Project : Health care

I. Project Overview:

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

Metrics:

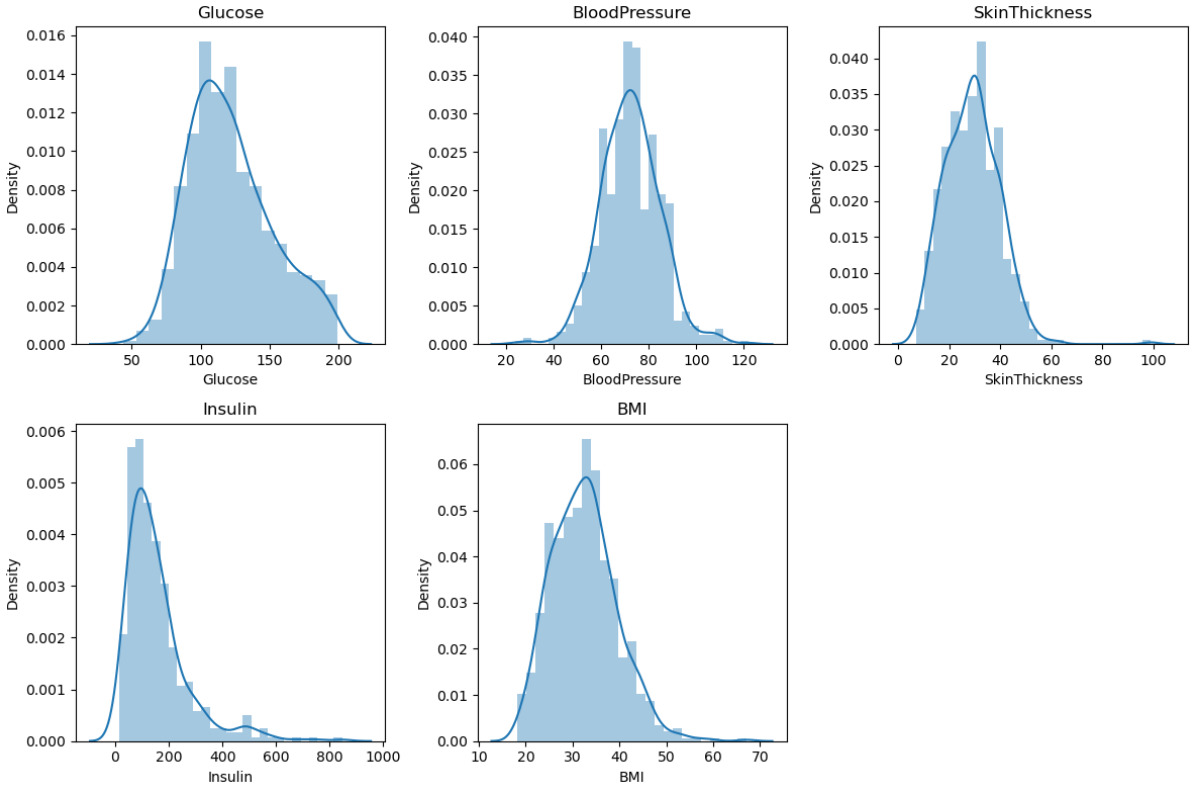
We'll evaluate the model based on the accuracy score of each classification model.

And for the model with highest accuracy score will evaluate these metrics such as confusion metrics, precision, recall, f1 score also will analyse AUC (ROC curve

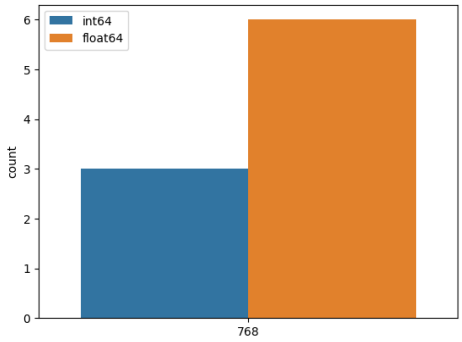
II. Analysis :

Data Exploration:

