Diabetes Prediction Using Naive Bayes

COURSE PRESENTER

(DR. Omaima Fallatah)

Submitted by:

Name	ID
Manar Ali Alsubhi	444003523
Jana Abdulraouf Allihyani	444001382

DEPARTMENT OF (INFORMATION SCIENCE-DATA SCIENCE)

COLLEGE OF COMPUTER AND INFORMATION SYSTEMS

UMM AL-QURA UNIVERSITY

- **Table of content**

.

1. Introduction

Diabetes is a chronic health condition characterized by high blood sugar levels, which can lead to serious health complications if not managed properly. Early diagnosis and prediction are essential for effective intervention and management. This report outlines the use of a Naive Bayes algorithm to predict whether a patient has diabetes based on various health metrics. The dataset includes features such as the number of pregnancies, plasma glucose concentration, blood pressure, skin thickness, insulin levels, body mass index (BMI), diabetes pedigree function, age, and the binary outcome indicating diabetes presence.

- ## Goals of the Study

1. To develop a predictive model using the Naive Bayes algorithm to assess the likelihood of diabetes in patients based on health metrics.
2. To provide insights that can assist healthcare professionals in identifying at-risk individuals for timely intervention.

2.ExploratoryDataAnalysis

Dataset Description

- **Pregnancies:** Number of pregnancies the patient has had.
- **Glucose:** Plasma glucose concentration 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).
- **Diabetes Pedigree Function:** A function that scores the likelihood of diabetes based on family history.
- **Age:** Age of the patient (years).
- **Outcome:** Class variable (0 or 1) indicating the absence or presence of diabetes.

The dataset:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

```
[ ] data.shape
(768, 9)
```

