جامعة أم القرى
UMM AL-QURA UNIVERSITY

Data Analysis 2

COURSE PRESENTER
(DR. Omima Fallatah)

SUBMITTED BY:

Name	ID
Jood Mohmmed Algarni	444005790
Noura Nawar Alhuthali	444001743

DEPARTMENT OF (DATA SCIENCE)
COLLEGE OF COMPUTERS
UMM AL-QURA UNIVERSITY

Data

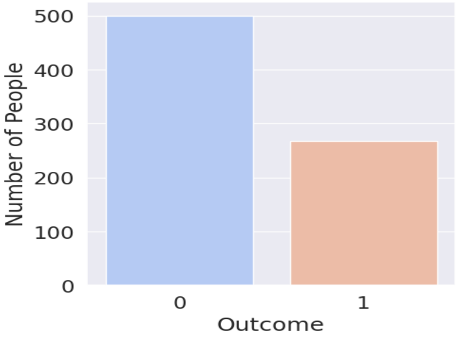
We will use the Pima indian diabetes dataset. The data is available at Kaggle and can be downloaded from here. The datasets nine columns: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age and Outcome.

Defining Objectives

The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset and maximizing the model's ability to correctly classify cases (diabetes patients)

Data Exploration

First we use some method to print the name of the columns and to know the type all columns was (int64) expect two column where (float64) then we try to know the actual value of how many people have diabetes or not by using value_counts() and use plot to show that
People who do not have diabetes: 500
People who have diabetes: 268



We also visualize distributions of features to help us understand data characteristics, identify patterns and outliers, and guide our modelling decisions.

Data Processing

After loading the data and displaying it to ensure that all columns and rows are present, it was confirmed whether the data contained any missing values, and the result was that there were no missing values. However, some columns contained cells with the value zero, which is not logically acceptable. Therefore, the zero values were replaced with null. Subsequently, the null values in some columns were replaced with the mean, while in others, they were replaced with the median.

Data Splitting

We need to separate the columns into target (**Outcome**) and features variables (**Eight features**). **X** contains **features variable** which is training data , and **y** contains **target variable** which is the test data.

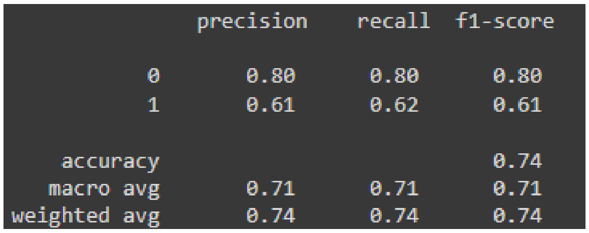
(<u>Naive Bayes</u>)

Choosing Models

Naive Bayes is a classification algorithm, which uses Bayes theorem of probability for prediction of unknown class we will use this algorithm to predict whether or not the patients in the dataset have **diabetes** or **not.**

Evaluating Model

To evaluate the model, we will check the (classification_report)using actual and predicted values



```
              precision    recall  f1-score

           0       0.80      0.80      0.80
           1       0.61      0.62      0.61

    accuracy                           0.74
   macro avg       0.71      0.71      0.71
weighted avg       0.74      0.74      0.74
```

1- **Precision:**

- Measures the ratio of correct predictions among all positive predictions made by the model.
- For class 0: 0.80 (80% of the instances predicted as class 0 were true).
- For class 1: 0.61 (61% of the instances predicted as class 1 were true).

2- **Recall:**

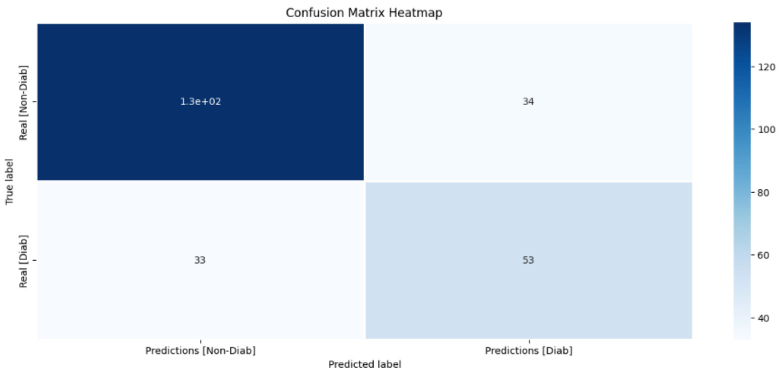
- Measures the ratio of correct predictions among all actual cases present in the data.
- For class 0: 0.80 (80% of actual class 0 instances were correctly identified).
- For class 1: 0.62 (62% of actual class 1 instances were correctly identified and missed 38%).

