Report for Capstone project

Healthcare - Diabetes

Mahaalakshmi M.

Context:

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Problem Statement:

Build a model to accurately predict whether the patients in the dataset have diabetes or not?

Dataset Description:

The datasets consists of several medical predictor variables and one target variable, `Outcome`. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

Pregnancies: Number of times pregnant
Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
BloodPressure: Diastolic blood pressure (mm Hg)
SkinThickness: Triceps skin fold thickness (mm)
Insulin: 2-Hour serum insulin (mu U/ml)
BMI: Body mass index (weight in kg/(height in m)^2)
DiabetesPedigreeFunction: Diabetes pedigree function
Age: Age (years)
Outcome: Class variable (0 or 1) 268 of 768 are 1, the others are 0

Project Task: Week 1
Data Exploration:

1. Perform descriptive analysis. Understand the variables and their corresponding values. On the columns below, a value of zero does not make sense and thus indicates missing value:

- **Glucose**
- **BloodPressure**
- **SkinThickness**
- **Insulin**
- **BMI**

2. Visually explore these variables using histograms. Treat the missing values accordingly.
3. There are integer and float data type variables in this dataset. Create a count (frequency) plot describing the data types and the count of variables.

Data Exploration:

4. Check the balance of the data by plotting the count of outcomes by their value. Describe your findings and plan future course of action.
5. Create scatter charts between the pair of variables to understand the relationships. Describe your findings.
6. Perform correlation analysis. Visually explore it using a heat map.

Project Task: Week 2

Data Modeling:

1. Devise strategies for model building. It is important to decide the right validation framework. Express your thought process.
2. Apply an appropriate classification algorithm to build a model.
3. Compare various models with the results from KNN algorithm.
4. Create a classification report by analyzing sensitivity, specificity, AUC (ROC curve), etc.

Please be descriptive to explain what values of these parameter you have used.

Data Reporting:

5. Create a dashboard in tableau by choosing appropriate chart types and metrics useful for the business. The dashboard must entail the following:

- Pie chart to describe the diabetic or non-diabetic population
- Scatter charts between relevant variables to analyze the relationships
- Histogram or frequency charts to analyze the distribution of the data
- Heatmap of correlation analysis among the relevant variables
- Create bins of these age values: 20-25, 25-30, 30-35, etc. Analyze different variables for these age brackets using a bubble chart.

EDA on the dataset:

```
In [9]: df.head()
```

Out[9]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figure 1: Dataset sample

```
In [10]: df.describe()
```

Out[10]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Figure 2: Dataset Description

As seen in the data description, features like glucose, Blood pressure, Insulin and BMI have a values of ‘0’ as their minimum. As it is not practical to have these values in real life we treat them as outliers

