Capstone Project: Healthcare PGP

Problem Statement

- NIDDK (National Institute of Diabetes and Digestive and Kidney Diseases) research creates knowledge about and treatments for the most chronic, costly, and consequential diseases.
- The dataset used in this project is originally from NIDDK. The objective is to predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset.
- Build a model to accurately predict whether the patients in the dataset have diabetes or not.

Solution:

1. Data Wrangling

Descriptive analysis: Descriptive analysis of the dataset has been carried out using barcharts. It has been observed that there are missing values in the dataset in the form of zeros.

Missing Value Treatment: The missing values (indicated by zeros) have been replaced with the mean values. Mean values are calculated for each paramenter by excluding zero values. Missing value treatment has not been performed on Outcome variable.

2. EDA

Count Plot for Data types: The dataset contained only two types of data type i.e. Integer and Floating points. A count (frequency) plot describing the data types and the count of variables. The given data set had 2 integer variables and 7 Float variables.

Plotting the count of outcomes by their value: Count plot for Outcome variable has been plotted. 268 patients are positive for Diabetes against 500 negative cases. The outcome field may be a little biased towards negative outcomes.

Scatter plot between variables: From the scatter plots we can say that other than Glucose and Blood Pressure and other features have a positive skewness.

Correlation Analysis: Correlation analysis performed on the dataset and visualised using Heat map. The heatmap shows that all variables have positive correlation with the target variable. blood pressure is least correlated and glucose is most correlated with target variable followed by BMI.

3. ML Models

Model Building strategy : Based on the EDA and data analysis, the dataset has been prepared for modelling using different ML algorithms. The Outcome variable is our Target variable and rest are Predictor variables. The outcome variable and Predictor variable has been segregated and the then the whole dataset has been sliced into Test and Train data in 3:1 ratio.

Then the Logistic, SVM, Decision Tree and KNN algorithms have been trained with the prepared dataset and the performance of the models have been judged based on performance etrics- Accuracy, Precision, Recall and F1 score

Comparison with KNN model: The performance metrics of Logistic, SVM and Decision Tree have been compared with KNN and visually presented with the help of bar charts.

<u>Analyzing Sensitivity, Specificity, AUC (ROC curve)</u>: The model demonstrates good performance with high specificity (ability to correctly identify negatives) and moderate sensitivity (ability to correctly identify positives). The ROC-AUC value of 0.8436 indicates strong discriminatory power, suggesting effective separation between positive and negative instances.

Data Reporting: Tableau dashboard has been created which presents following:-

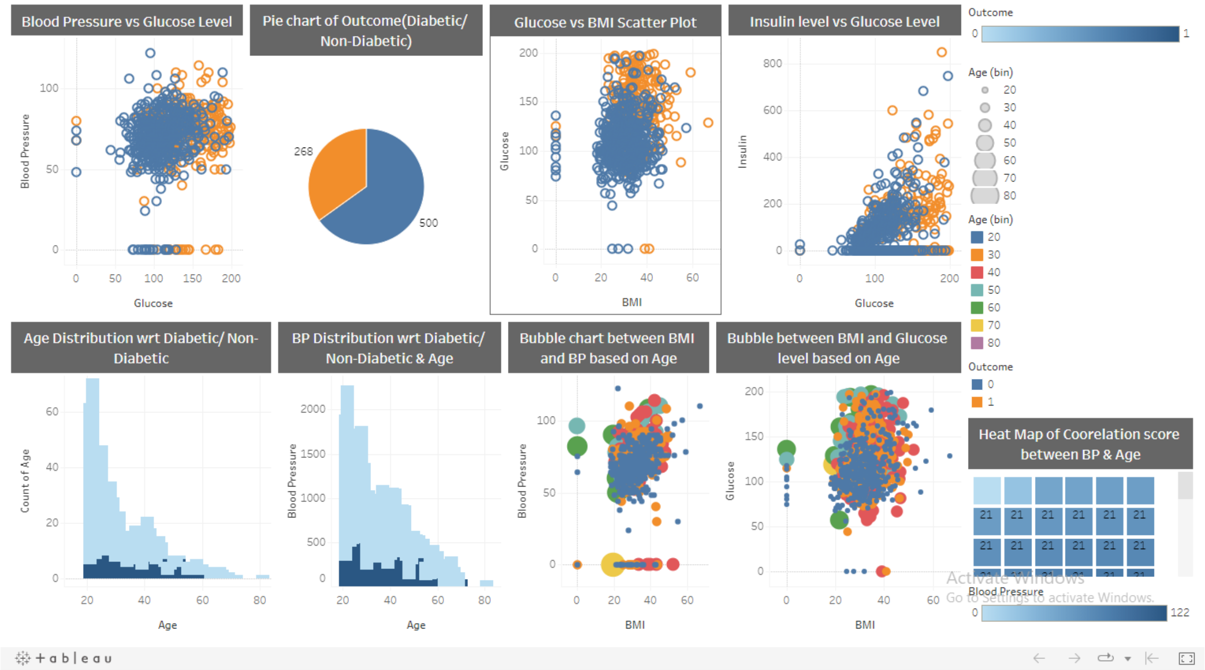
a. Pie chart to describe the diabetic or non-diabetic population

b. Scatter charts between relevant variables to analyze the relationships

c. Histogram charts to analyze the distribution of the data

d. Heatmap of correlation analysis among the relevant variables

e. Bubble charts between relevant variables wrt Age bins of size 10.



Link to Tableau Dashboard: [Capstone | Tableau Public]()