Data Science Capstone Project
Project 2: Healthcare
Auther: PV-J

Note: This document summarizes the outcome of the notebook written for given dataset of this project

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

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

Blood Pressure: Diastolic blood pressure (mm Hg)

Skin Thickness: Triceps skin fold thickness (mm)

Insulin: 2-Hour serum insulin (mu U/ml)

BMI: Body mass index (weight in kg/(height in m)^2)

Diabetes Pedigree Function: Diabetes pedigree function

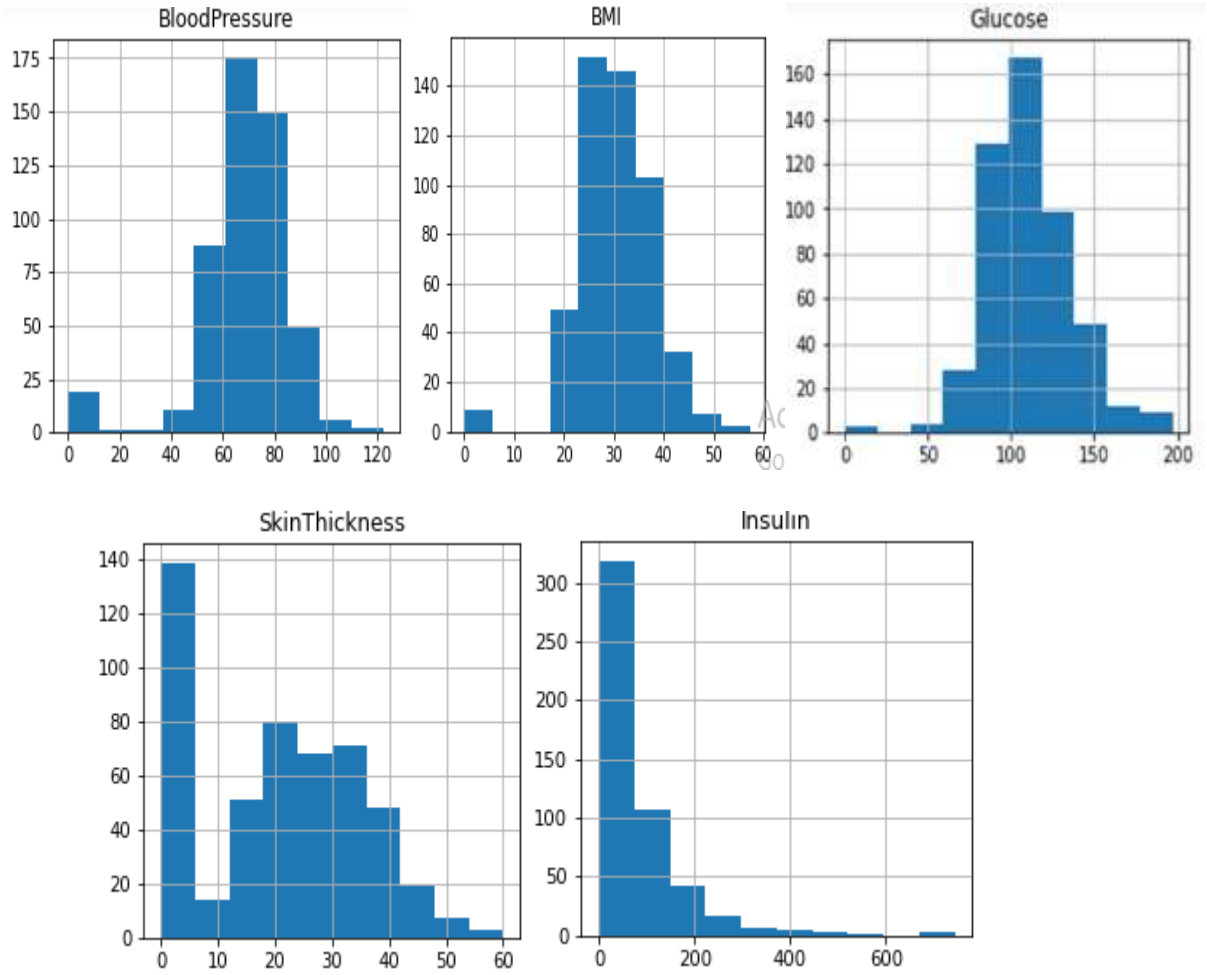
Age: Age (years)

Outcome: Class variable (0 or 1) 268 of 768 are 1, the others are 0

Step 1:

Descriptive analysis:

To perform descriptive analysis. It is very important to understand the attributes and corresponding values. During pre-processing of the minimum value of below listed columns are zero or not that checked on these columns, a value of zero does not make sense and thus indicates missing value. For this histogram for all attributes w.r.t 'Outcome' generated as follows:



After analyzing the histogram we can identify that there are some outliers in some columns. For Example:-

Blood Pressure - A living person cannot have a diastolic blood pressure of zero.

Glucose - Zero is invalid number as fasting glucose level would never be as low as zero.

Skin Fold Thickness - For normal people, skin fold thickness can't be less than 10 mm better yet zero.

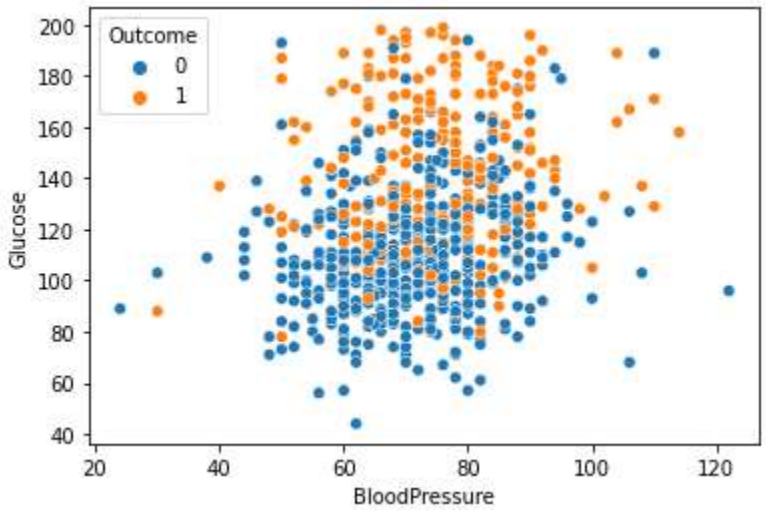
BMI: Should not be 0 or close to zero unless the person is really underweight which could be life-threatening.

