Diabetes Predictions

Project by Gaurav Kumar Rai

Reference:

[Aditya-Mankar/Diabetes-Prediction: Predict Diabetes using Machine Learning. (github.com)](https://github.com/Aditya-Mankar/Diabetes-Prediction)

GitHub URL:

[GauravRai1512/UCDPA\_GauravKumarRai (github.com)](https://github.com/GauravRai1512/UCDPA_GauravKumarRai)

Abstract:

This has been very common health issue with age >45 due to our unbalanced unhygienic lifestyle. To predict whether the patient has diabetes or not based on the dataset featured from the **National Institute of Diabetes** and **Digestive and Kidney Disease** become very easy through Machine Learning. Dataset has different columns like Glucose, Blood Pressure, Skin Thickness, Insulin, Age & BMI which will help us to predict whether the patient has diabetes or not. Prediction will be done through Machine Learning.

This diabetes problem could be a classic example of binary classification and used KNN model to predict the accuracy. The datasets consist of feature variables and one target variable, **Outcome**.

We have loaded the csv file which we took from **Kaggle** and then started **Exploratory Data Analysis** (EDA) through checking head (), info (), shape () function to get the actual rows and columns details.

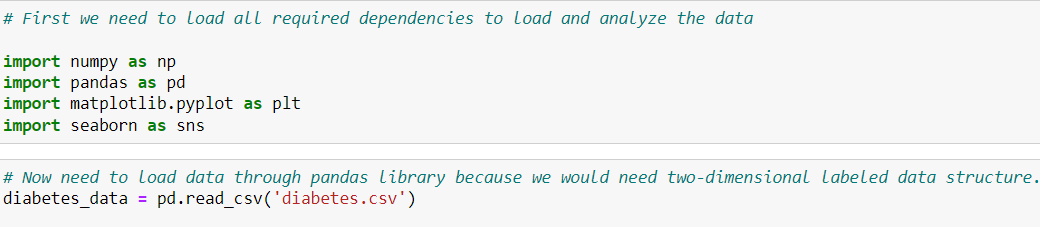
Then Data Preparation where Data cleaning process will come into the picture where we must check and clean all possible data which has 0 value NaN and duplicate values.

We build a machine learning model which is **KNN classifier model** of supervised learning to train the data, fit the data, predict the data whether the patients in the dataset have diabetes or not.

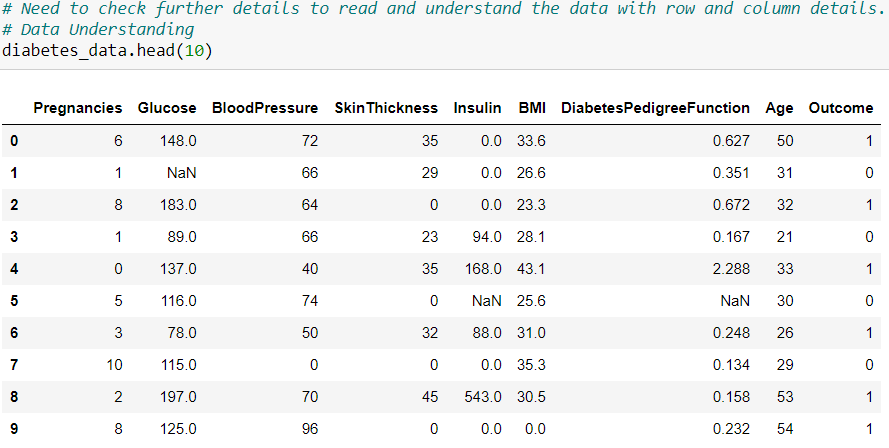
Introduction:

Implementation Process:

1. Loading important libraries which will help us to execute the all-important functions.



1. Data Understanding is also an important and crucial part to analyze the data





**1.** **Data Collection**:

We have used Kaggle ([93% test acc- data cleaning+sampling+ ANN | Kaggle](https://www.kaggle.com/code/vuppalaadithyasairam/93-test-acc-data-cleaning-sampling-ann/notebook)) to get the data set and little bit modified to feature NaN & Duplicate scenarios . downloaded csv file from mentioned linked and saved under project folder with the name Diabetes.csv with total **769 rows** and **9 columns.**

**2. Data Cleaning:**

As mentioned, few rows and columns has NaN value also few rows were duplicates as well so cleaning of such data would be required. NaN values were very less as less than 10% so have used dropna() function of pandas to drop all NaN rows not column as dropping column will completely reduce data for analysis so have not used axis=1 feature to drop the columns.

TO remove duplicates first we have identified how many duplicates were present

After cleaning