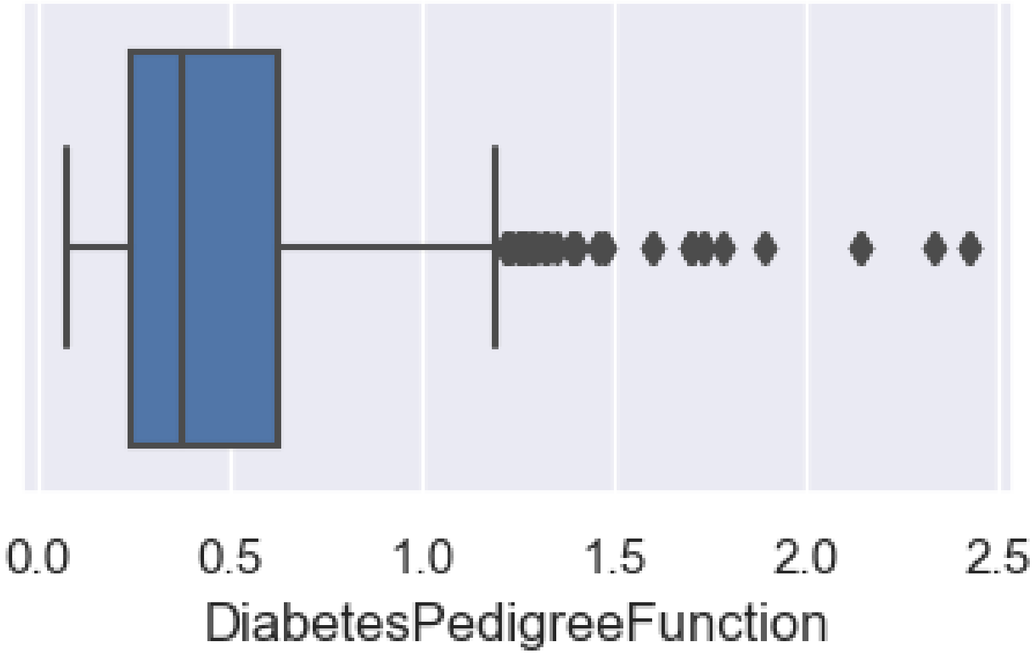
Countplotting


Outcome
0
1
0
250
500
750
1000
1250
count

Boxplots to see outliers


0
200
400
600
Insulin


0
200
400
600
Insulin


0.0
0.5
1.0
1.5
2.0
2.5
DiabetesPedigreeFunction

Removing Outliers

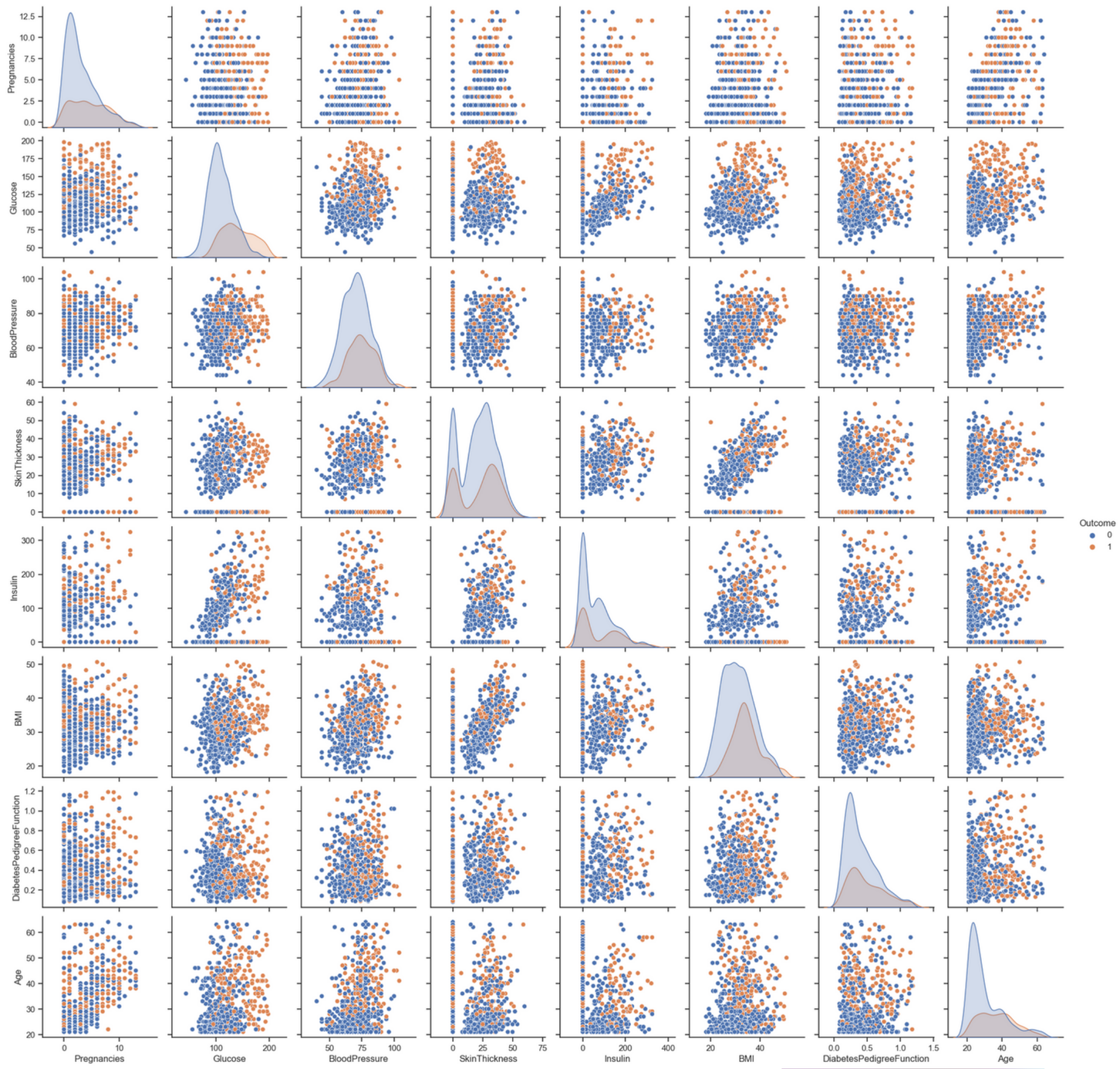
Looking the data on scatter matrix

```
# Outlier removing

Q1=df.quantile(0.25)
Q3=df.quantile(0.75)

IQR=Q3-Q1

# The IQR describes the middle 50% of values when ordered from lowest to highest.
# To find the interquartile range (IQR), first find the median (middle value)
# of the lower and upper half of the data.
# These values are quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference
# between Q3 and Q1.
```


Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age
Outcome
0
1

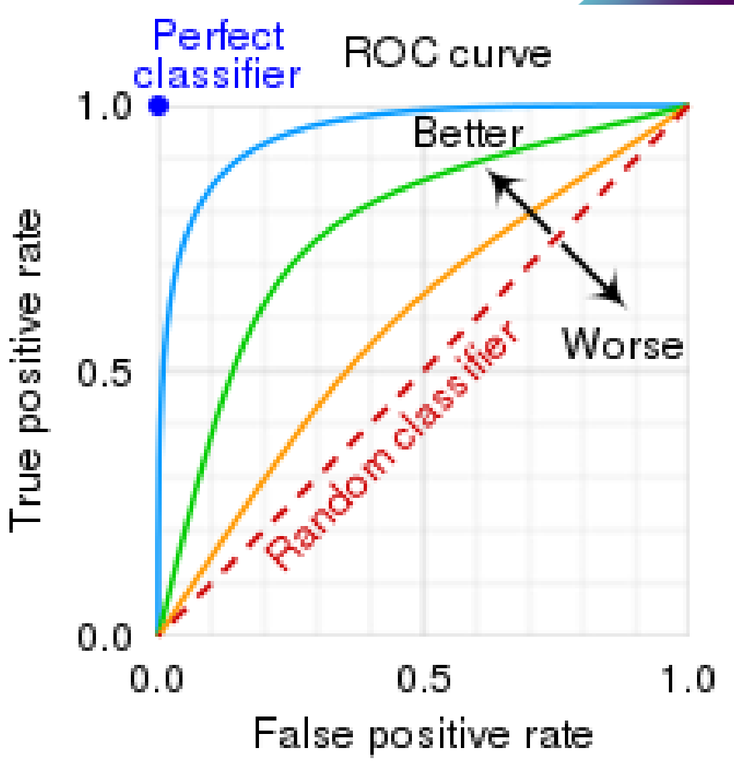
Used Models

- Logistic Regression,
- Support Vector Machine
- K-nearest neighbors algorithm
- Random Forest
- Naive Bayes Theorem
- Gradient Boosting Classifier

Metrics to Compare Results

- Accuracy
- ROC

ROC:



Reminders:

ACCURACY:

The number of classifications a model correctly predicts divided by the total number of predictions made

Logistic Regression Results

Accuracy:

0.8066465256797583

ROC:

0.7433641578902304

TP: [24 21 24 26 21 21 15 23 24 23]

TN: [79 79 85 80 85 86 87 79 83 80]

FN: [18 20 17 15 20 20 26 19 18 19]

FP: [12 12 6 11 6 5 4 11 7 10]

Support Vector Machine Results

Accuracy:

0.8006042296072508

ROC:

0.7281231269800497

TP: [24 21 19 25 21 21 18 23 25 23]

TN: [80 81 83 80 85 86 88 80 84 82]

FN: [18 20 22 16 20 20 23 19 17 19]

FP: [11 10 8 11 6 5 3 10 6 8]

K-Nearest Neighbors Algorithm Results

Accuracy:

0.8731117824773413

ROC:

0.8484887404743556

TP: [33 33 28 34 27 31 26 30 29 30]

TN: [79 76 85 80 85 79 82 79 77 81]

FN: [9 8 13 7 14 10 15 12 13 12]

FP: [12 15 6 11 6 12 9 11 13 9]

Random Forest Results

Accuracy:

0.9879154078549849

ROC:

[illegible]03921568627452

TP: [42 37 39 36 37 38 36 40 42 38]

TN: [91 87 90 88 85 91 88 89 87 90]

FN: [0 4 2 5 4 3 5 2 0 4]

FP: [0 4 1 3 6 0 3 1 3 0]

Naive Bayes Theorem Results

Accuracy:

0.7734138972809668

ROC:

0.7302209093244285

TP: [29 25 25 27 23 24 16 26 27 29]

TN: [75 73 80 77 81 78 80 76 75 73]

FN: [13 16 16 14 18 17 25 16 15 13]

FP: [16 18 11 14 10 13 11 14 15 17]

Gradient Boosting Classifier Results

Accuracy:

0.8912386706948641

ROC:

0.8615891771555784

TP: [34 33 33 32 30 34 24 28 32 30]

