Healthcare Capstone

Domain: Health/Medicine

Objective:

- 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.

Analysis to be done:

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
   - Blood Pressure
   - Skin Thickness
   - 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.

Project Task: Week 2

Data Exploration:

1. Check the balance of the data by plotting the count of outcomes by their value. Describe your findings and plan future course of action.

2. Create scatter charts between the pair of variables to understand the relationships. Describe your findings.

3. Perform correlation analysis. Visually explore it using a heat map.

Project Task: Week 3

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. Compare various models with the results from KNN algorithm.

Project Task: Week 4

Data Modeling:

1. 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:

2. 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 or non-diabetic population

   b. Scatter charts between relevant variables to analyze the relationships

   c. Histogram or frequency charts to analyze the distribution of the data

   d. Heat map of correlation analysis among the relevant variables

   e. Create bins of these age values: 20-25, 25-30, 30-35, etc. Analyze different variables for these age brackets using a bubble chart.

Approach:

Project Task: Week 1

- Understand the dataset and identify the missing values.
- Remove missing values to make the data more useful, by assigning the mean of the entire variable to the missing values.
- Datatypes can be described by plotting a bar chart.

Project Task: Week 2

- A bar chart of outcome can be plotted to understand the balance.
- Scatter plots can be created between variables to understand the relationships.
- Explore variables using a heat map to identify best relationship.

Project Task: Week 3

- The Outcome variable is a categorical variable, hence KNN, Logistic Regression, Random Forest is best suited model for this data.
- We can apply Logistic Regression, Random Forest and compare the results with KNN

Project Task: Week 4

- An AUC (ROC curve) can be plotted and classification report can be created which is used to compare the 3 models' precision, recall and AUC curve score to determine the best model among them.
- A tableau dashboard can be created accommodating the requirements (Pie, scatter, frequency, bubble charts and heat map).