



## DATA ANALYSIS

COURSE PRESENTER  
(DR. Omainah)

### Pima Indians Diabetes Database

| STUDENT     | ID        |
|-------------|-----------|
| SHAHAD AMER | 444005703 |
| AREEJ TALEB | 444002403 |

DEPARTMENT OF (INFORMATION SCIENCE-DATA SCIENCE)  
COLLEGE OF COMPUTER AND INFORMATION SYSTEMS  
UMM AL-QURA UNIVERSITY

2024

# INTRODUCTION

The Pima Indians Diabetes Database is a pivotal dataset widely utilized in diabetes research and predictive modelling. The dataset consists of several medical predictor variables, including the number of pregnancies a patient has had, Body Mass Index (BMI), insulin levels, age, and glucose concentration. These variables are instrumental in assessing the risk factors associated with diabetes. The target variable, known as "Outcome," indicates whether the individual has been diagnosed with diabetes (1) or not (0), serving as the primary focus of predictive modeling efforts.

# OBJECTIVE

The main objective of this project is to build a model that accurately predict whether the patients in the dataset have diabetes or not based on certain diagnostic measurements included in the dataset.

# Data preparation

## Import Libraries

Import the most important libraries that facilitate the analysis process.

## Read the File

Load the dataset from a specified file.

## Display Initial Data

Display the first few rows of the Data Frame to understand its structure.

## Identify Missing Values

Identify any zero values in the dataset, which may indicate missing data.

## Replace Zero Values

Replace the zero values in specific columns with missing values (pd.NA).

## Fill Missing Values

- Fill missing values with the mean for the columns **Glucose** and **Blood Pressure**.
- Fill missing values with the median for the columns **Skin Thickness**, **Insulin**, and **BMI**.

This approach helps ensure that analyses or models built on this data are more robust and informative.

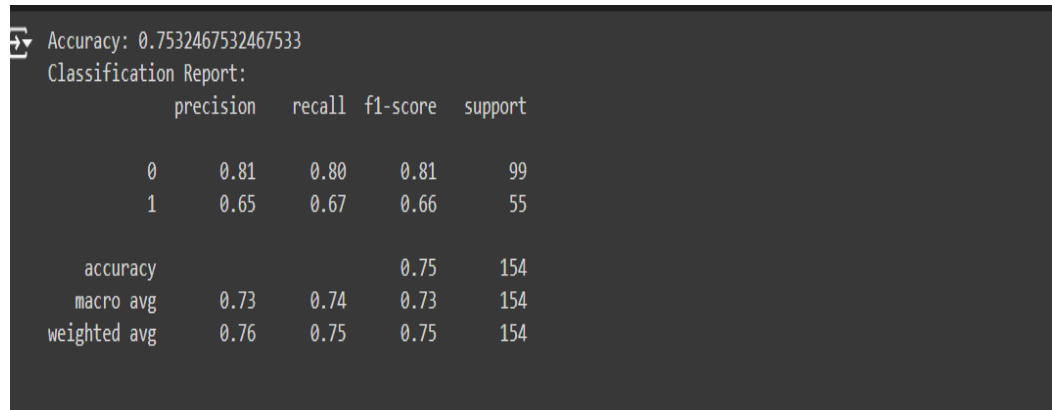
## Naive Bayes model

In this analysis, we implemented a **Gaussian Naive Bayes model** to predict diabetes outcomes using the Diabetes dataset, after splitting the dataset into training and testing sets (with 80% for training and 20% for testing), we standardized the features to ensure the model's effectiveness.

The model achieved an accuracy of approximately **75.32%**, indicating that it correctly classified about three-quarters of the test samples. This level of accuracy reflects the model's ability to distinguish between patients with and without diabetes based on the provided diagnostic measurements.

The classification report provides further insight into the model's performance:

- **Precision:** For class 0 (non-diabetic), the precision was **0.81**, suggesting that when the model predicted a patient as non-diabetic, it was correct 81% of the time. However, for class 1 (diabetic), the precision was lower at **0.65**, indicating some challenges in correctly identifying diabetic patients.
- **Recall:** The recall for class 0 was **0.80**, meaning the model correctly identified 80% of actual non-diabetic cases. For class 1, the recall was **0.67**, which shows that the model missed about 33% of the diabetic cases.
- **F1-Score:** The F1-score, which balances precision and recall, was **0.81** for class 0 and **0.66** for class 1, further highlighting the model's stronger performance in predicting non-diabetic cases.



```
Accuracy: 0.7532467532467533
Classification Report:

```

|              | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0            | 0.81      | 0.80   | 0.81     | 99      |
| 1            | 0.65      | 0.67   | 0.66     | 55      |
| accuracy     |           |        | 0.75     | 154     |
| macro avg    | 0.73      | 0.74   | 0.73     | 154     |
| weighted avg | 0.76      | 0.75   | 0.75     | 154     |