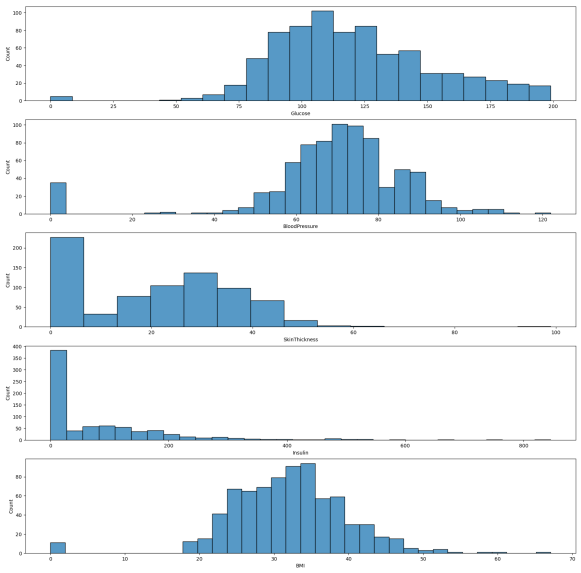

Figure 3: Histograms for NaN/‘0’ values

From the histogram graphs above, it can be seen that insulin has a skewness to it. To fill the NaN values for 'Insulin', a median value is used. The rest of the features are filled with their mean values as their distribution resemble normal distribution, and the mean value will remain the same is mean is used to fill the missing values.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	763.000000	733.000000	541.000000	394.000000	757.000000	768.000000	768.000000	768.000000
mean	3.845052	121.686763	72.405184	29.153420	155.548223	32.457464	0.471876	33.240885	0.348958
std	3.369578	30.535641	12.382158	10.476982	118.775855	6.924988	0.331329	11.760232	0.476951
min	0.000000	44.000000	24.000000	7.000000	14.000000	18.200000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	64.000000	22.000000	76.250000	27.500000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	29.000000	125.000000	32.300000	0.372500	29.000000	0.000000
75%	6.000000	141.000000	80.000000	36.000000	190.000000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Figure 4: Dataset description after converting to NaN values.

```
Pregnancies                 0
Glucose                     5
BloodPressure              35
SkinThickness             227
Insulin                   374
BMI                        11
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

Figure 5: Null values

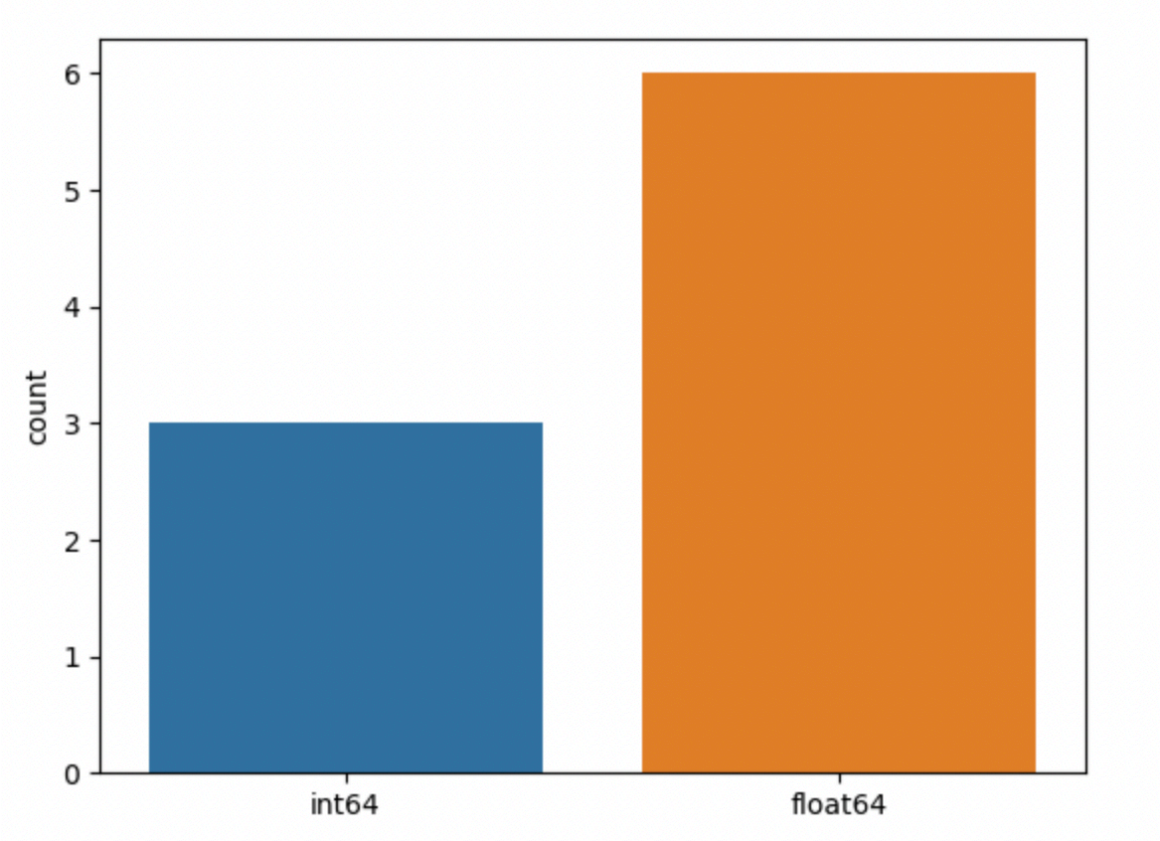

Figure 6: Histogram for datatypes in the dataset

The figure above shows the distribution of types of data that are present in the dataset

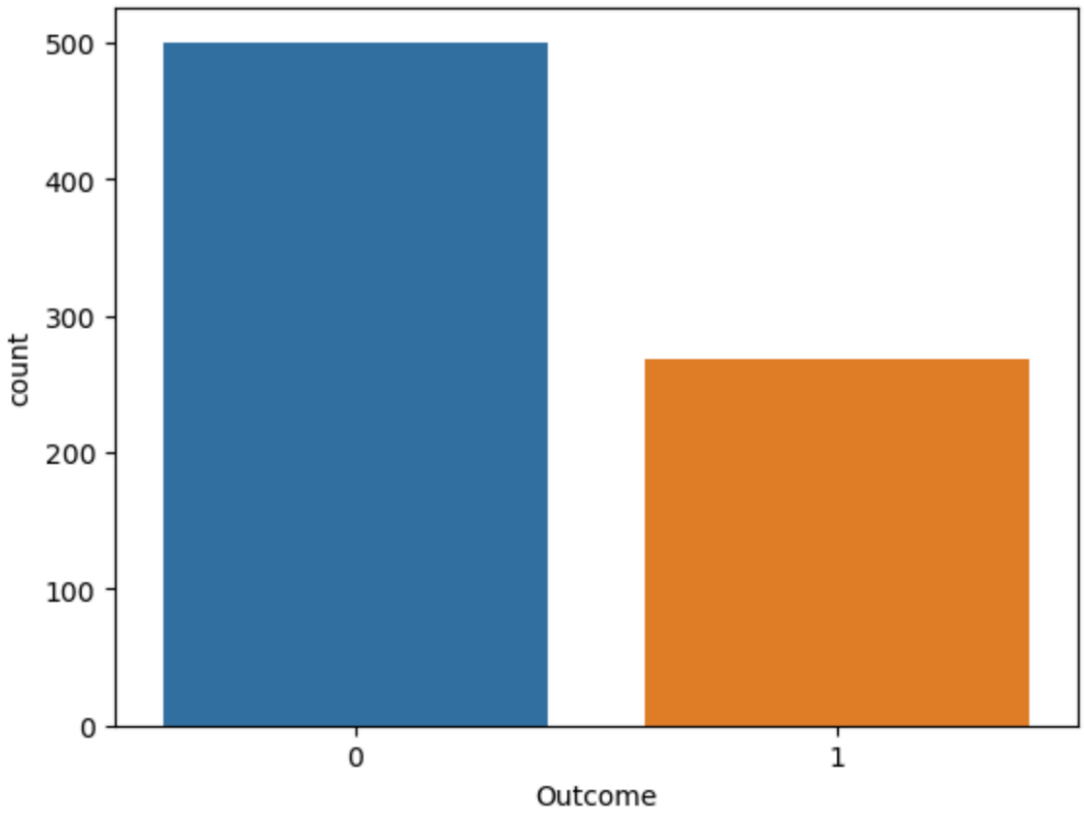

Figure 7: Checking the balance of dependent variable

From the histogram above, it is clear that the target variable is highly imbalanced. It is important to balance this data as it affects the accuracy and precision during fitting the model. Imbalanced data, causes the machine learning model to learn more features about the value with higher count, and the prediction value of the lower count value will not be very accurate. To avoid this, data balancing is done using Borderline SMOTE.

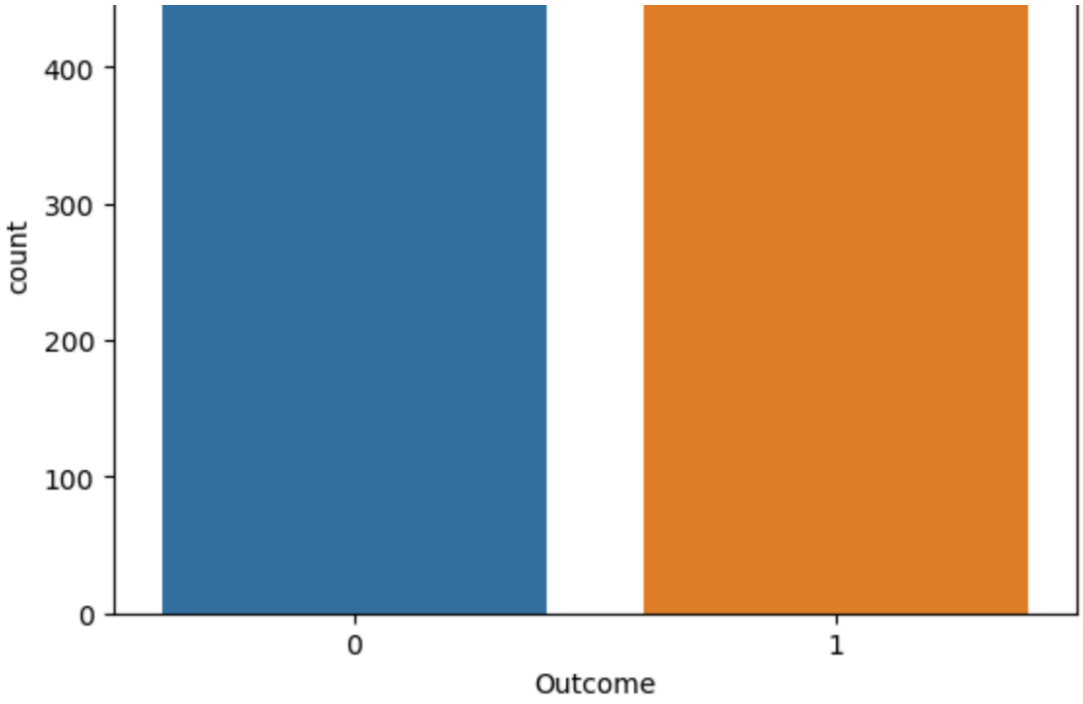

Figure 8: Checking the balance after using SMOTE