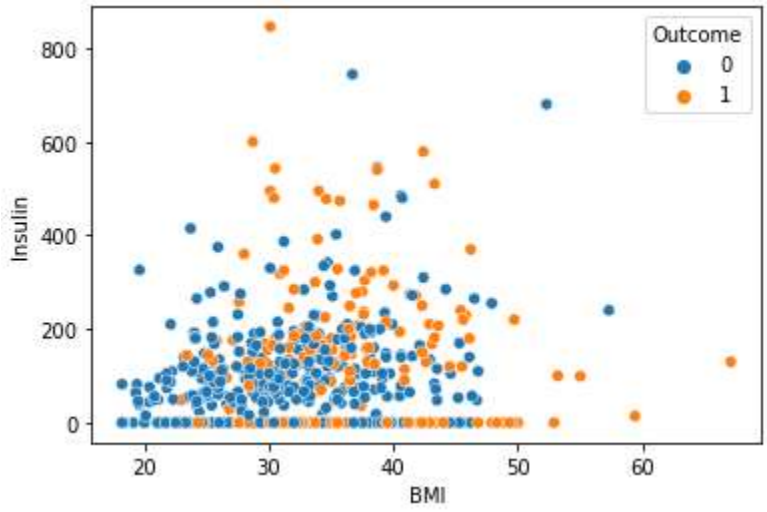
Insulin: In a rare situation a person can have zero insulin but by observing

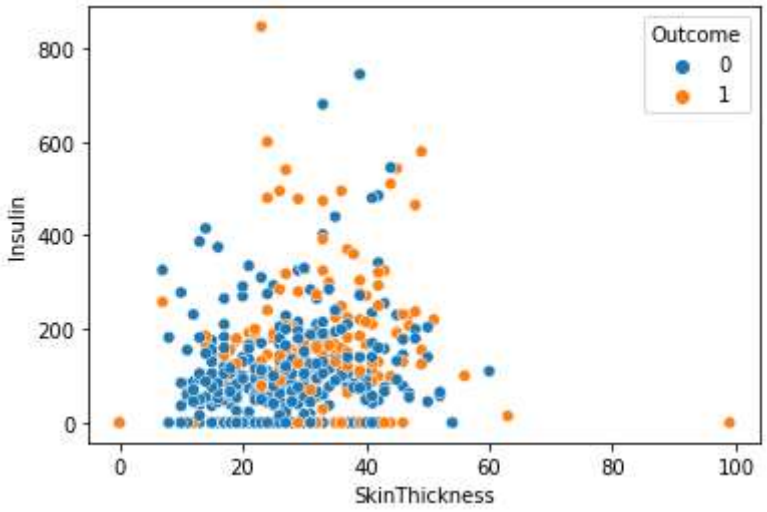
Step 2: Checking the balance of the data by plotting the count of outcomes by their value. Describe your findings and plan future course of actions

	count	mean	std	min	25%	50%	75%	max
Pregnancies	724.0	3.866022	3.362803	0.000	1.000	3.000	6.0000	17.00
Glucose	724.0	121.882597	30.750030	44.000	99.750	117.000	142.0000	199.00
BloodPressure	724.0	72.400552	12.379870	24.000	64.000	72.000	80.0000	122.00
SkinThickness	724.0	21.443370	15.732756	0.000	0.000	24.000	33.0000	99.00
Insulin	724.0	84.494475	117.016513	0.000	0.000	48.000	130.5000	846.00
BMI	724.0	32.467127	6.888941	18.200	27.500	32.400	36.6000	67.10
DiabetesPedigreeFunction	724.0	0.474765	0.332315	0.078	0.245	0.379	0.6275	2.42
Age	724.0	33.350829	11.765393	21.000	24.000	29.000	41.0000	81.00
Outcome	724.0	0.343923	0.475344	0.000	0.000	0.000	1.0000	1.00

Step 3: Create scatter charts between the pair of variables to understand the relationships.