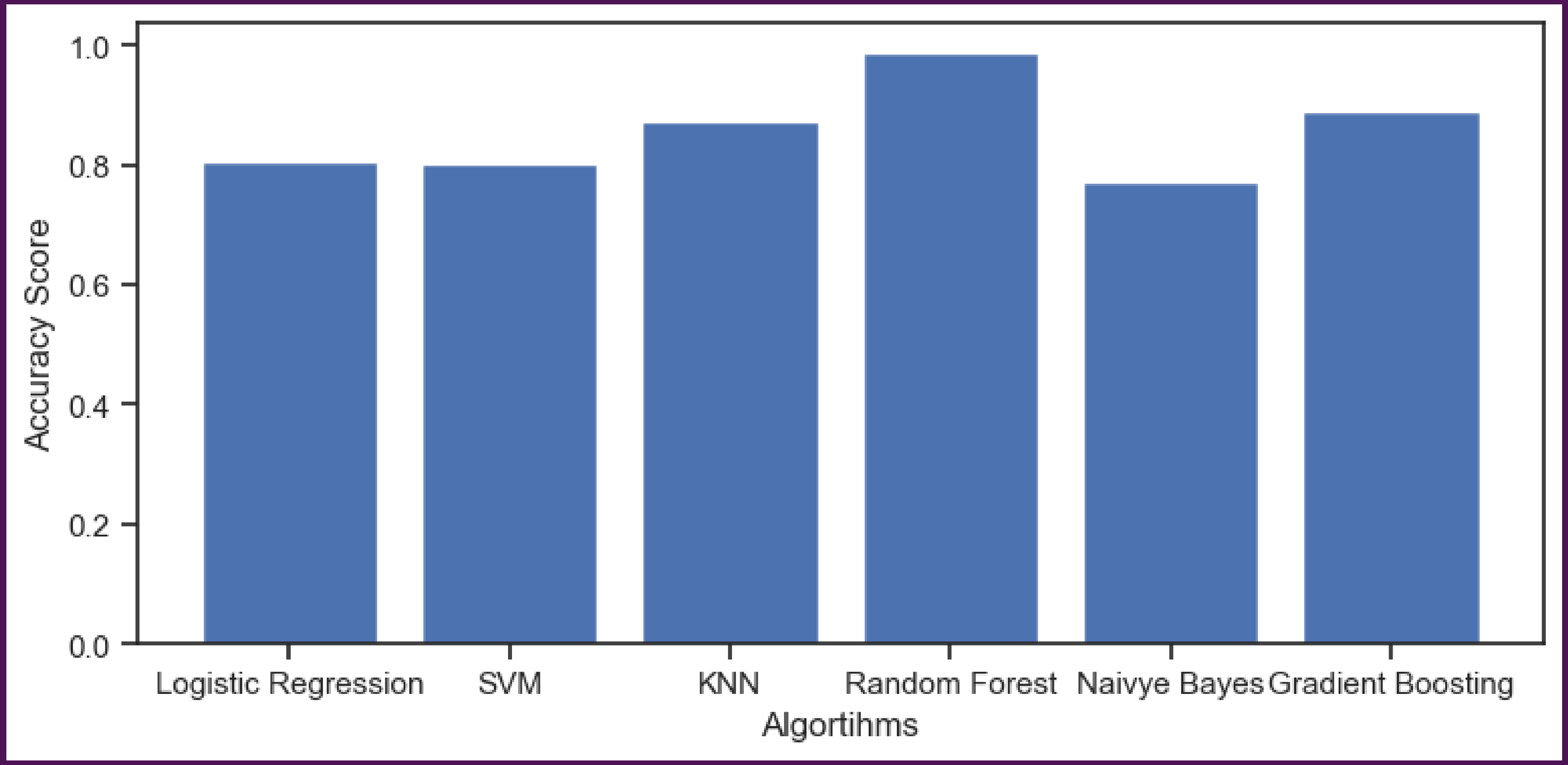
TN: [84 83 87 87 86 85 87 84 83 86]