It is important to understand how data relates with each other, and can be helpful to check if any variable has importance against others to determine the dependent variable. From the figure below, it can be said that glucose and BMI are relatively of higher importance than the other variables. It can also be seen that all of the variables will be needed to predict the output variable and cannot depend on just BMI and Glucose for the prediction. To understand the data better, it is required to perform a correlation analysis on the data.

Figure 9: Scatter plot analysis of all variables

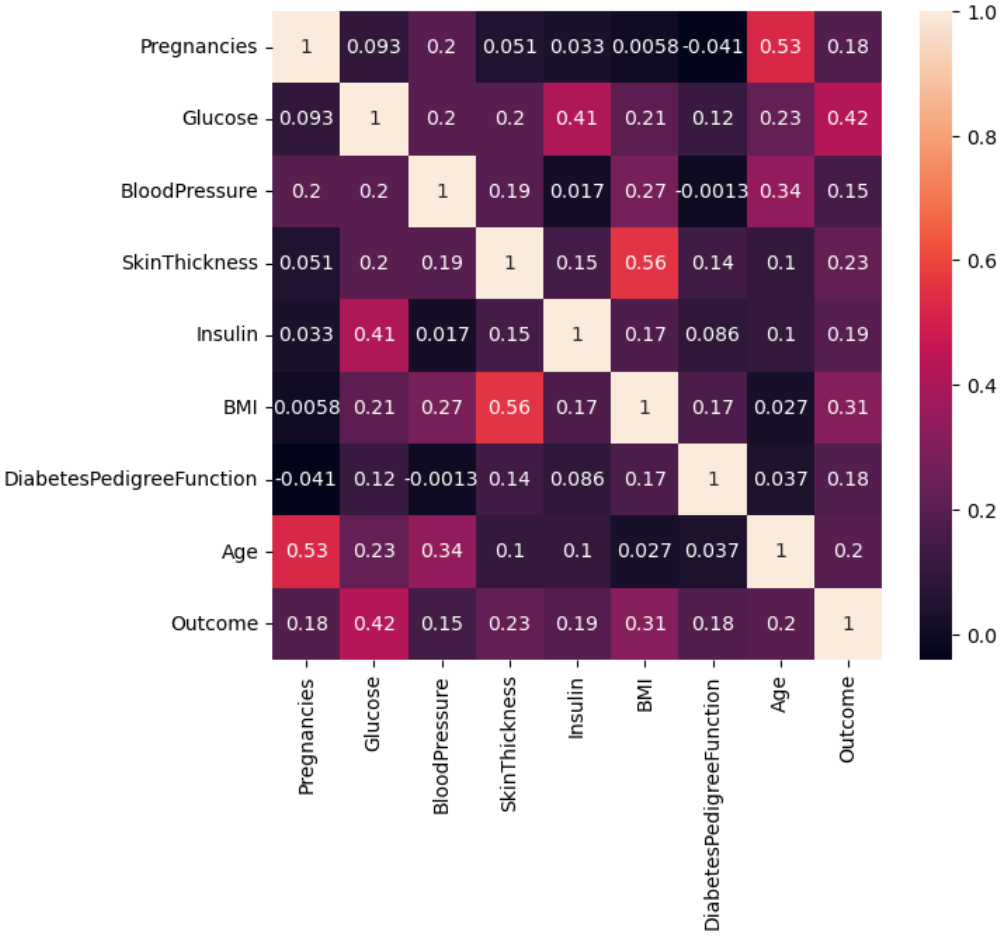

Figure 10: Heatmap of all variables

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.127911	0.208522	0.082989	0.025047	0.021565	-0.033523	0.544341	0.221898
Glucose	0.127911	1.000000	0.218367	0.192991	0.419064	0.230941	0.137060	0.266534	0.492928
BloodPressure	0.208522	0.218367	1.000000	0.192816	0.045087	0.281268	-0.002763	0.324595	0.166074
SkinThickness	0.082989	0.192991	0.192816	1.000000	0.154678	0.542398	0.100966	0.127872	0.215299
Insulin	0.025047	0.419064	0.045087	0.154678	1.000000	0.180170	0.126503	0.097101	0.203790
BMI	0.021565	0.230941	0.281268	0.542398	0.180170	1.000000	0.153400	0.025519	0.311924
DiabetesPedigreeFunction	-0.033523	0.137060	-0.002763	0.100966	0.126503	0.153400	1.000000	0.033561	0.173844
Age	0.544341	0.266534	0.324595	0.127872	0.097101	0.025519	0.033561	1.000000	0.238356
Outcome	0.221898	0.492928	0.166074	0.215299	0.203790	0.311924	0.173844	0.238356	1.000000

Figure 11: Correlation values between all variables

From the heat map and correlation value table, it is seen that there is some correlation between BMI and skin thickness, and age and pregnancies.

Data modelling:

The dependent variable here is a categorical variable, hence classification algorithms are needed to predict the output.

In the project the following algorithms are used to predict the dependant 'y' variable:

- Random forest
- Support vector machine
- Naive bayes
- K-NN

The accuracy, confusion matrices, ROC(AUC) graphs, precision and recall graphs are shown to assess different models.

After each model is fitted and predicted, a grid search is used to check the best parameters, and cross validation is performed to check the models' validity.

The parameters for each of the model used after grid search are listed below:

- Random forest: n_estimators = 200, criterion = 'entropy', max_features = 'log2', max_depth = 8
- SVM: kernel = 'rbf', C=1, gamma = 0.01
- K-NN: n_neighbors = 1

The values selected above were based on grid search to give the best accuracy.