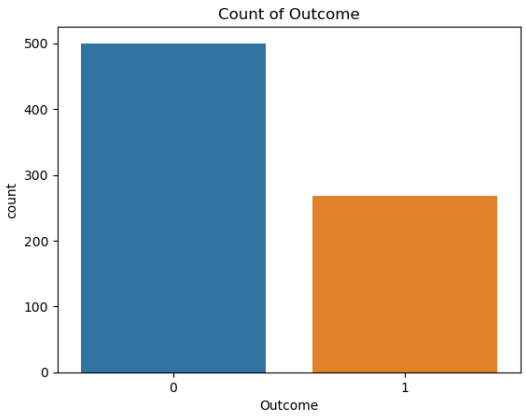
1. Performed the descriptive analysis , where we found that the variables Glucose, Blood pressure, Skin thickness ,Insulin and BMI have minimum value of zero which does not make sense and thus it indicates missing value. Hence we transformed these zero Values to Nan Values
2. Visually explored the data distribution for these variables using histogram. The data was not normally distributed hence imputed the missing values with its corresponding Median value



3. The data set has 3 integer datatype and 4 float data type with 768 patients observation

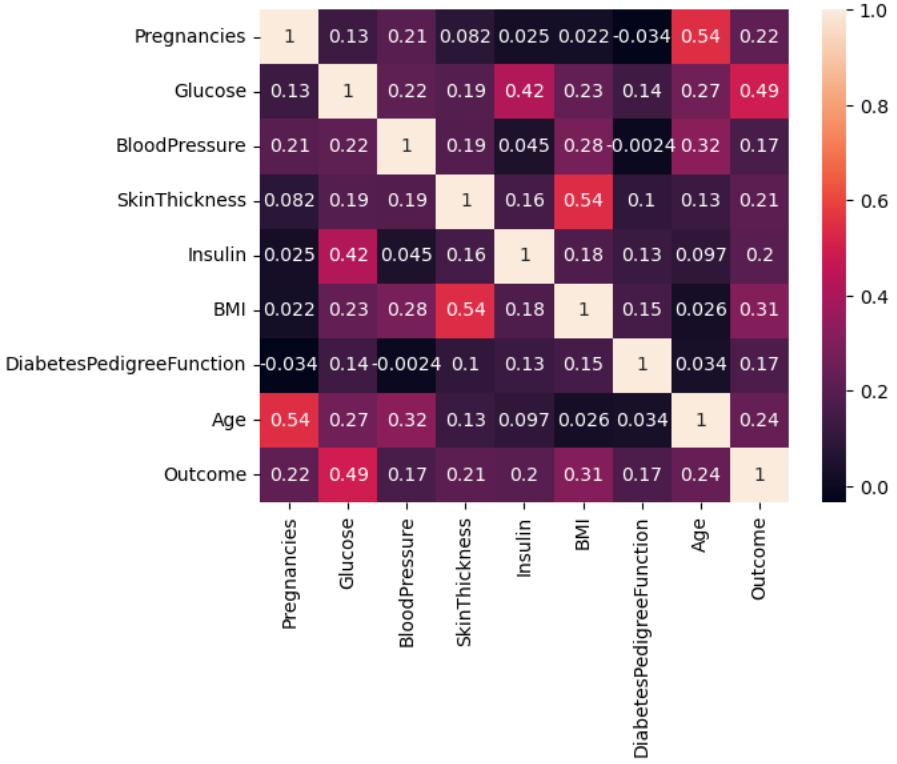


4. When we checked balance of target variable , out of 768 observations , 268 observations (i.e, 35%) are diabetic and 500 observations (i.e, 65%) are Non-Diabetic . Hence it is required to balance the data before training the model by using SMOTE method



5. Prepared the scatter plot between the pair of variables to understand the relations using the pair plot and found that all the independent variables are having weak linear correlation

6. Performed the correlation analysis and visually explored it with heatmap, where correlation coefficients varies between -0.034 to 0.54 which display weak correlation between variables



Algorithms and Techniques:

Strategy for Model building :

This is a classification problem , hence build the models using Logistic regression, KNN, Support vector Machine, decision tree, Random forest, Gradient Boost and Adaboost . Cross validation technique will help in lowering the bias and variances and using Grid search tune the hyperparameters.