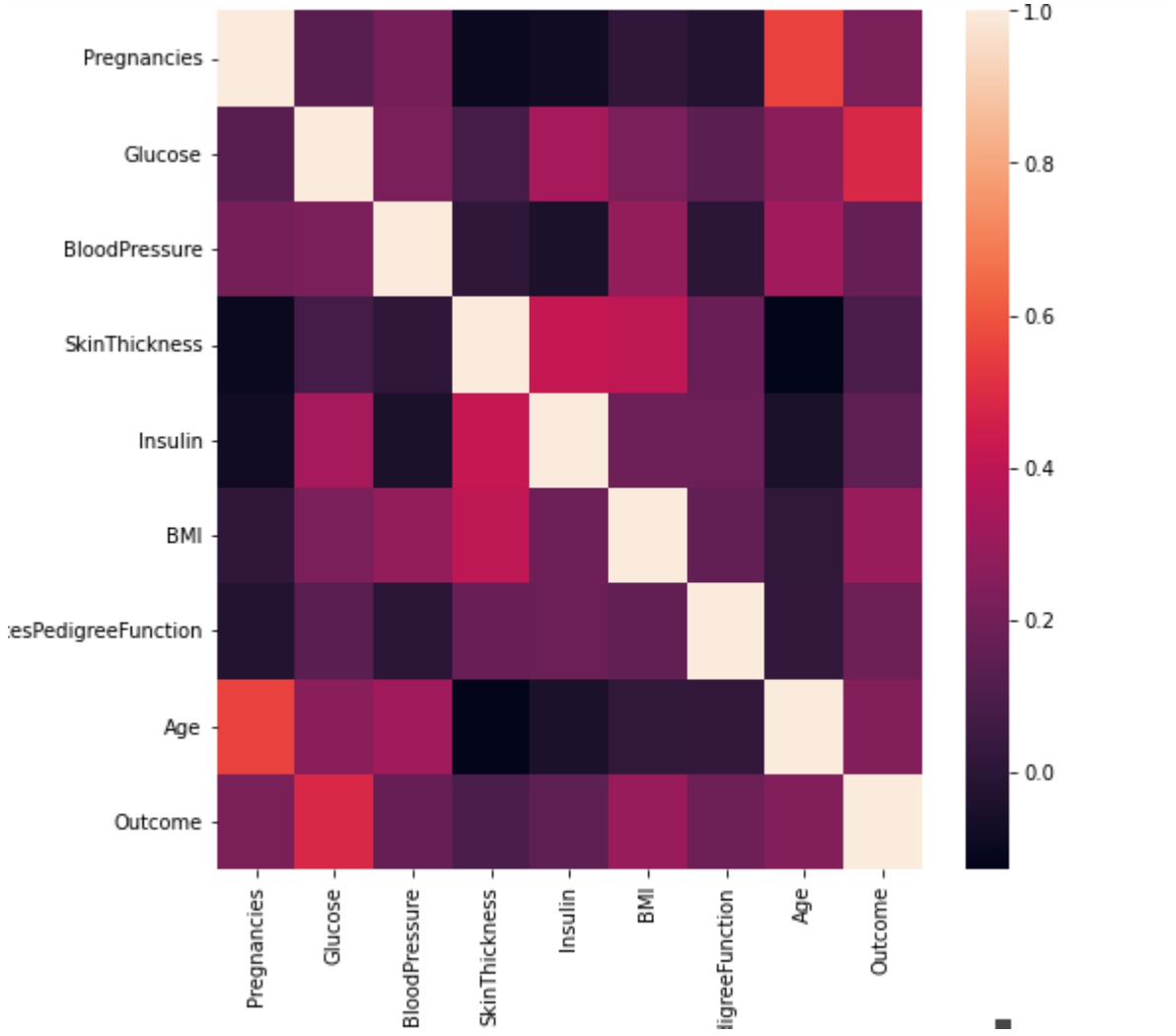




Above scatter plot shows that relationship between Glucose and Blood Pressure, Insulin and BMI and Insulin and Skin Thickness are highly correlated with each other w.r.t diabetic and non-diabetic patient. This can be observed from following table also.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.134915	0.209668	-0.095683	-0.080059	0.012342	-0.025996	0.557066	0.224417
Glucose	0.134915	1.000000	0.223331	0.074381	0.337896	0.223276	0.136630	0.263560	0.488384
BloodPressure	0.209668	0.223331	1.000000	0.011777	-0.046856	0.287403	-0.000075	0.324897	0.166703
SkinThickness	-0.095683	0.074381	0.011777	1.000000	0.420874	0.401528	0.176253	-0.128908	0.092030
Insulin	-0.080059	0.337896	-0.046856	0.420874	1.000000	0.191831	0.182656	-0.049412	0.145488
BMI	0.012342	0.223276	0.287403	0.401528	0.191831	1.000000	0.154858	0.020835	0.299375
DiabetesPedigreeFunction	-0.025996	0.136630	-0.000075	0.176253	0.182656	0.154858	1.000000	0.023098	0.184947
Age	0.557066	0.263560	0.324897	-0.128908	-0.049412	0.020835	0.023098	1.000000	0.245741
Outcome	0.224417	0.488384	0.166703	0.092030	0.145488	0.299375	0.184947	0.245741	1.000000

Step 4: Performing correlation analysis. Visually explore it by using a heat map

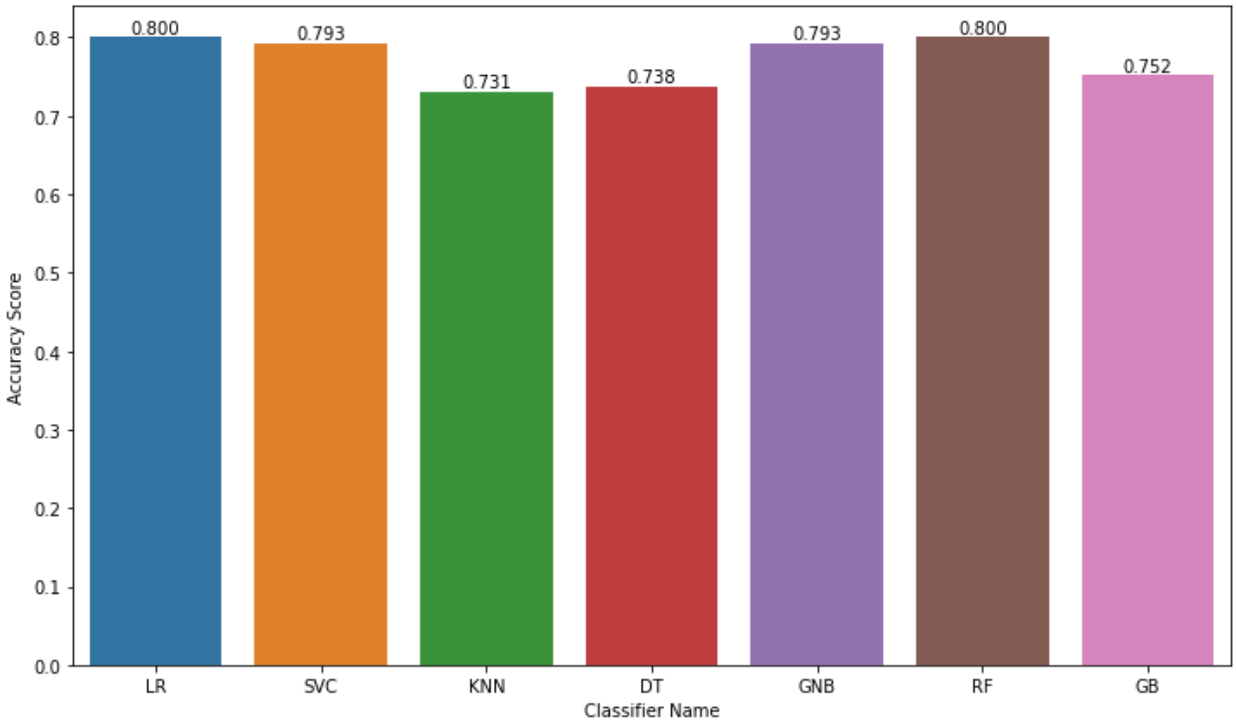


Step 5: Building Model

Implemented various algorithm like Logistic Regression, Support vector machine, K –Nearest Neighbour, Decision tree, Gaussian Naive Bayes, Random forest and Gradient boosting algorithms and relevant score and accuracy achieved are shown below

	Name	Score	Accuracy Score
0	LR	0.770294	0.800000
1	SVC	0.768566	0.793103
2	KNN	0.804836	0.731034
3	DT	1.000000	0.737931
4	GNB	0.751295	0.793103
5	RF	1.000000	0.800000
6	GB	0.929188	0.751724

Graphical Representation

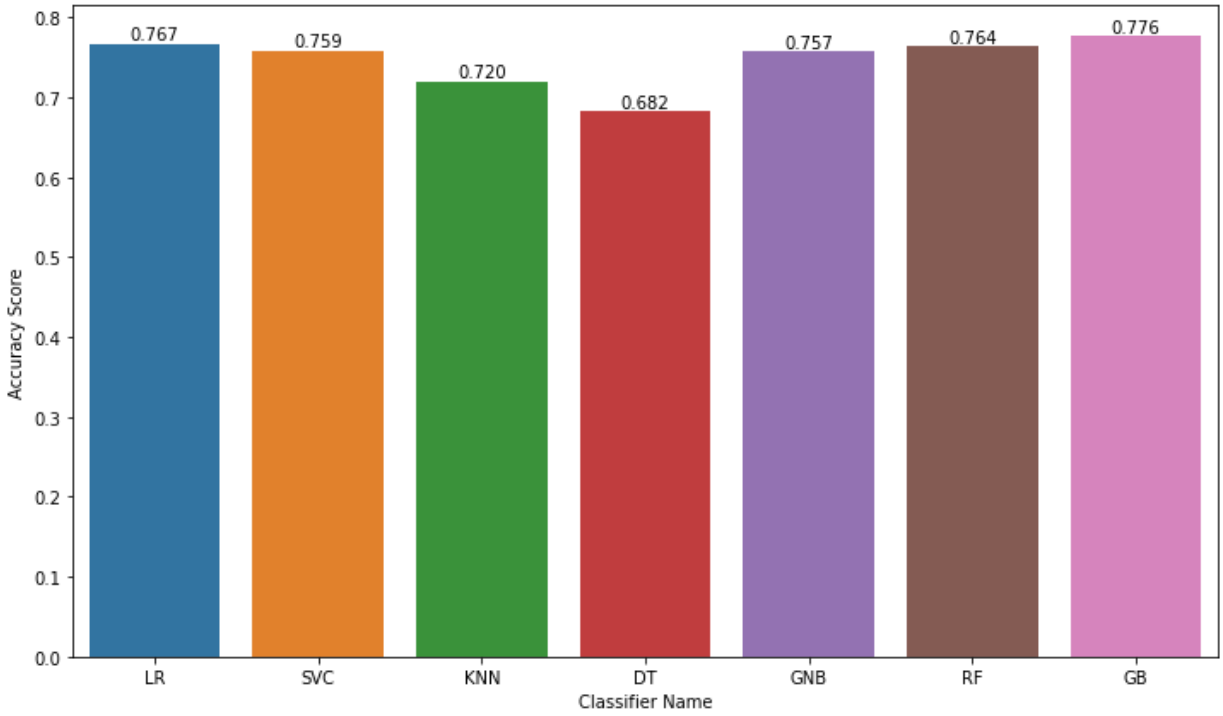


Applying K-Fold Cross Validation with Scikit Learn

With K-Fold cross validation as it is more accurate and use the data efficiently. Training the models using 10 fold cross validation and calculated the mean accuracy of the models. "k_fold_cross_val_score" provides its own training and accuracy calculation interface.