Number of colums:9

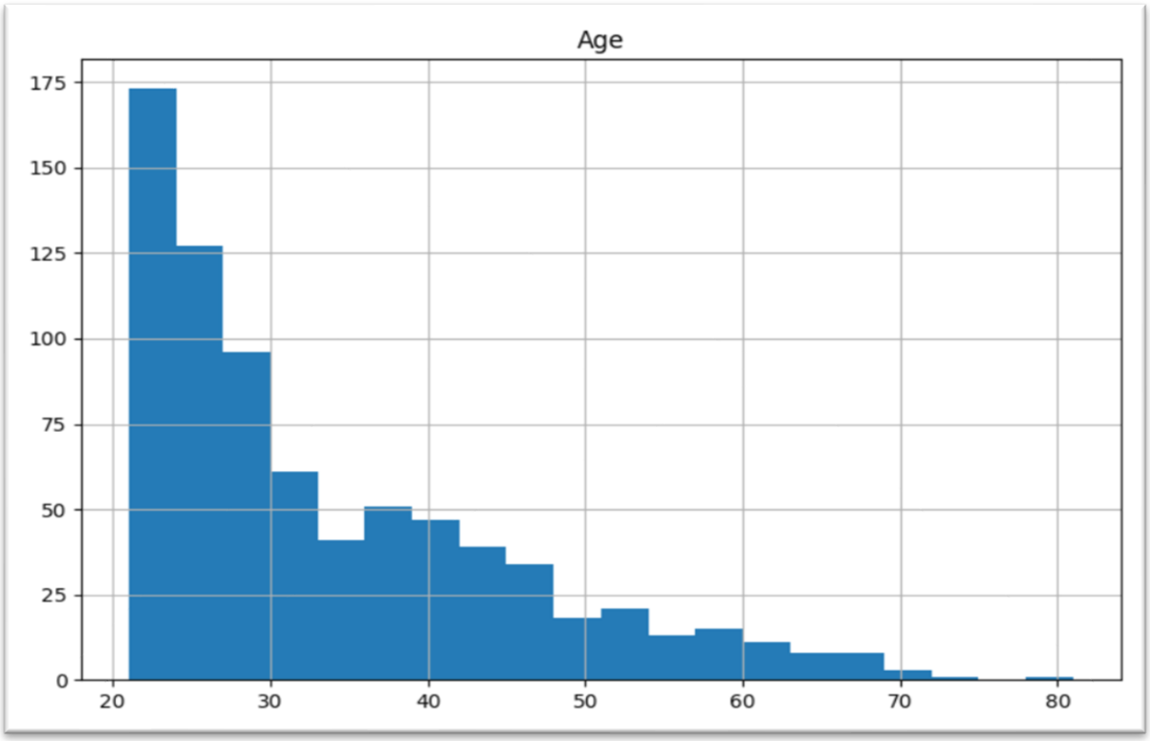
Number of rows:768

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

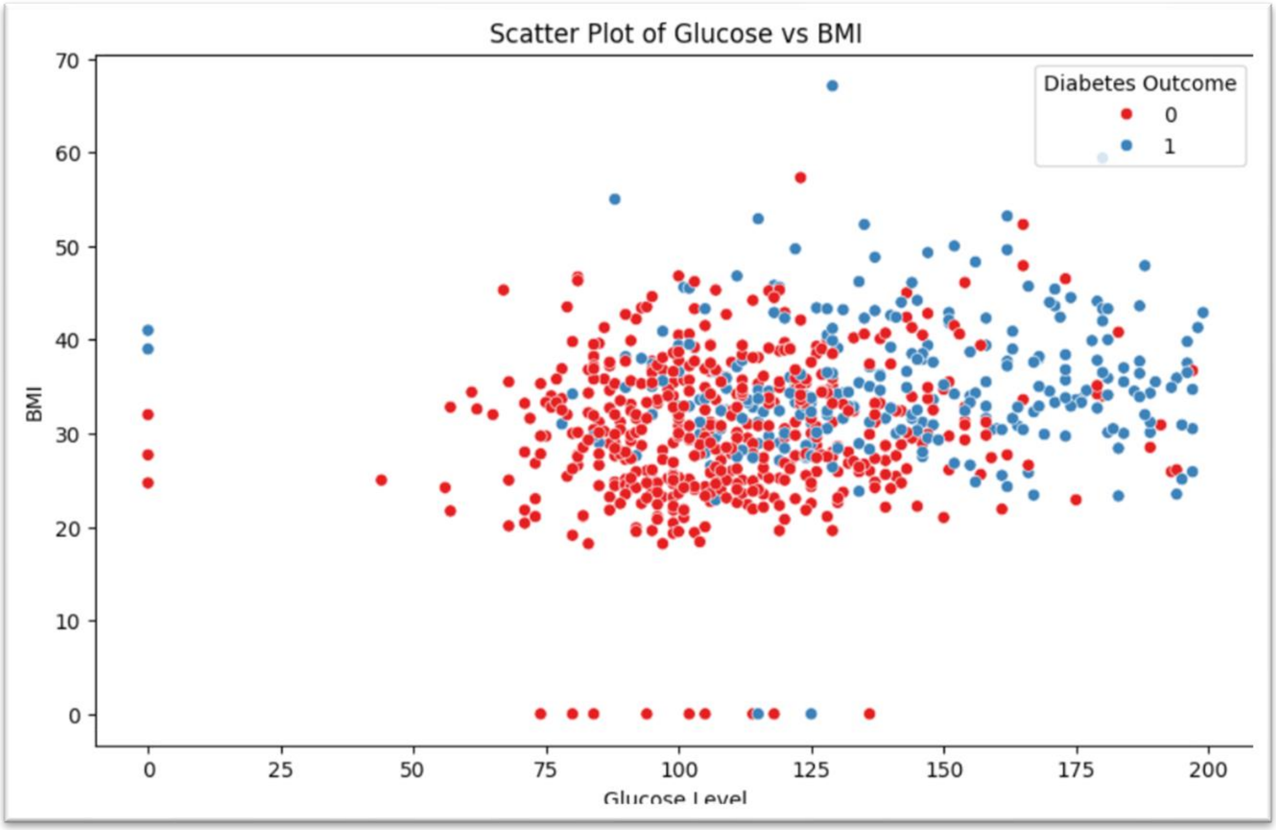
- All columns have 768 non-null entries, indicating there are no missing values in the dataset.
- The dataset contains both integer (int64) and float (float64) data types