```
[[100  26]
 [ 34  90]]

0.76
```

Figure 12: Accuracy and confusion matrix for Random forest classifier

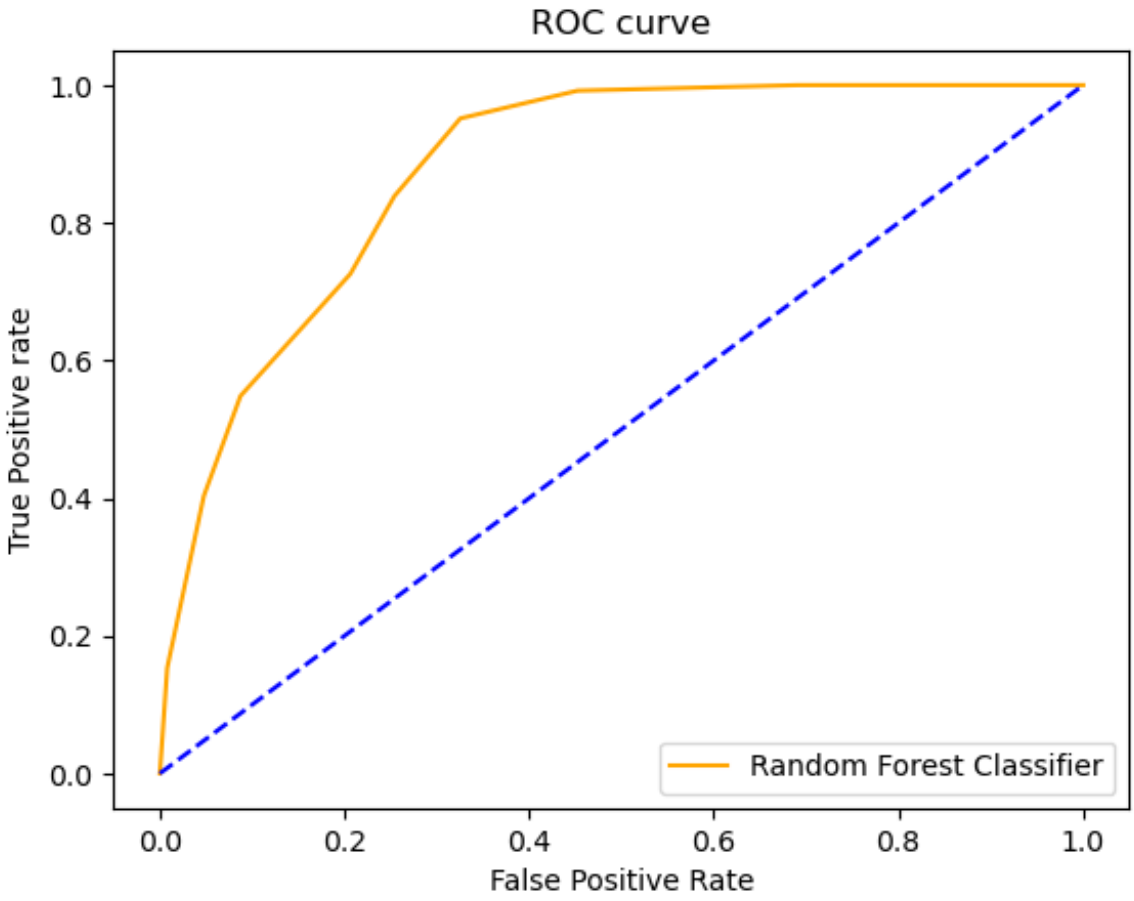

Figure 13: ROC (AUC) for random forest classifier

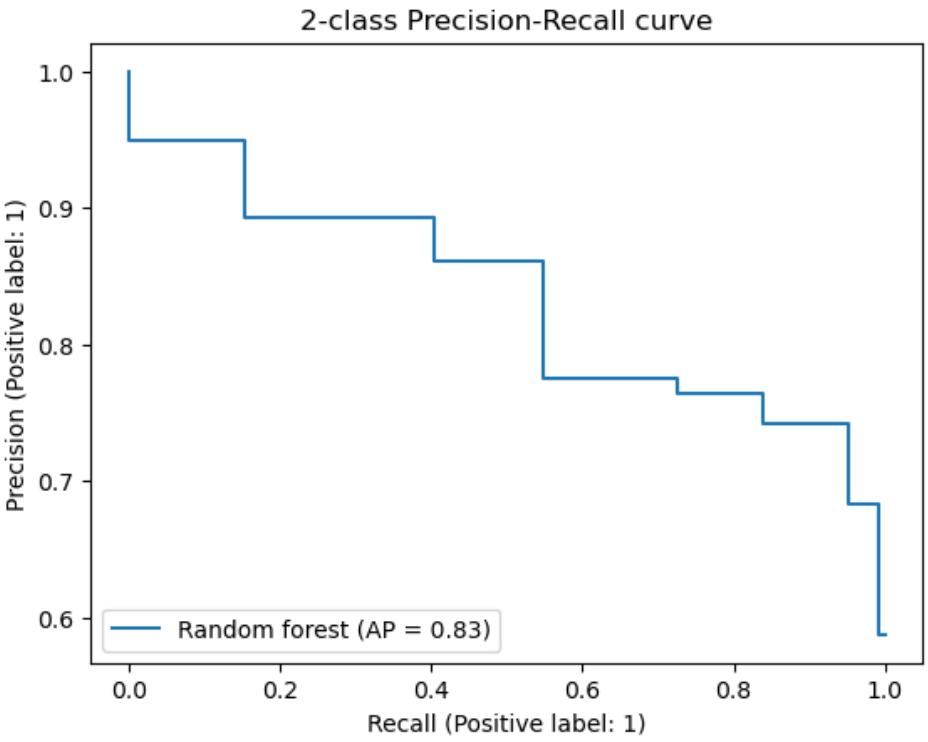

Figure 14: Precision and recall for Random forest classifier

```
[[102  24]
 [  7 117]]

