Student Name: Muhammad Maqsud Hussain
Course Cod: CIND 820

Proposal

Diabetes prediction using Machine Learning

Context:

Diabetes mellitus refers to a group of diseases that affect how the body uses blood glucose. Glucose is an important source of energy for the cells that make up the muscles and tissues. It's also the brain's main source of fuel. Diabetes can lead to excess sugar in the blood. Too much sugar in the blood can lead to serious health problems, such as heart disease, vision loss, and kidney diseases etc. As per the International Diabetes Federation, 537 million adults (20-79 years) are living with diabetes and there are more than 5.7 million Canadians living with diabetes.

The Dataset:

For the project, the **Diabetes prediction dataset** is chosen which is a collection of medical and demographic data from patients, along with their diabetes status (positive or negative). The data includes features such as *age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level,* and *blood glucose level*. There are 10,000 instances/rows in this dataset. The dataset can be retrieved from
https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

Problem:

This dataset can be used to build machine learning models to predict diabetes in patients based on their medical history and demographic information. Some probable research questions can be as follows:

- Which factors are highly correlated with positive diagnosis of diabetes?
- Does probability of being a diabetic varies with age and/or gender?
- How smoking history is related to the probability of becoming a diabetic patient in future?
- Is blood sugar level of an obese/overweight person always higher than a person with normal BMI?
- How accurately can diabetes be predicted, and which model works best?

Techniques:

Predictive modeling techniques such as logistic regression, k-nearest neighbor regression, decision tree classifier, random forest classifier support vector model classifier etc. can be used in this project to solve the research problems. A comparison among these models can be made to find the accuracy of each model.

Tools:

It is planned that python will be used predominantly for performing the analysis. RStudio may be used to draw some statistical inference. Finally, Tableau can be used to visualize the findings of the analysis.