Benchmark

For benchmarking chosen KNN Model. This gives the accuracy score of 77%

III. Methodology:

Data Pre-processing:

1. Split the data into X and Y variables to build the model
2. Scaled the X variables data using standard scalar
3. Applied the SMOTE to balance the data . while balancing considered the parameter sampling strategy as *Minority* to resample only the minority class i.e, diabetic
4. Split the data into Train and Test with 70:30 ratio

Model Building:

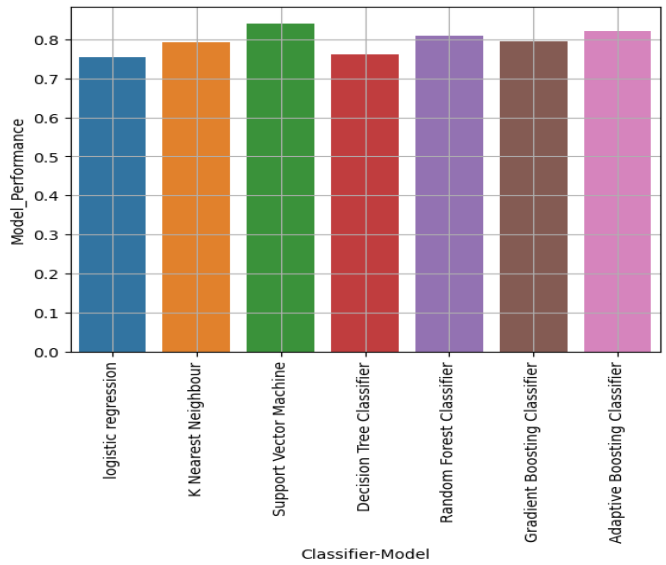
Built Models –

a. Logistic Regression,
b. K Nearest Neighbour (KNN) with n_neighbours as 4,
c. Support Vector Machine with the best params as C=10 and gamma =1,
d. Decision Tree with the best params as min_sample_leaf=7
e. Random Forest classifier with best params as min_sample _leaf=3 and n_estimators=150
f. Gradient boosting
g. Adaptive Boosting Classifier with best params as Decision tree Classifier (max_depth=5), n_estimators=200

IV. Results:

Accuracy scores for all the models built are as given below

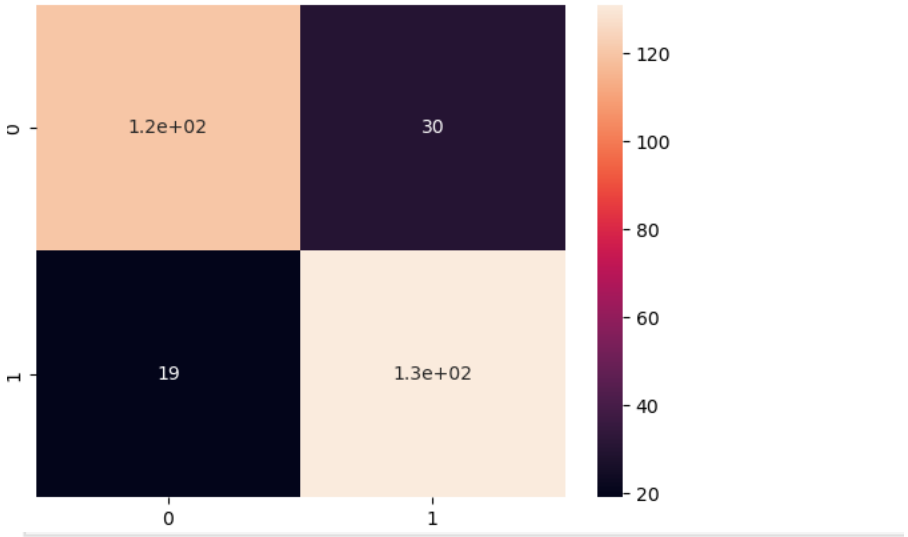
	Classifier-Model	Model_Performance
0	logistic regression	0.755
1	K Nearest Neighbour	0.792
2	Support Vector Machine	0.840
3	Decision Tree Classifier	0.760
4	Random Forest Classifier	0.810
5	Gradient Boosting Classifier	0.794
6	Adaptive Boosting Classifier	0.821



V. Conclusion:

With the above comparison of accuracy score Support Vector Machine has accuracy score of 83% hence chose that as final classifier. Then evaluated Confusion Matrix, Precision, recall and F1 score for the Model as shown below.