0.876
```

Figure 15: Accuracy for random forest after Cross validation

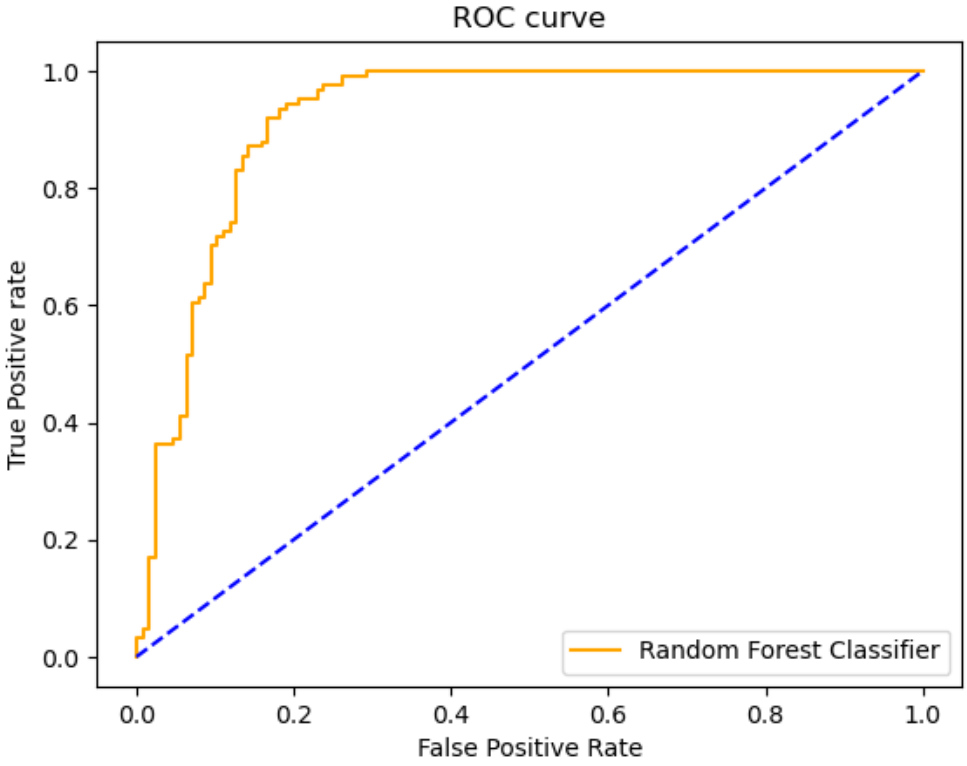

Figure 16: ROC(AUC) of random forest classifier after cross validation

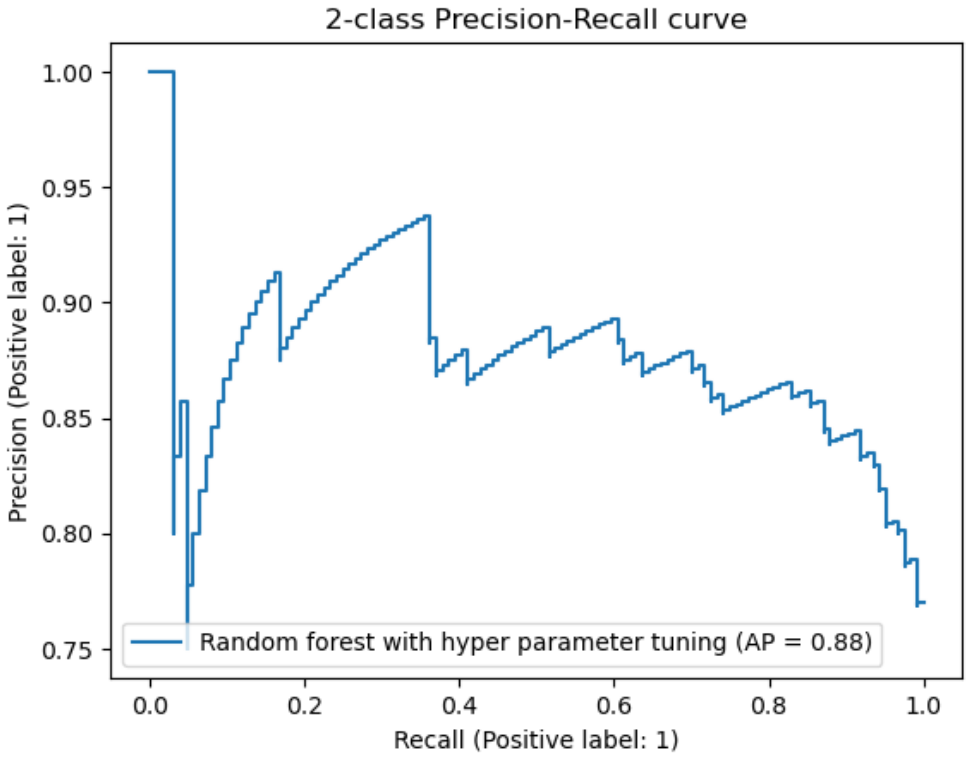

Figure 17: Precision and recall for random forest after cross validation

```
[[93 33]
 [33 91]]

0.736
```

Figure 18: Accuracy for SVM classifier

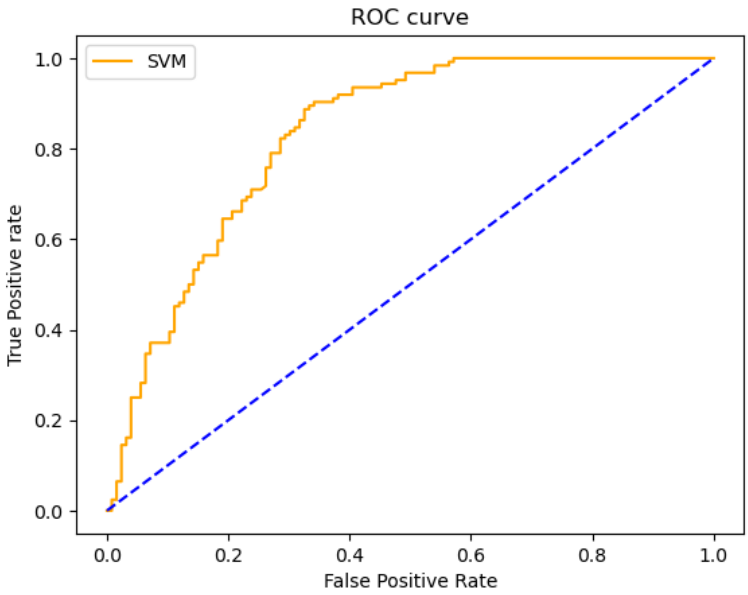

Figure 19: ROC(AUC) for SVM classifier

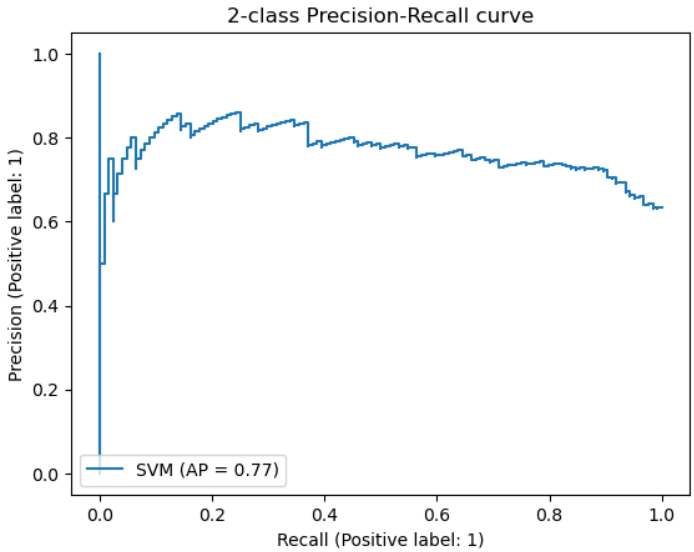

Figure 20: Recall and

```
[[109  17]
 [ 40  84]]

0.772
```

accuracy for SVM classifier

Figure 21: Accuracy of SVM after Cross Validation

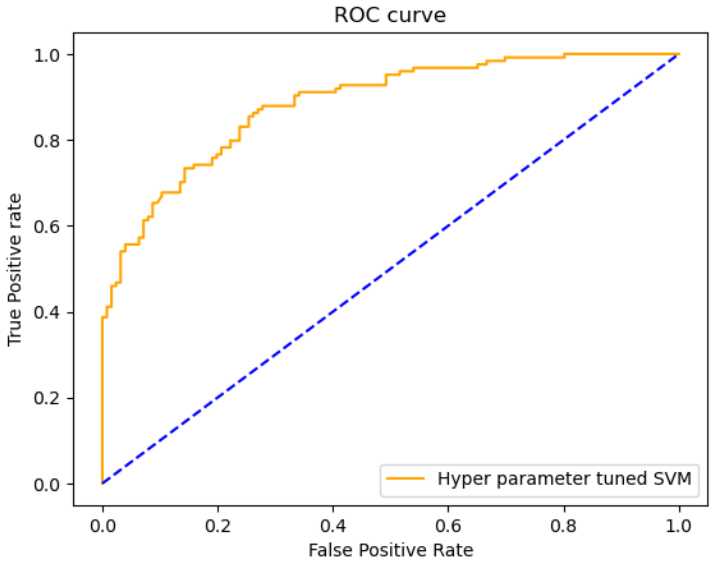

Figure 22: ROC(AUC) of SVM after Cross validation

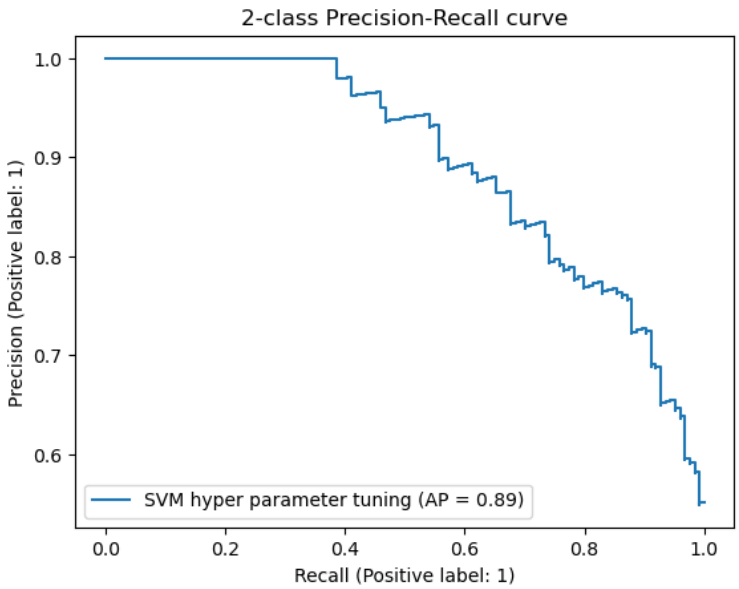

Figure 23: Recall and recision for SVM after cross validation

```
[[102  24]
 [ 42  82]]

0.736
```

Figure 24: Accuracy for Naive Bases

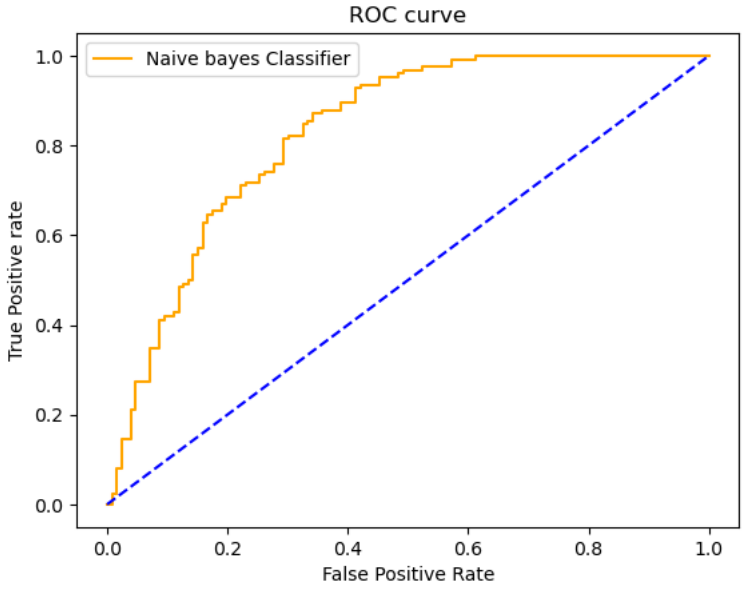

Figure 25: ROC(AUC) for Naive Bayes classifier

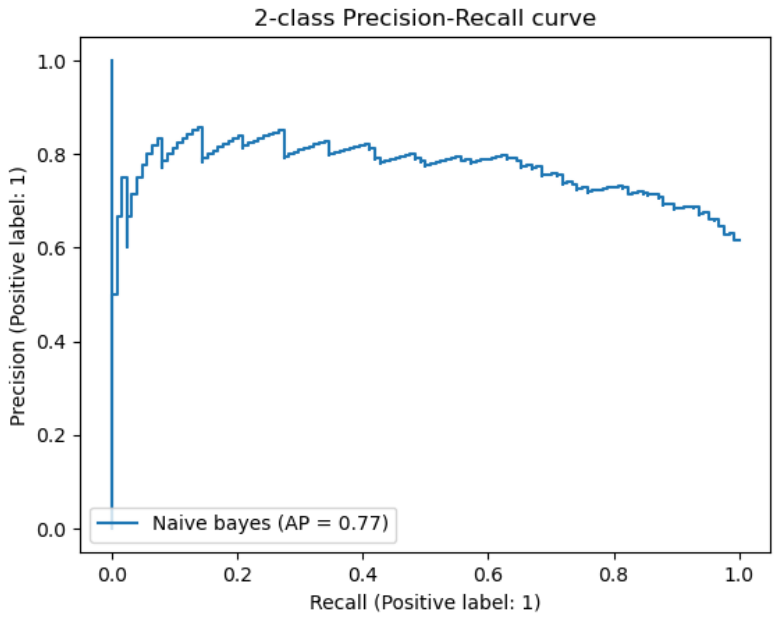

Figure 26: Recall and precision for Naive Bayes classifier

```
[[100  26]
 [ 35  89]]

