Data Science Capstone for Healthcare

**Problem Statement**

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**

The datasets consists of several 8 medical predictor variables and one target variable (Outcome) for 176 individual observations. Predictor variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and more.

**Data Exploration**

Check for shape of the data whether it shows 176 observations with 9 predictor variables

Check for descriptive statistics like min, max, std, IQVs etc for all predictive variables

Check for data types of data . Found Int and float data types

Create frequency chart for data types for visual representation

Check for null or empty predictive data. No empty data found

Check for zeros as data types are numeric. Found some observations with zeros and replaced with mean or median based on distribution of data

Created a frequency chart for outcome variables. Found two values for outcome

Create scatter plot using pair variable and correlation. There is some relationship between some pair variable and no relationship using correlation. Better to consider all variables

**Data Modeling**

To build model and decide model

Pre-processed the data using standardiser

Split the data into train and test data sets

Applied linear regression and check for the accuracy, confusion matrix

Applied KNN algorithm and check for accurarcy, confusion matrix

Found KNN is giving more accuracy compare to Linear Regression and applied different neighbour points to find best model

Created classification report ROC. As ROC is near to 1 , it is a best score to consider KNN at neighbour 25

**Data Reporting:**

Exploratory analysis done using Tableau with different kind of analysis

Created Pie chart. It described the diabetic or non-diabetic population. Around 60% for diabetic

Created Scatter charts to analyze between relevant variables to analyze the relationships. Found relationship by Age, Sugar level

Created Histogram charts to analyze the distribution of the data. Found by age wise outcome is diabetic

Created Heat map of correlation analysis among the relevant variables. There is less correlation between predictive variables vs. out come

Created bins of these age values: 20-25, 25-30, 30-35, etc. Analyzed different variables for these age brackets using a bubble chart. By Age, out come varied

<https://public.tableau.com/profile/venkatrao.m#!/vizhome/FirstDashboardonHealthcareproject/Dashboard1?publish=yes>