3- **F1-Score:**

- A metric that combines precision and recall, particularly useful when dealing with unbalanced classes.
- For class 0: 0.80 (indicating a balanced performance).
- For class 1: 0.61 (lower than class 0, indicating lower performance).

overall accuracy of 74% indicates that while the model is reasonably effective, there is an opportunity to enhance its predictive capabilities and that's why we are traying a new model

Confusion Matrixes used that shows the number of correct and incorrect forecasts for each category This matrix is represented by a heat map.
the model performs quite well in identifying non-diabetic but has some difficulty accurately classifying diabetic as indicated by the false positives and false negatives, so we need to try different algorithm to improve the results.



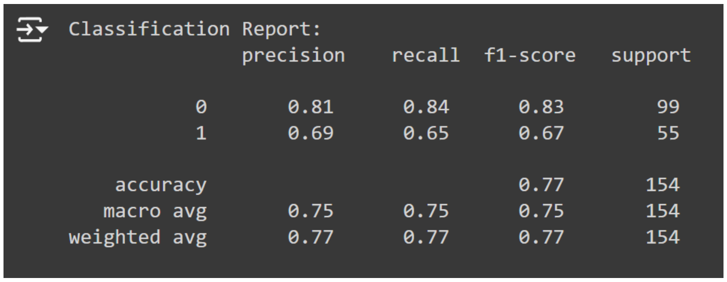
(**<u>Logistic Regression</u>**)

Choosing Models

Logistic Regression is a statistical technique used to estimate the likelihood of a certain event occurring, and it is often applied in fields such as medicine and social sciences.
By using logistic regression, it is possible to infer the impact of each variable on the likelihood of developing diabetes, aiding in understanding the contributing factors of the disease.

Evaluating Model

To evaluate the model, we will check the (classification_ report) using actual and predicted values



```
Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.84      0.83        99
           1       0.69      0.65      0.67        55

    accuracy                           0.77       154
   macro avg       0.75      0.75      0.75       154
weighted avg       0.77      0.77      0.77       154
```

1-Precision:

- Measures the ratio of correct predictions among all positive predictions made by the model.
- For class 0: 0.81 (81% of the predictions were correct).
- For class 1: 0.69 (69% of the predictions were correct).

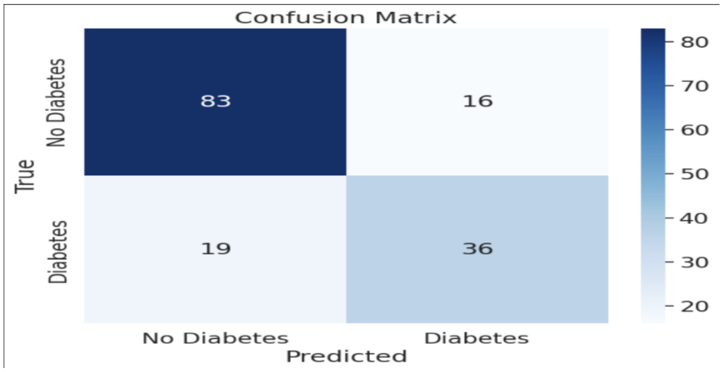
2-Recall:

- Measures the ratio of correct predictions among all actual cases present in the data.
- For class 0: 0.84 (84% of the actual cases were correctly identified).
- For class 1: 0.65 (65% of the actual cases were correctly identified).

3-F1-Score:

- A metric that combines precision and recall, particularly useful when dealing with unbalanced classes.
- For class 0: 0.83 (a good value indicating a balance between precision and recall).
- For class 1: 0.67 (indicates some challenges in retrieving positive cases).

The **Confusion Matrix** is a tool used to evaluate the performance of a machine learning model in classifying data. The matrix illustrates how the model classifies true cases, helping to understand the errors and correct predictions.



-True Negatives (TN): 83 cases were correctly identified as not having diabetes.
-False Positives (FP): 16 cases were incorrectly identified as having diabetes, while they do not have it.
-False Negatives (FN): 19 cases were incorrectly identified as not having diabetes, while they do have it.
-True Positives (TP): 36 cases were correctly identified as having diabetes.

Conclusion:

Based on the accuracy results of the two models, the logistic regression algorithm had a higher accuracy than the Naive Bayes model.
The results from the logistic regression in the confusion matrix showed that the model made fewer incorrect predictions compared to Naive Bayes.
Therefore, the logistic regression algorithm is better in this analysis than Naive Bayes.