0.756
```

Figure 27: Confusion matrix and accuracy for K-NN

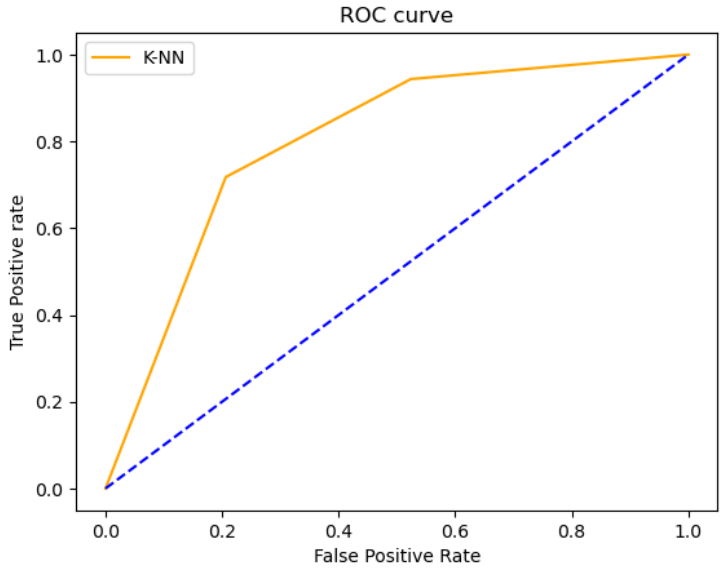

Figure 28: ROC(AUC) for KNN

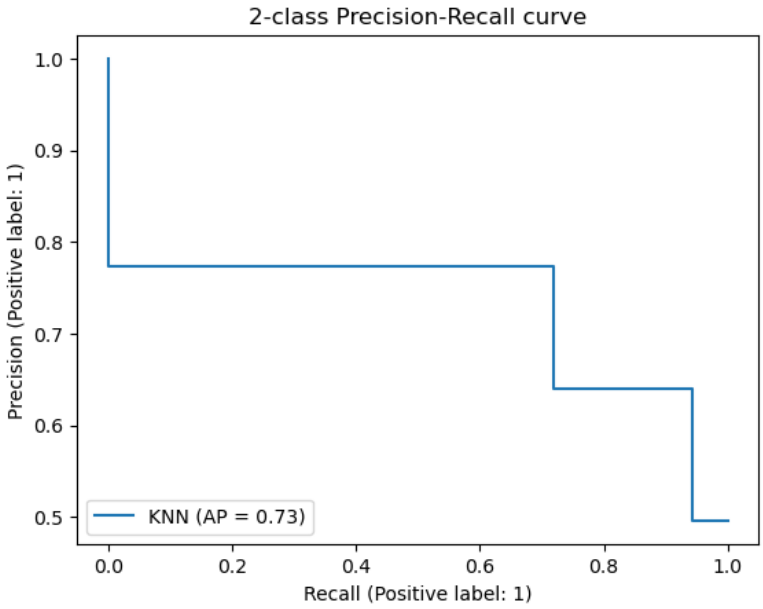

Figure 29: Recall and precision for KNN

```
[[ 89  37]
 [ 22 102]]