	0
Pregnancies	0
Glucose	0
BloodPressure	0
SkinThickness	0
Insulin	0
BMI	0
DiabetesPedigreeFunction	0
Age	0
Outcome	0

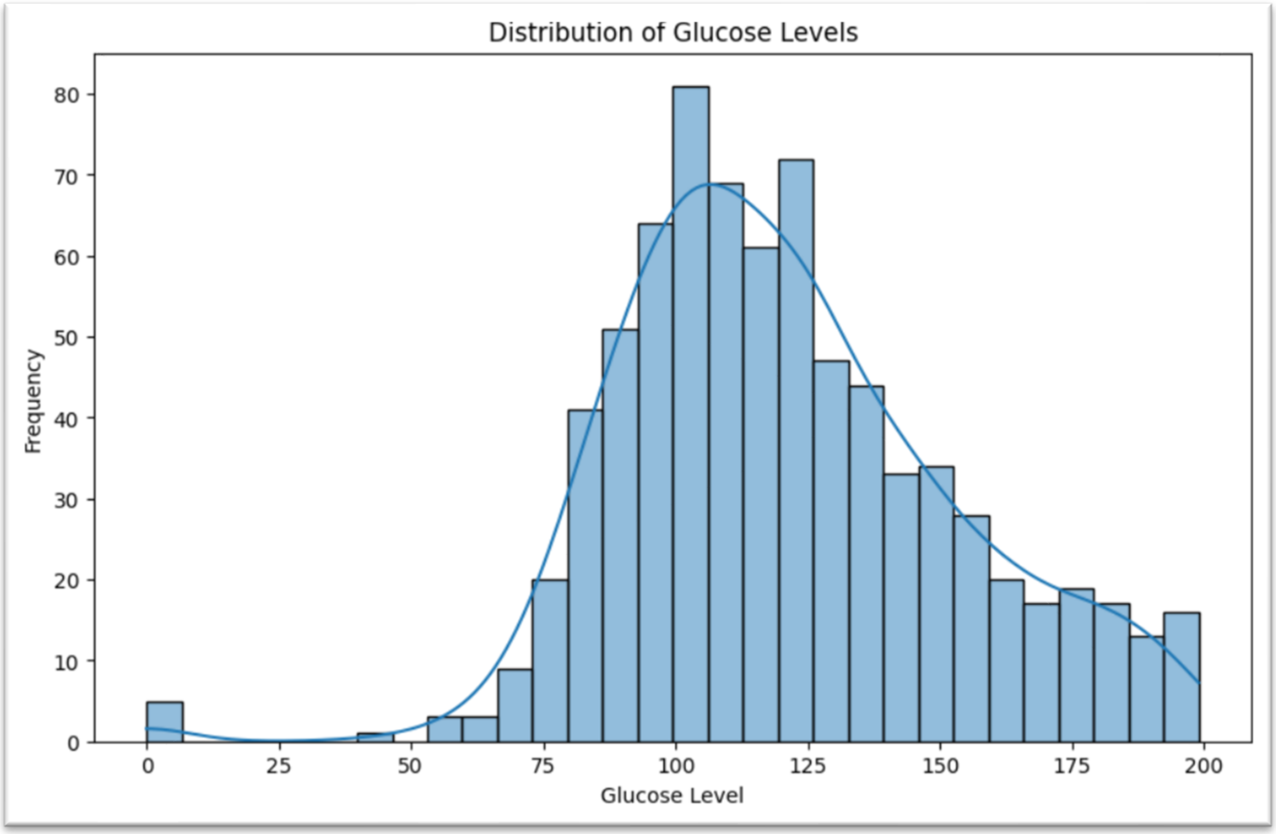
There are no null values



The histogram of Age shows a right-skewed distribution, with most individuals in their 20s and 30s



- The scatter plot of **Glucose** vs. **BMI** displays two groups based on diabetes outcome, with red dots (indicating diabetes presence) generally clustering at higher glucose levels and varying BMI, while blue dots (indicating absence) show a wider spread, suggesting a potential correlation between higher glucose and diabetes status.



- This histogram provides valuable insights into the glucose levels of participants, highlighting the prevalence of elevated levels that may be indicative of diabetes risk.

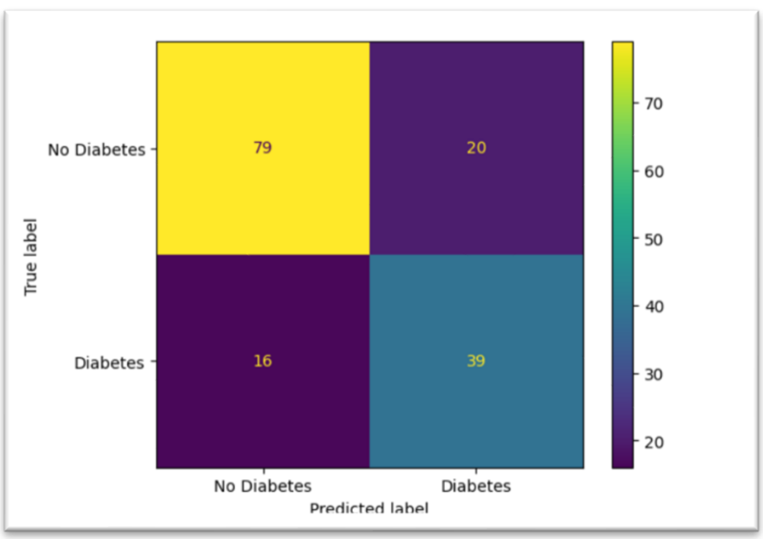
3. Naive Bayes Classifier

Naive Bayes algorithm is used for classification problems, Predicts the probability of an instance belongs to a class with a given set of feature value. It is a probabilistic classifier. It is because it assumes that one feature in the model is independent of existence of another feature.

- Model predict:

The data was split into train and test, X part have all the features that will be used for predicting. (y) will represent the outcome

We used (80%) of the data for training and (20%) for testing




```
accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy: {accuracy:.2f}')
print(classification_report(y_test, y_pred))

# Print confusion matrix
conf_matrix = confusion_matrix(y_test, y_pred)
print('Confusion Matrix:')
print(conf_matrix)

Accuracy: 0.77
              precision    recall  f1-score   support

           0       0.83      0.80      0.81        99
           1       0.66      0.71      0.68        55

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

Confusion Matrix:
[[79 20]
 [16 39]]
```

For this model the accuracy was 0.77 (77%) it means that the model performance is good.

- Naive Bayes Classifier VS Decision tree

```
Naive Bayes Accuracy: 0.7662337662337663

Naive Bayes Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.80      0.81        99
           1       0.66      0.71      0.68        55

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

Decision Tree Accuracy: 0.7467532467532467

Decision Tree Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.76      0.79        99
           1       0.62      0.73      0.67        55

    accuracy                           0.75       154
   macro avg       0.73      0.74      0.73       154
weighted avg       0.76      0.75      0.75       154
```

Naive Bayes	Accuracy: 0.76 (76%)
Decision tree	Accuracy:0.74 (74%)

Overall, it appears that Naive Bayes slightly outperforms Decision Tree in terms of accuracy and overall performance.