	Name	Score
0	LR	0.766781
1	SVC	0.758581
2	KNN	0.719787
3	DT	0.682420
4	GNB	0.757021
5	RF	0.764041
6	GB	0.776427

Graphical Representation



Above visualization of chart shows that Logistic Regression, Gaussian Naive Bayes, Random Forest and Gradient Boosting have performed better than the SVC, KNN, DT and GNB. From the base level we can observe that the Logistic Regression performs better than the other algorithms.

At the baseline Logistic Regression managed to achieve a classification accuracy of 77.64 %.

Step 6: Create a classification report by analysing sensitivity, specificity, AUC (ROC curve) etc. Please try to be as descriptive as possible to explain what values of these parameter you settled for? any why?

A classification report generated analyzing sensitivity, specificity, AUC (ROC curve), etc - AUC-ROC curve helps us visualize how well our machine learning classifier is performing.

	precision	recall	f1-score	support
0	0.79	0.90	0.84	475
1	0.74	0.54	0.62	249
accuracy			0.78	724
macro avg	0.76	0.72	0.73	724
weighted avg	0.77	0.78	0.77	724

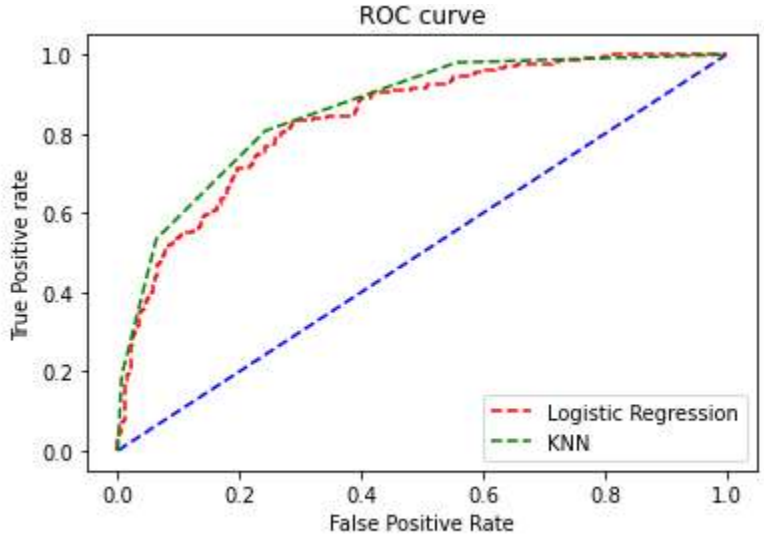
The classification report visualizer displays the precision, recall, F1, and support scores for the model. The classification report shows a representation of the main classification metrics on a per-class basis. This gives a deeper intuition of the classifier behaviour over global accuracy which can mask functional weaknesses in one class of a multiclass problem.

Metrics of importance in this project

The recall is the measure of our model correctly identifying True Positives.
Mathematically:

Recall = True Positives /(True Positive + False Negative)

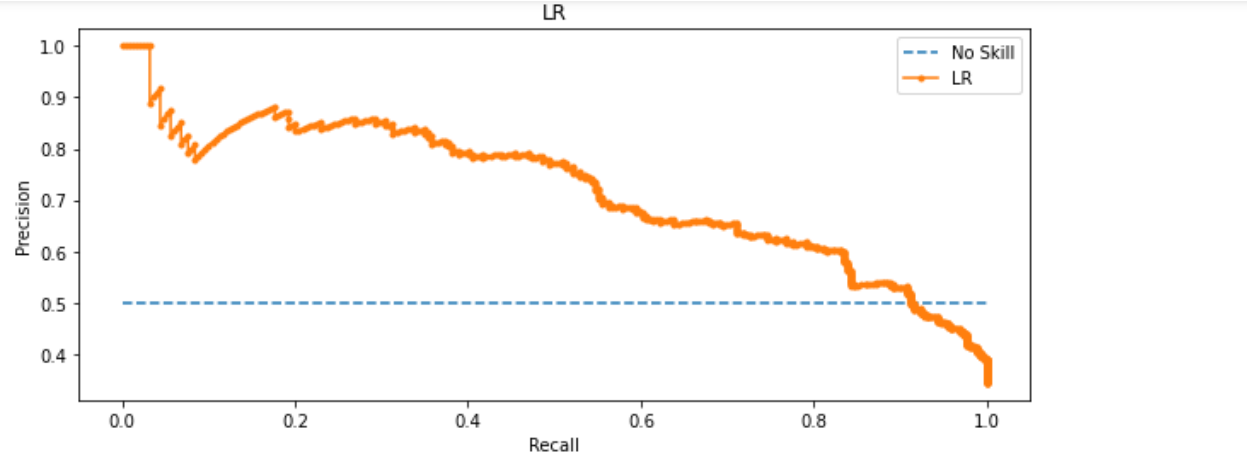
Recall also gives a measure of how accurately our model is able to identify the relevant data. We refer to it as Sensitivity or True Positive Rate. Higher sensitivity (recall) is more desirable for hospitals because it is more crucial to correctly identify “high risk” patients who are likely to be readmitted than identifying “low risk” patients.




```
AUC LR: 0.83796 AUC KNN: 0.86121
```

Above ROC curve shows that KNN performs better than LR

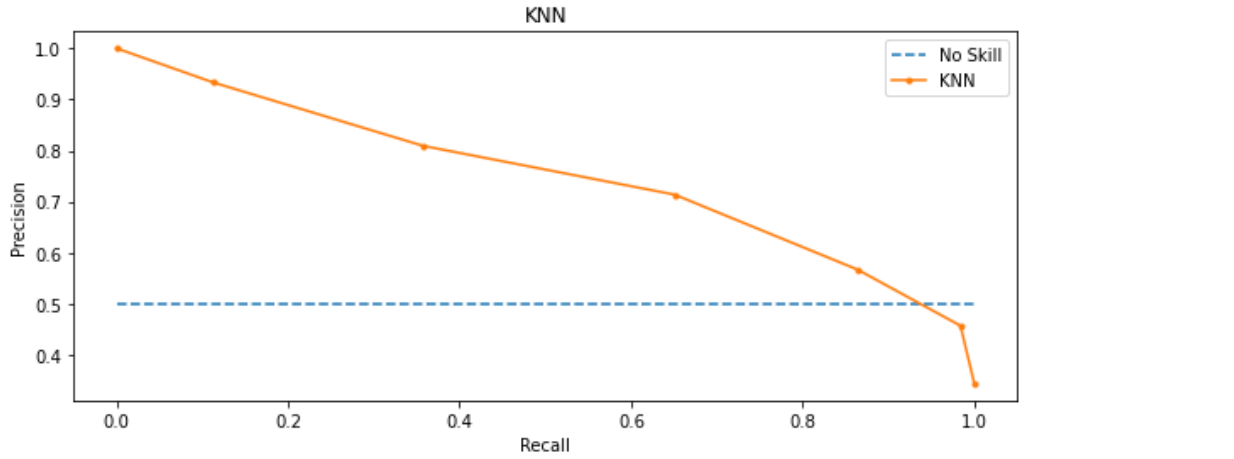
Precision Recall Curve for LR



LR calculated value : F1 Score =0.625, Area Under the Curve=0.722, Average Precision=0.723

The above precision-recall curve plot is showing the precision/recall for each threshold for a LR model (orange) compared to a no skill model (blue).

