

Data Science

Project Report

Project Title:  **Predicting Diabetes**

Supervisor: Dr. Zeeshan Ahmed

Submitted By: Shahrukh Bashir

Student ID: k164055

Section: GR-2

**Problem Statement:**

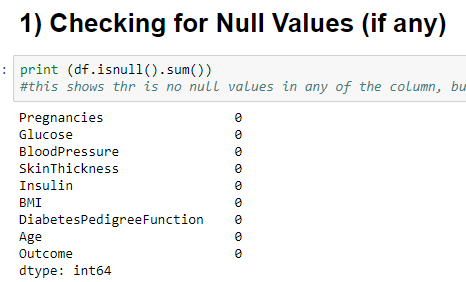
Diabetes is one of the deadliest diseases in the world. It is not only a disease but also creator of different kinds of diseases like heart attack, blindness etc. The normal identifying process is that patients need visit a diagnostic center, consult their doctor, and sit tight for a day or more to get their reports. So, the objective of this project is to identify whether the patient has diabetes or not based on diagnostic measurements. The dataset used has been obtained from UCI Machine Learning Repository having 769 record of Female Patients exclusively.

**Solution:**

As we have to classify the data into patients having diabetes or not, the best method which can be used is decision tree (using gini index), K-Nearest Model and Naives Byes because in this, the dataset is divided into training and testing data. Further we can easily classify and predict the outcome using nodes and internodes .The rules derived will be helpful for doctors to identify patients suffering from diabetes. Further predicting the disease early leads to treating the patient before it becomes critical.

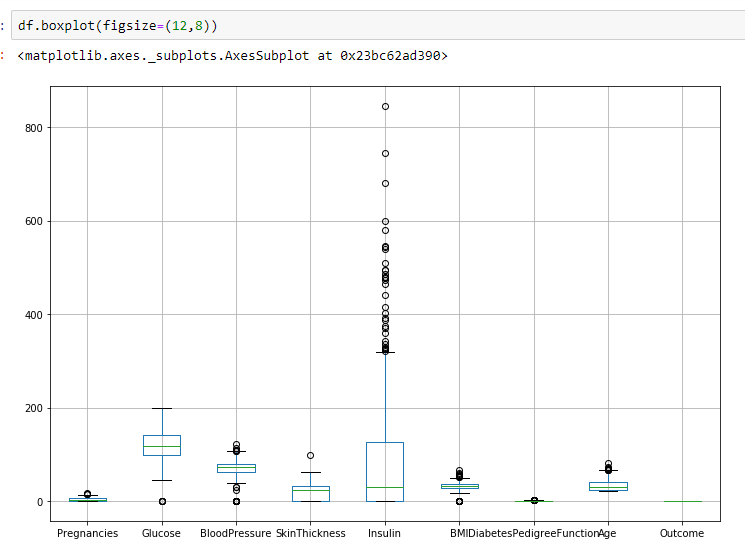
**Data Cleaning Process:**

**Null Values:** After data visualization, using a python function for finding null values, I found that there exists no null values in the dataset.



**Dealing with Zeros**: I made a method which will return the mean value of the coumn pass as parameter, all columns are traversed and if any zero is found that is replaced with the mean value.

**Dealing with Outliers:** Used Boxplot to identify the outliers.



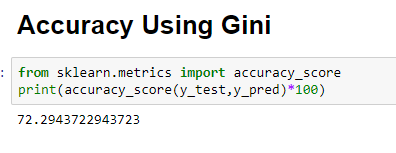
**Spliting of Dataset**:

**Training dataset**: 70%

**Test Dataset**: 30%

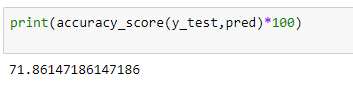
**Model Fitting**:

1. Gini:



1. **KNN**:

Accuracy Using KNN



1. **Naïve Byes:**

Accuracy Using Naïve Byes:

