**BISI CST2101 Project - Diabetes Analysis**

**Context**

Diabetes is one of the most frequent diseases worldwide and the number of diabetic patients are growing over the years. The main cause of diabetes remains unknown, yet scientists believe that both genetic factors and environmental lifestyle play a major role in diabetes.

A few years ago research was done on a tribe in America which is called the Pima tribe (also known as the Pima Indians). In this tribe, it was found that the ladies are prone to diabetes very early. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients were females at least 21 years old of Pima Indian heritage.

**Objective**

Here, we are analysing different aspects of Diabetes in the Pima tribe by doing Exploratory Data Analysis.

**Data Dictionary**

The dataset has the following information:

\* Pregnancies: Number of times pregnant

\* Glucose: Plasma glucose concentration over 2 hours in an oral glucose tolerance test

\* Blood Pressure: Diastolic blood pressure (mm Hg)

\* Skin Thickness: Triceps skin fold thickness (mm)

\* Insulin: 2-Hour serum insulin (mu U/ml)

\* BMI: Body mass index (weight in kg/(height in m)^2)

\* DiabetesPedigreeFunction: A function which scores likelihood of diabetes based on family history.

\* Age: Age in years

\* Outcome : Class variable (0: person is not diabetic or 1: person is diabetic)

**Models Used to train and test the data**

* Logistic Regression
* RandomForestClassifier

**Steps used to create the notebook**

1. For creating the notebook I opened Anaconda software.
2. After opening anaconda I got an dashboard and then I clicked on jupyter notebook.
3. I started writing the context and objective.
4. Then after these things I started writing the code and started from uploading the csv file in the notebook.
5. Then I checked the last 5 records of the dataset using Tail command.
6. Then I checked first 5 records of the dataset using Head command.
7. After these things the dimension of the Pima data frame using Pima. Shape.
8. Then the next thing I did is I checked the size of the Pima data frame using Pima. size.
9. After doing these steps I checked the data types of all the variables in the data set using Pima . dtypes .

So there are so many steps I did in this project as you can easily see and understand in my notebook.

**Conclusion**

The performance of a model relies on its accuracy, with a more accurate model having better performance. In this case, the Random Forest classifier has a higher accuracy of 79%, while the Logistic Regression model has an accuracy of 76%.