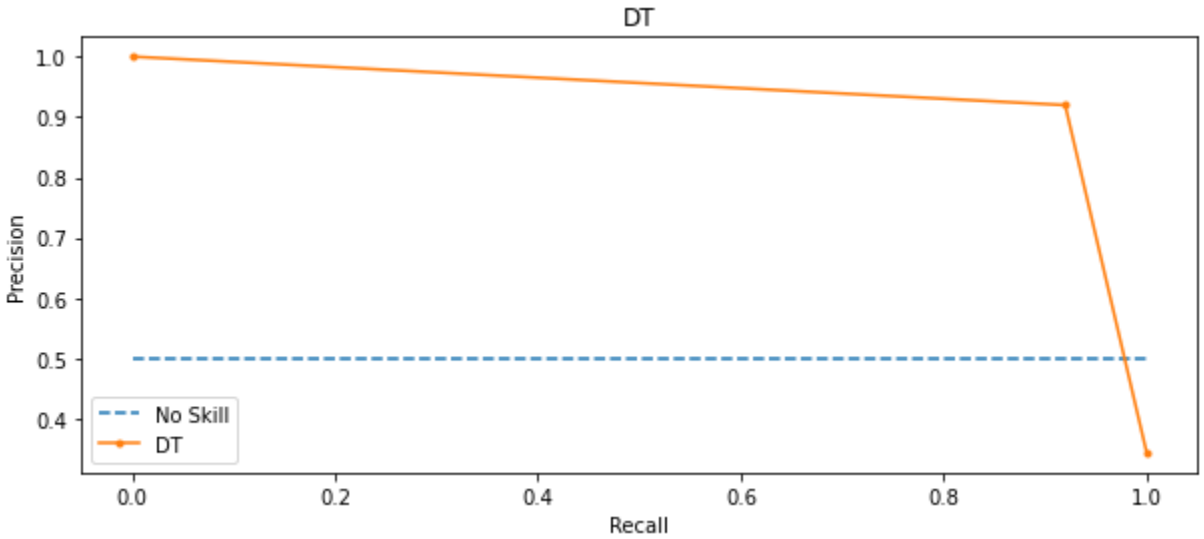
Precision Recall Curve for KNN



KNN calculated value : F1 Score =0.681, Area Under the Curve=0.750, Average Precision=0.694

The above precision-recall curve plot is showing the precision/recall for each threshold for a KNN model (orange) compared to a no skill model (blue).

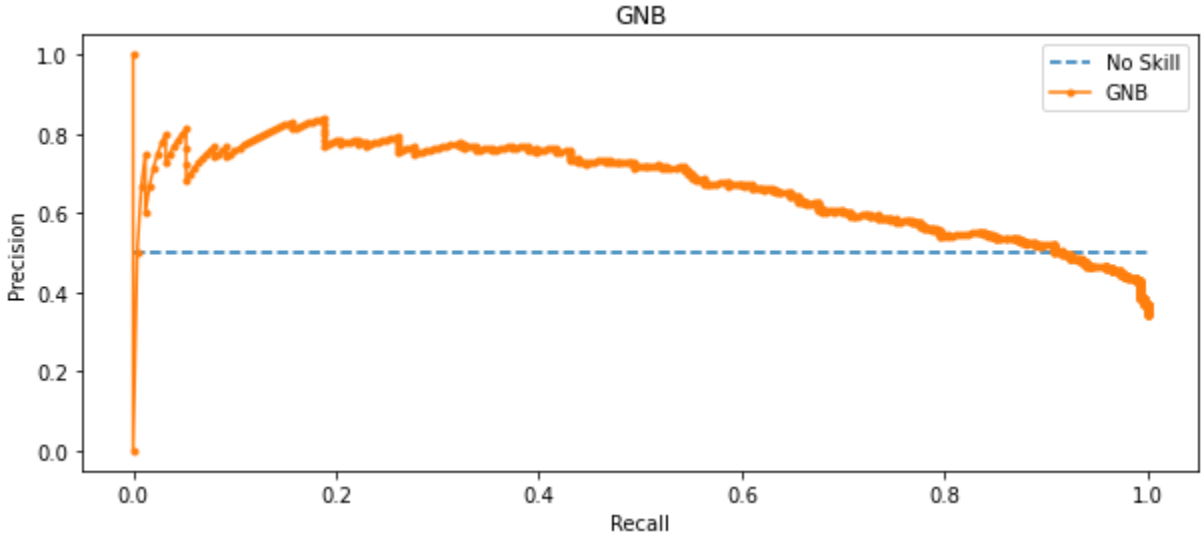
Precision Recall Curve for DT



DT calculated value : F1 Score =0.920, Area Under the Curve=0.933, Average Precision=0.873

The above precision-recall curve plot is showing the precision/recall for each threshold for a DT model (orange) compared to a no skill model (blue).

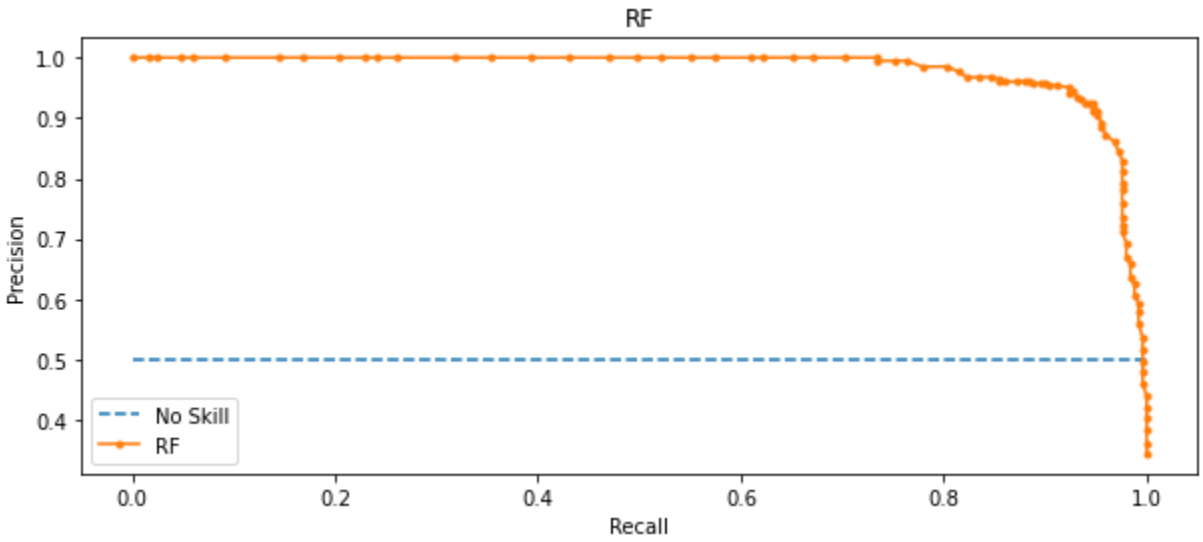
Precision Recall Curve for GNB



GNB calculated value : F1 Score =0.637, Area Under the Curve=0.671, Average Precision=0.674

The above precision-recall curve plot is showing the precision/recall for each threshold for a GNB model (orange) compared to a no skill model (blue).

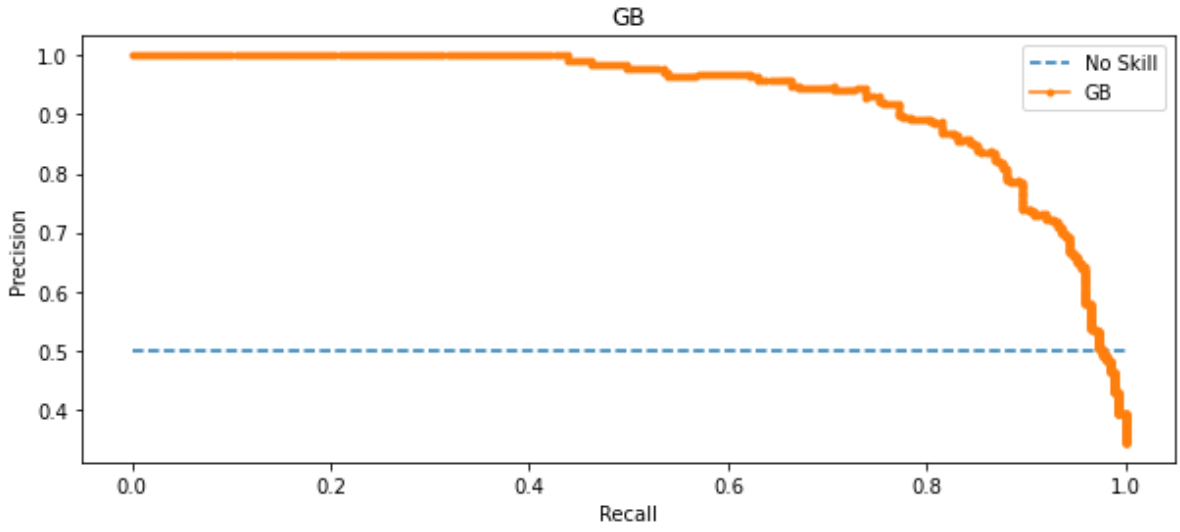
Precision Recall Curve for RF



RF calculated value: F1 Score =0.928, Area Under the Curve=0.980, Average Precision=0.979

The above precision-recall curve plot is showing the precision/recall for each threshold for a RF model (orange) compared to a no skill model (blue).

Precision Recall Curve for GB



GB calculated value : F1 Score =0.832, Area Under the Curve=0.929, Average Precision=0.929

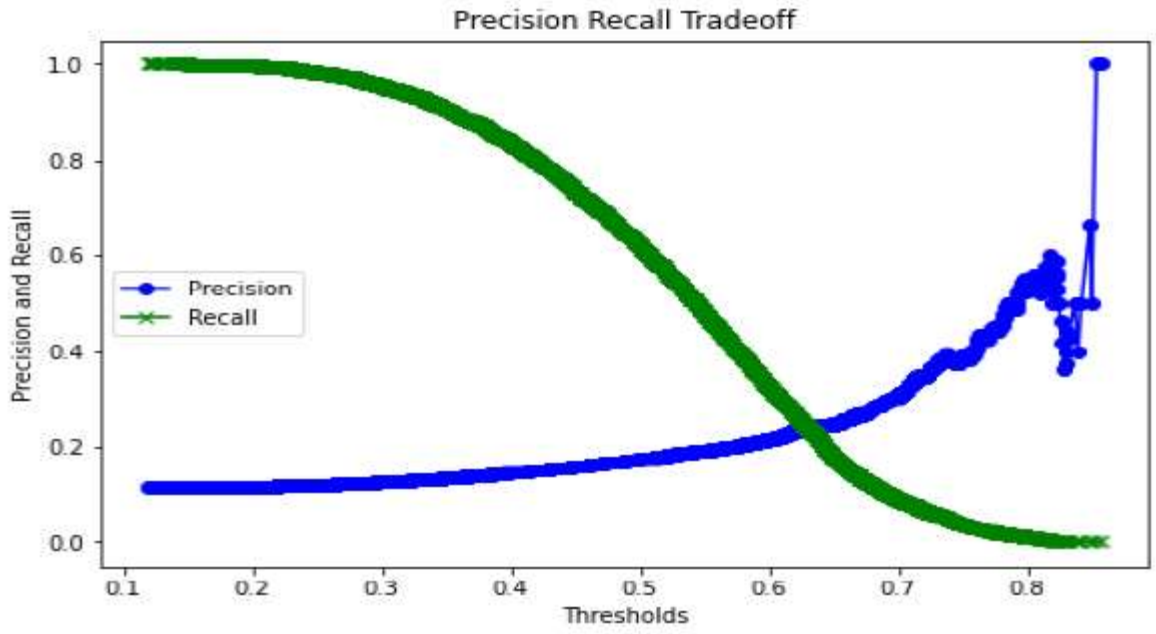
The above precision-recall curve plot is showing the precision/recall for each threshold for a GB model (orange) compared to a no skill model (blue).

Confusion Matrix:

```
array([[427,  48],
       [114, 135]], dtype=int64)
```

Above is the confusion matrix given by the final model on the test dataset.

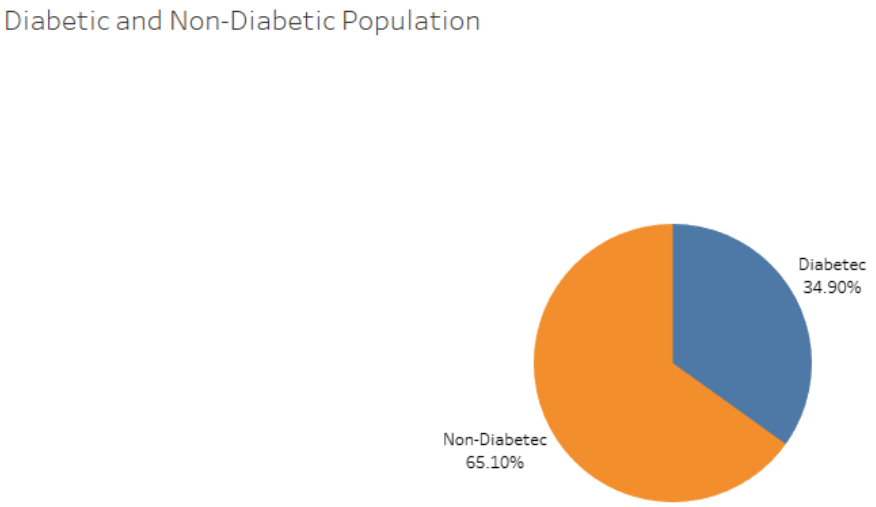
	precision	recall	f1-score	support
0	0.79	0.90	0.84	475
1	0.74	0.54	0.62	249
accuracy			0.78	724
macro avg	0.76	0.72	0.73	724
weighted avg	0.77	0.78	0.77	724