AUC Value measures the degree of separability in the classification model. For this model the AUC value is 0.91 which is between 0.9-1 which is considered as good score.



	precision	recall	f1-score	support
0	0.86	0.80	0.83	150
1	0.81	0.87	0.84	150
accuracy			0.84	300
macro avg	0.84	0.84	0.84	300
weighted avg	0.84	0.84	0.84	300

Tableau Dashboard:

The Dashboard have following visuals

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

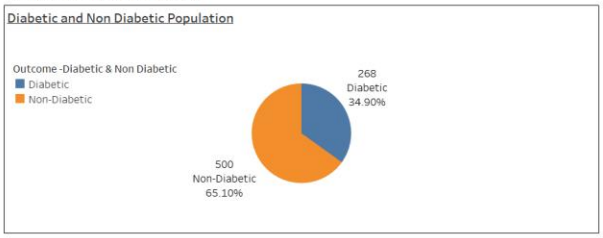
b) Scatter charts between relevant variables to analyse the relationships

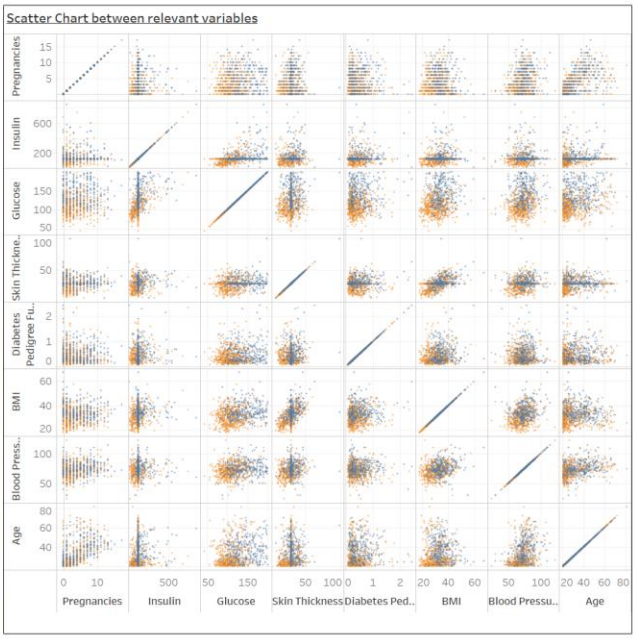
c) Histogram/frequency charts to analyse the distribution of the data

d) Heatmap of correlation analysis among the relevant variables

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.

Project: Health care (pg 1 of 2)

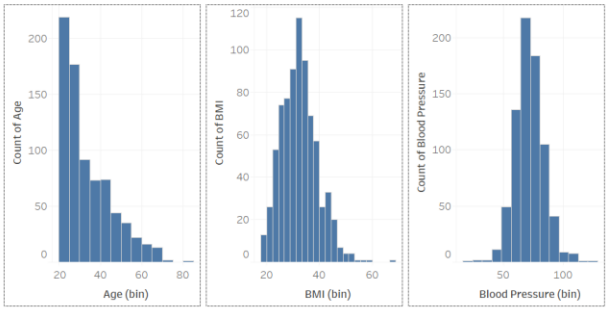


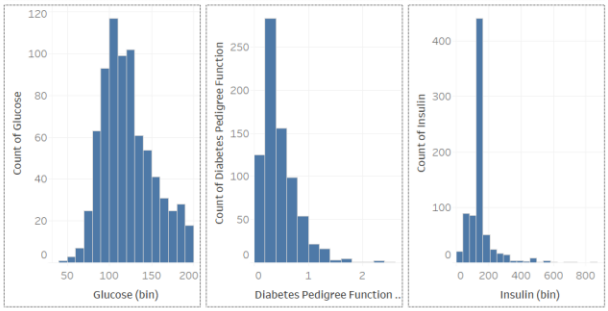


Dashboard 1 | Dashboard 2

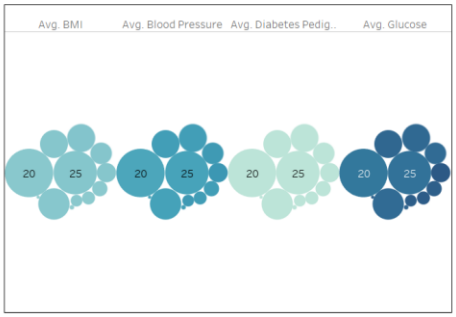
Project: Health care (pg 2 of 2)

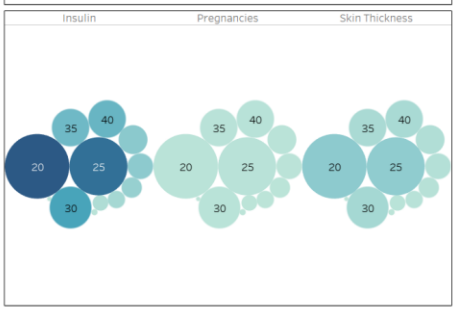
Histograms





Bubble chart





Dashboard 1 | Dashboard 2