**Capstone-2 Healthcare Project**

DESCRIPTION

**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.

**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 more.

|  |  |
| --- | --- |
| **Variables** | **Description** |
| Pregnancies | Number of times pregnant |
| Glucose | Plasma glucose concentration in an oral glucose tolerance test |
| BloodPressure | Diastolic blood pressure (mm Hg) |
| SkinThickness | Triceps skinfold thickness (mm) |
| Insulin | Two hour serum insulin |
| BMI | Body Mass Index |
| DiabetesPedigreeFunction | Diabetes pedigree function |
| Age | Age in years |
| Outcome | Class variable (either 0 or 1). 268 of 768 values are 1, and the others are 0 |

**Project Task: Week 1**

**Data Exploration:**

1. Performed 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

• BloodPressure

• SkinThickness

• Insulin

• BMI

2. Visuallised explored these variables using histograms. Treat the missing values accordingly.

3. There are integer and float data type variables in this dataset.I have Created a count (frequency) plot describing the data types and the count of variables.

**Project Task: Week 2**

**Data Exploration:**

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

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

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

**Project Task: Week 3**

**Data Modeling:**

1. Done with Devise strategies for model building. It is important to decide the right validation framework. Express your thought process.

2. Applied an appropriate classification algorithm to build a model. Compare various models with the results from KNN algorithm.

**Project Task: Week 4**

**Data Modeling:**

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

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

b. Created Scatter charts between relevant variables to analyze the relationships

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

d. Created Heatmap of correlation analysis among the relevant variables

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