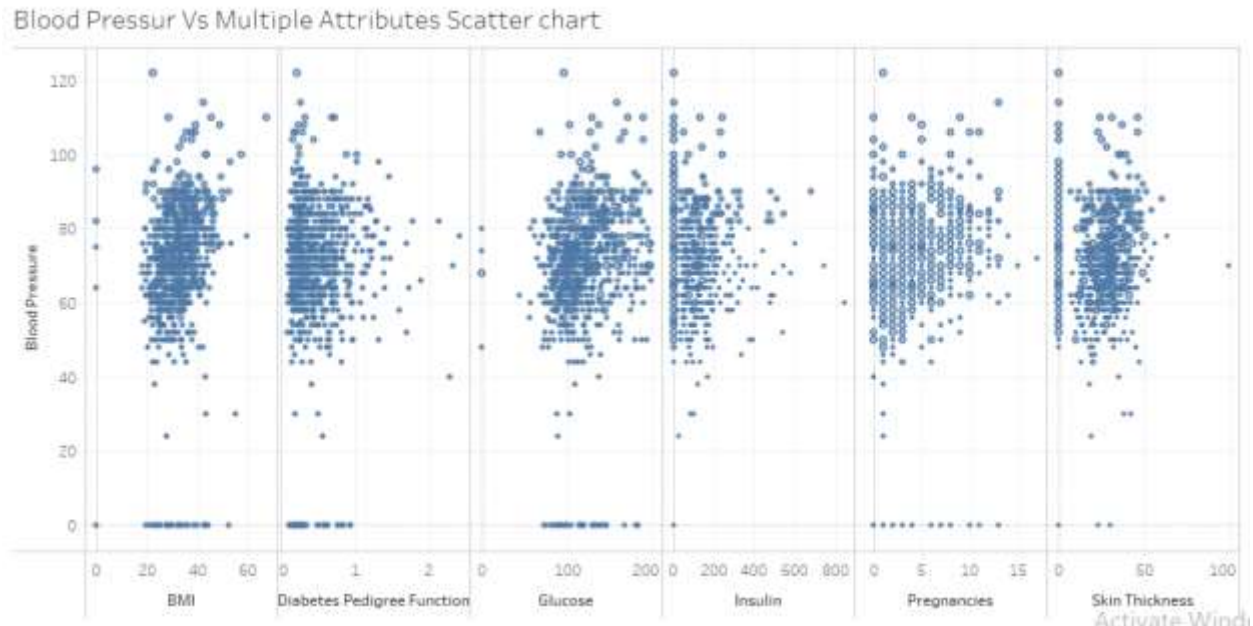
As we can see from performance curve above it also represents the Precision Recall Tradeoff in, as we increase the Recall, the precision decreases which means that if we want to reduce the number of False Negatives our False Positives will increase. Given that our primary metric is Recall, we have chosen the threshold that's giving us good recall while manageable precision.

Step 7: Create a dashboard in tableau by choosing appropriate chart types and metrics useful for the business. The dashboard must entail the following:

a) Pie chart to describe the diabetic/non-diabetic population

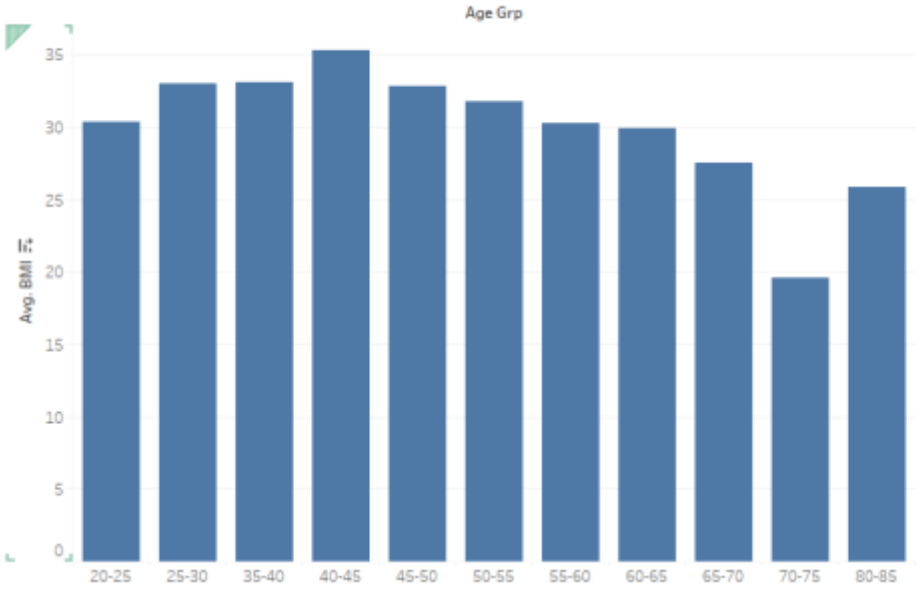


b) Scatter charts between relevant variables to analyse the relationships



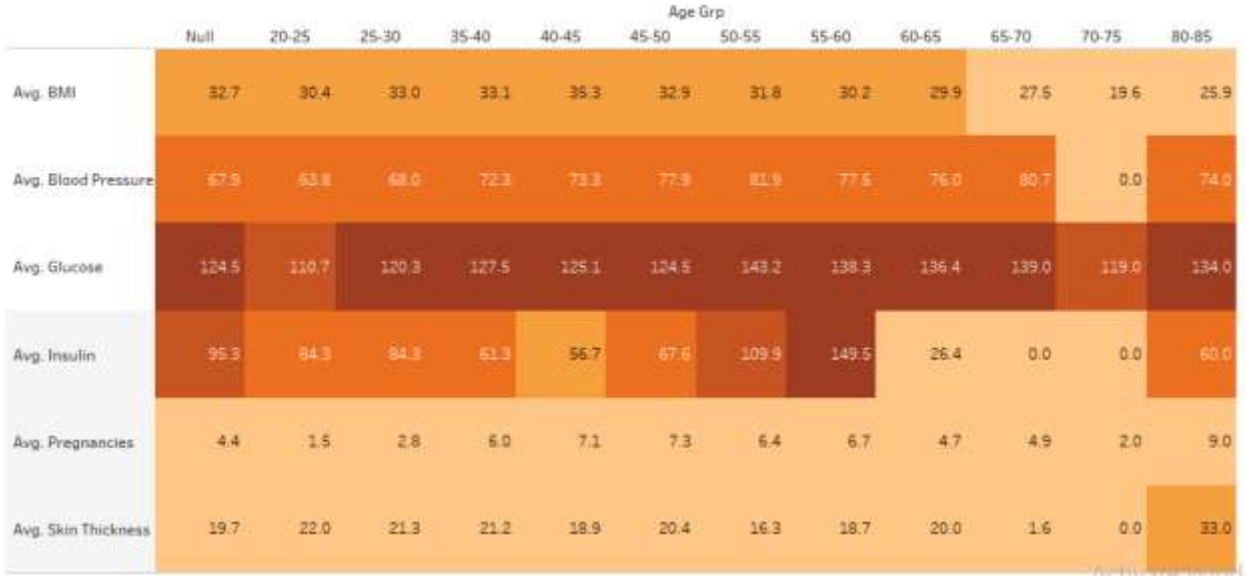
c) Histogram/frequency charts to analyse the distribution of the data