0.764
```

Figure 30: Confusion matrix and accuracy for K-NN after cross validation

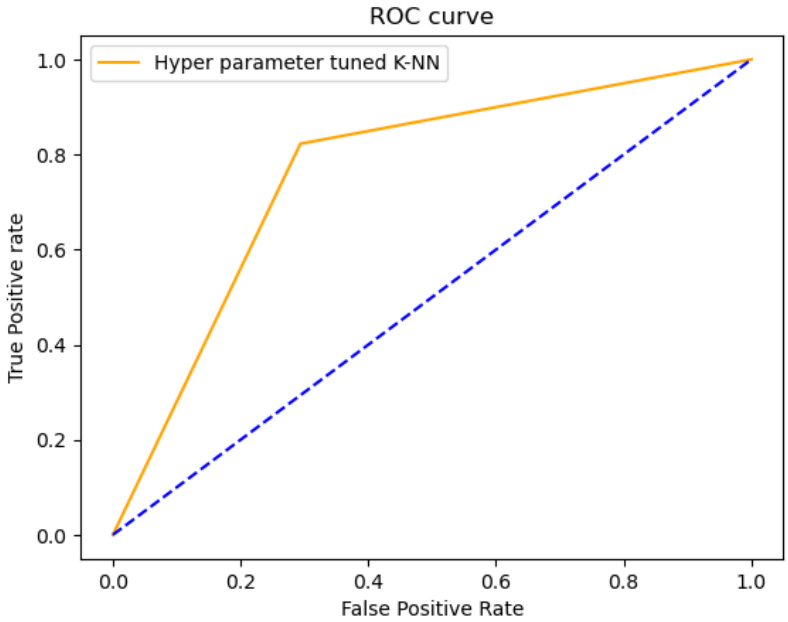

Figure 31: Recall and precision for K-NN after cross validation

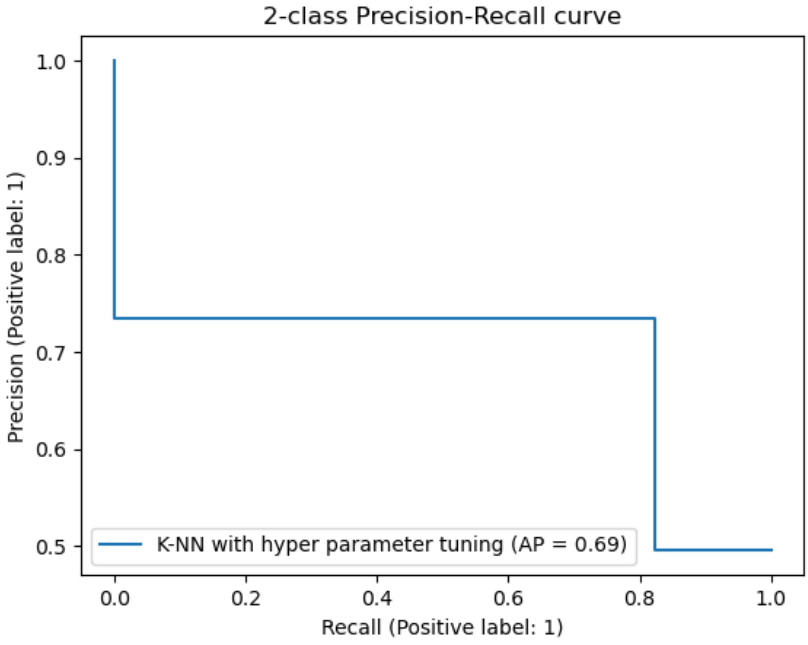

Figure 32: Precision and recall for K-NN after cross validation

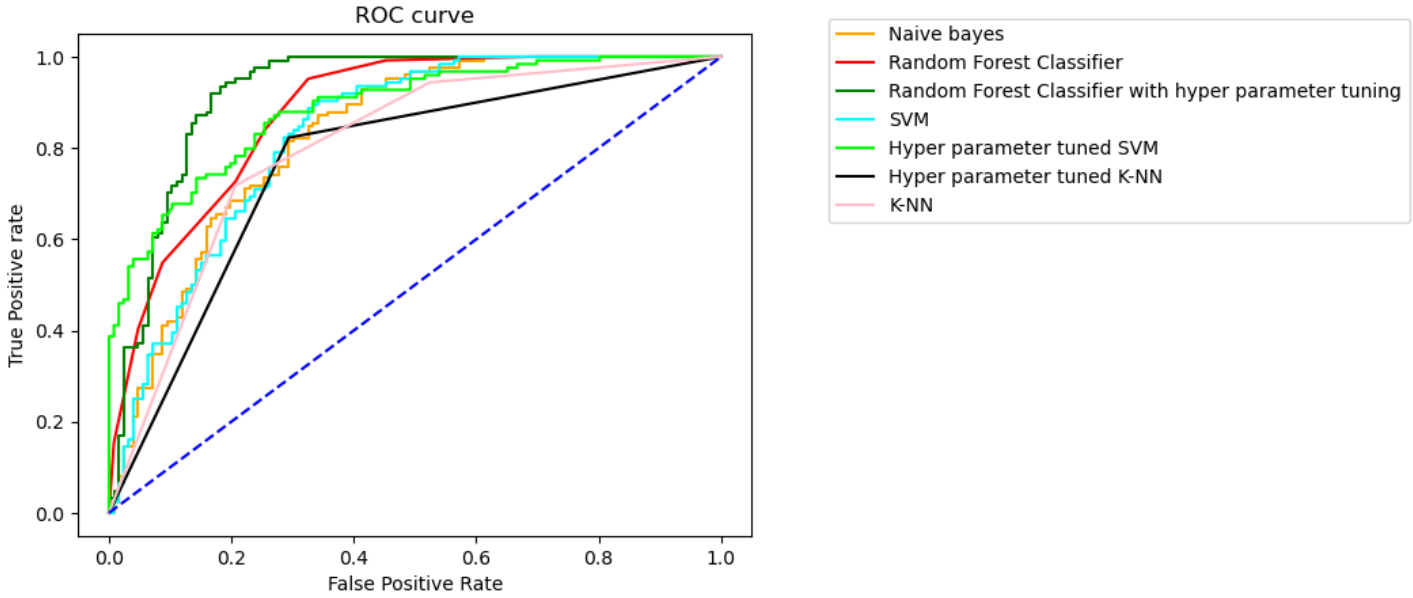

Figure 33: Comparison between ROC(AUC) for all the models

It is evident from the graphs above that random forest classifier performs the best out of all the other classification algorithms. The model, in theory performs better when the line is closer to the top left corner if the graph. The blue dashed line represents a baseline model. When the lines are closer to this, the model is under performing.

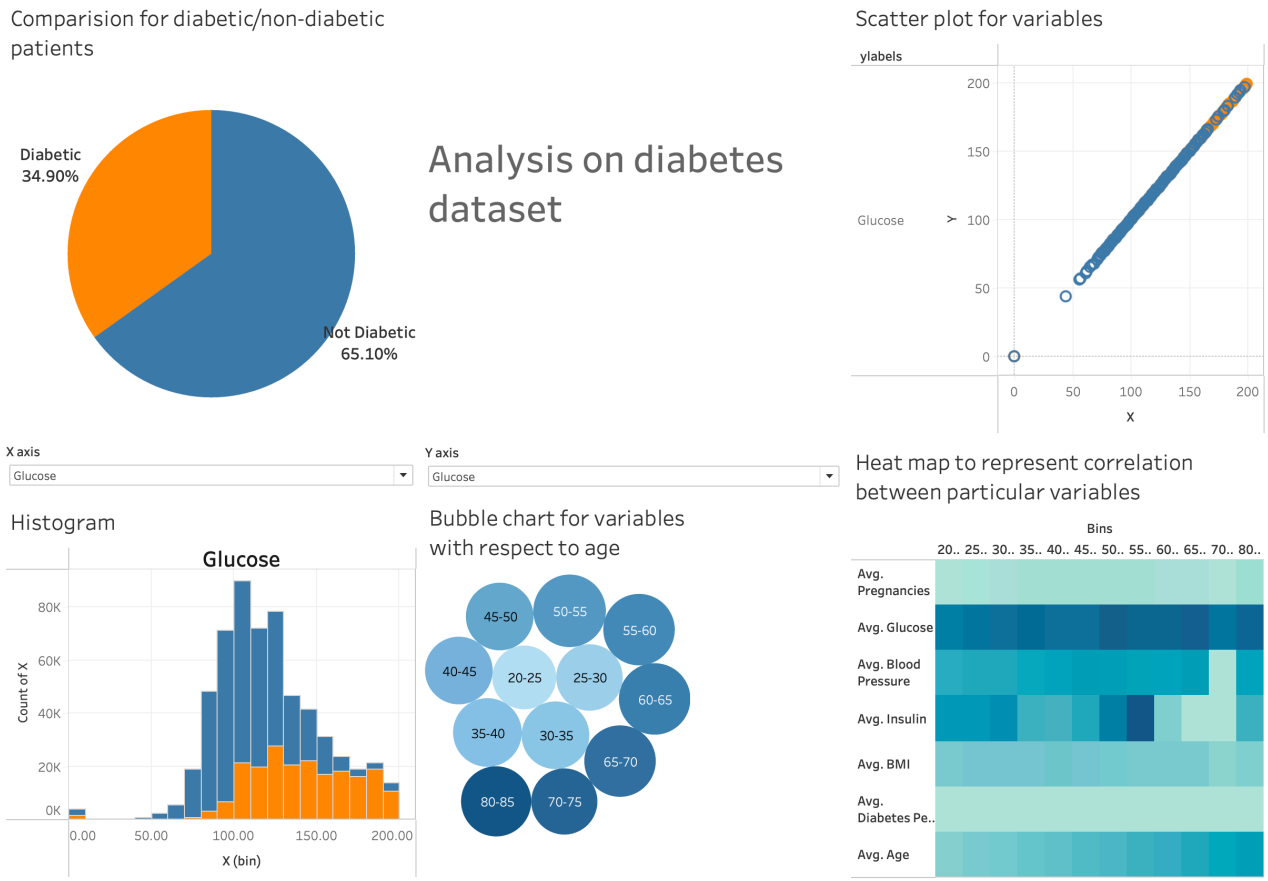

Figure 34: Tableau analysis

Here is a link for the visualisation workbook:

https://public.tableau.com/app/profile/lakshmi.m2649/viz/Capstone-Tableau_16766111249580/Dashboard1?publish=yes