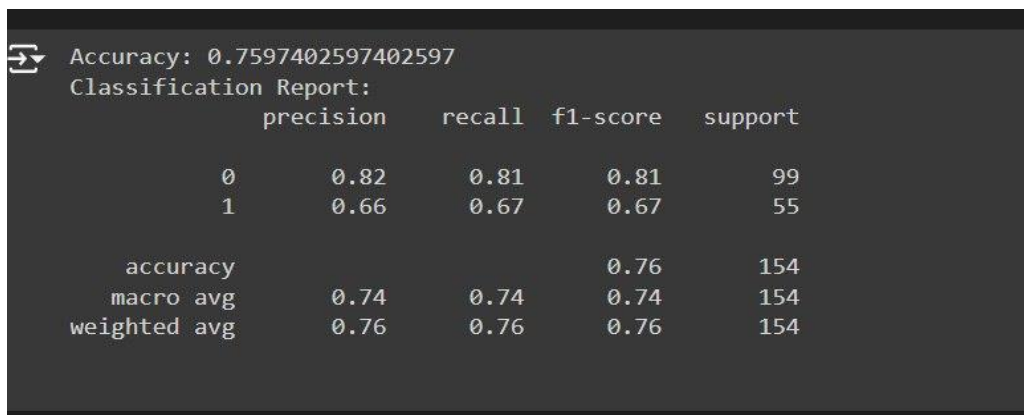
## Random Forest Model

In this phase, we applied the **Random Forest Classifier** to assess its performance in predicting diabetes outcomes using the Diabetes dataset.

The model achieved an accuracy of **75.97%**, indicating that it correctly classified approximately 76% of the samples in the test set, this accuracy reflects the model's ability to distinguish between patients with diabetes and those without, based on the available diagnostic measurements.

The classification report below provides further details about the model's performance:

- **Precision:** For class 0 (non-diabetic), the precision was **0.82**, meaning the model was accurate 82% of the time in predicting non-diabetic patients. For class 1 (diabetic), the precision was **0.66**, indicating that the model faced challenges in accurately identifying diabetic cases.
- **Recall:** The recall for class 0 was **0.81**, suggesting that the model correctly identified 81% of the non-diabetic cases. For class 1, the recall was **0.67**, indicating that the model missed about 33% of the diabetic cases.
- **F1 Score:** The F1 score, which reflects the balance between precision and recall, was **0.81** for class 0 and **0.67** for class 1. This shows strong performance for the non-diabetic class while highlighting the need for improvement in the diabetic class.

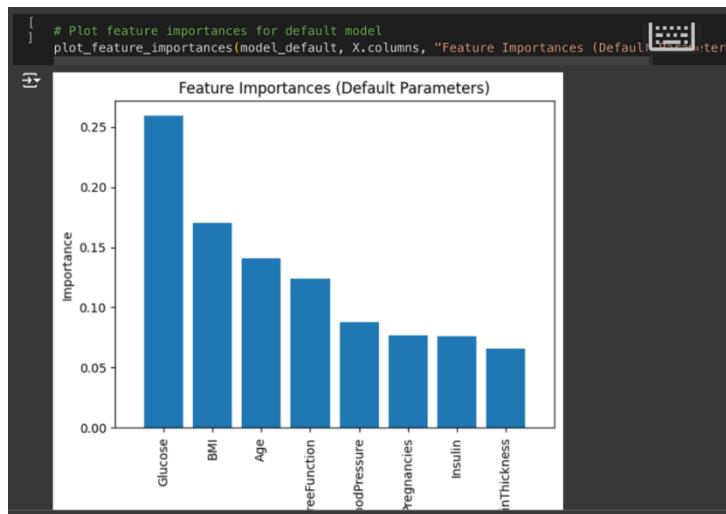
A screenshot of a terminal window showing a classification report. The report includes accuracy, precision, recall, f1-score, and support for two classes (0 and 1), as well as macro and weighted averages. The background is dark with light-colored text.

```
Accuracy: 0.7597402597402597
Classification Report:

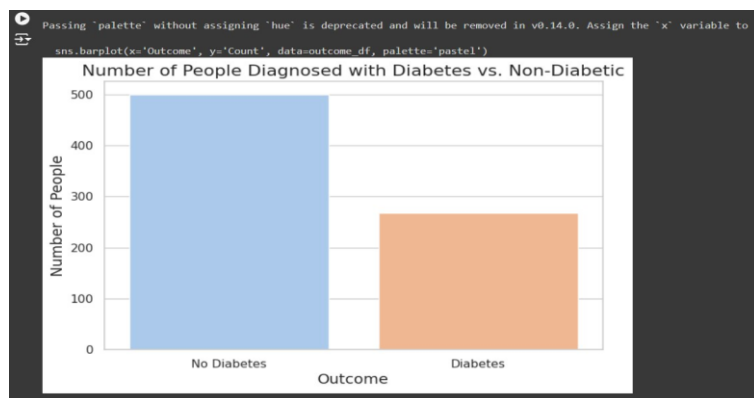
```

|              | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0            | 0.82      | 0.81   | 0.81     | 99      |
| 1            | 0.66      | 0.67   | 0.67     | 55      |
| accuracy     |           |        | 0.76     | 154     |
| macro avg    | 0.74      | 0.74   | 0.74     | 154     |
| weighted avg | 0.76      | 0.76   | 0.76     | 154     |

## Visualisation



From the plot we can see the most important features, which they are BMI and Glucose, they Are the most common reasons for diabetes.



From the graph we can see the number of people who have been diagnosed with diabetes and those who have not been diagnosed with diabetes.