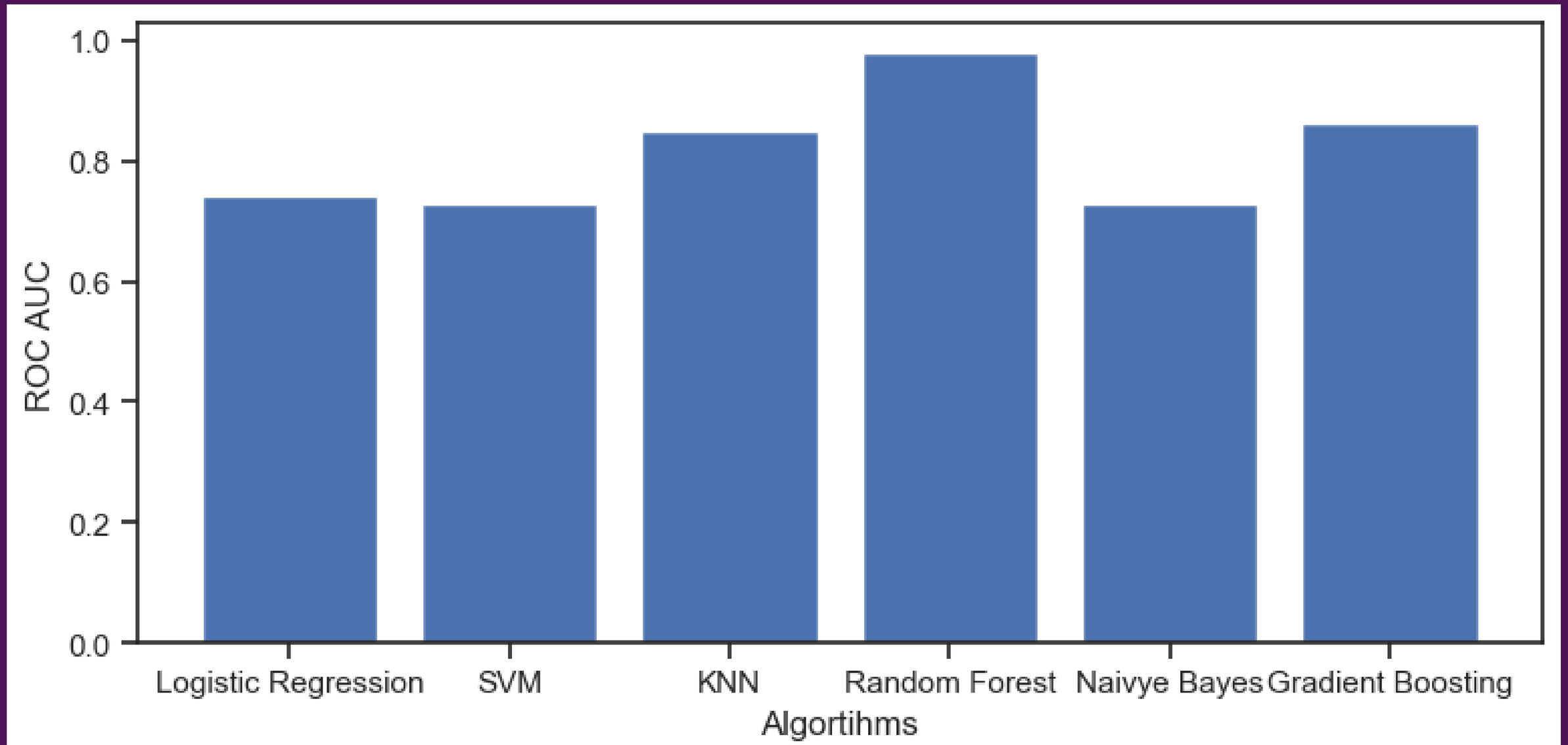
FN: [8 8 8 9 11 7 17 14 10 12]

FP: [7 8 4 4 5 6 4 6 7 4]

Results on Bar Graph

Accuracy Score and ROC AUC


1.0
0.8
0.6
0.4
0.2
0.0
Accuracy Score
Logistic Regression
SVM
KNN
Random Forest
Naivye Bayes
Gradient Boosting
Algortihms


1.0
0.8
0.6
0.4
0.2
0.0
ROC AUC
Logistic Regression
SVM
KNN
Random Forest
Naivye Bayes
Gradient Boosting
Algortihms

Thank you for listening...