**Healthcare PGP – Project 2**

DESCRIPTION

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. Perform descriptive analysis. Understand the variables and their corresponding values. Onthe columns below, a value of zero does not make sense and thus indicates missing value:

\*Glucose

\*BloodPressure

\*SkinThickness

\*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.

**Data Exploration:**

4.Check the balance of the data by plotting the count of outcomes by their value. Describeyour findings and plan future course of action.

5.Create scatter charts between the pair of variables to understand the relationships.

Describe your findings.

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

**Project Task: Week 2**

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.

3. Compare various models with the results from KNN algorithm.

4. 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.