Age Vs BMI histogram with count



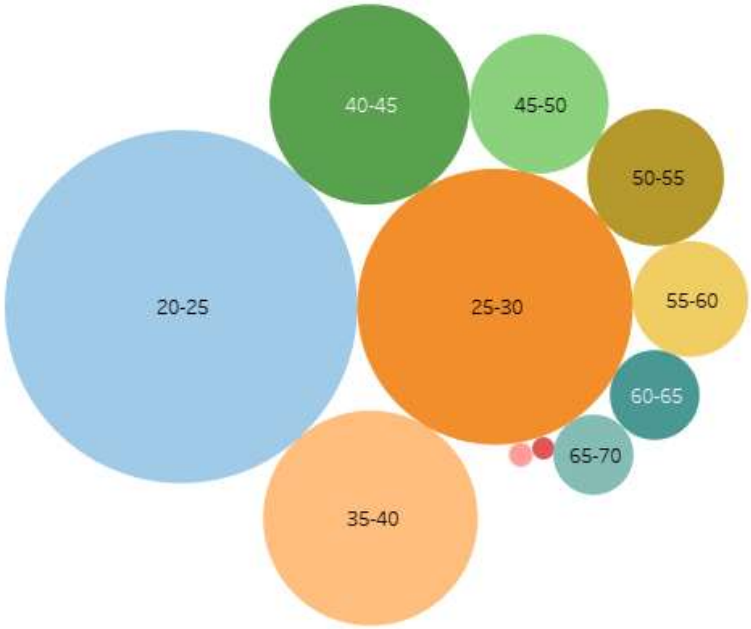
d) Heatmap of correlation analysis among the relevant variables

Age grp Vs Other Attribute Heatmap

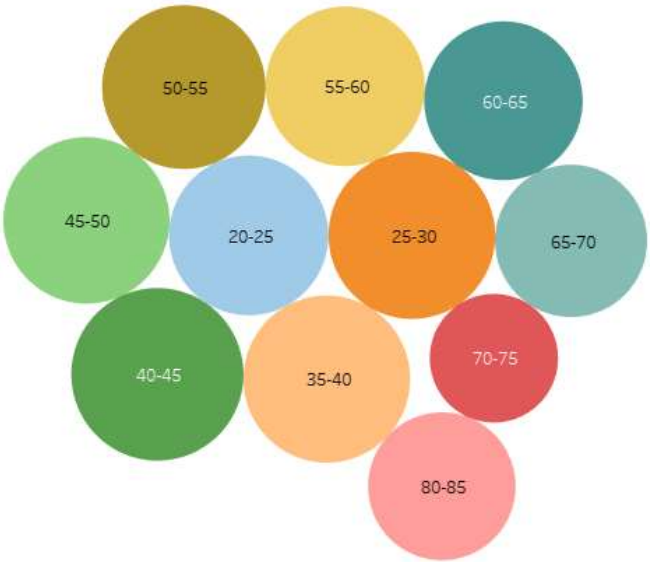


e) Create bins of Age values – 20-25, 25-30, 30-35 etc. and analyse different variables for these age brackets using a bubble chart.

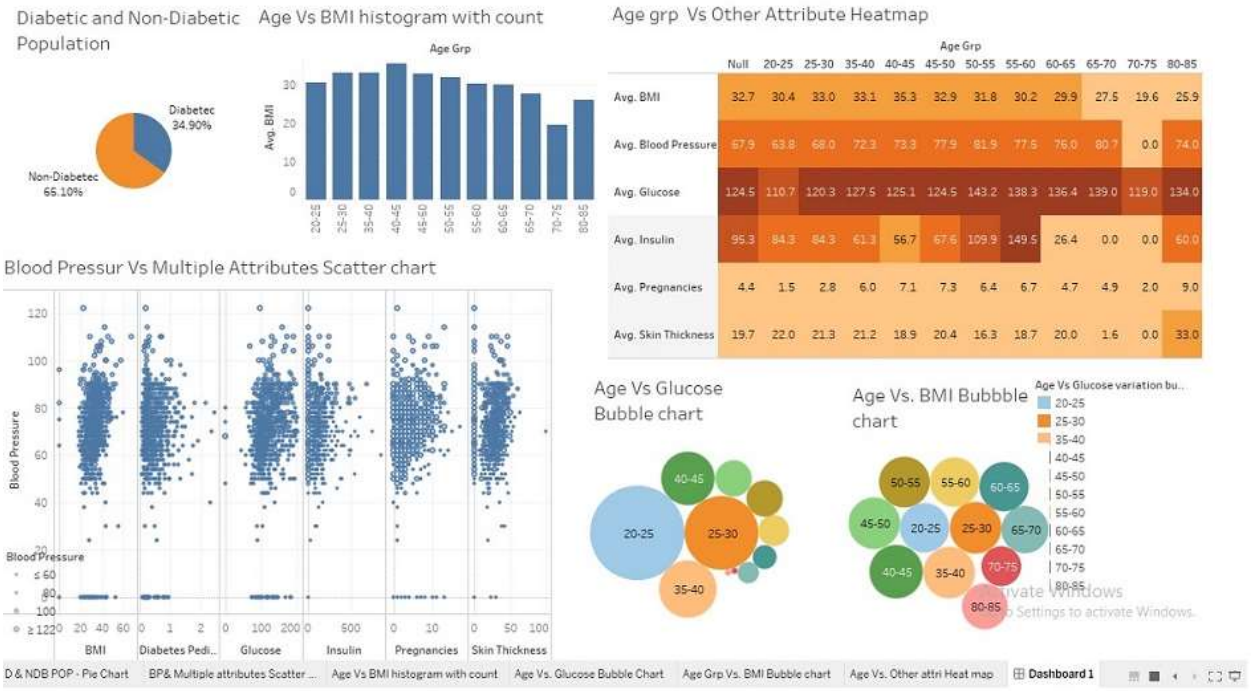
Age Vs Glucose Bubble chart



Age Vs. BMI Bubbble chart



Dashboard



Summary

- Building a machine learning-based classifier that predicts if a patient is diabetic or not, based on the information provided in the database.
- While building this predictor, initially common preprocessing steps such as feature scaling and imputing missing values are followed.
- Implemented various algorithm like Logistic Regression, Support vector machine, K – Nearest Neighbor, Decision tree, Gaussian Naive Bayes, Random forest and Gradient boosting algorithms, evaluated the performance measured using the accuracy score, comparing the performance between train and test data. Also parameter tuning improved using accuracy score, AUC.