

Naive Bayes

COURSE PRESENTER: (Dr. Omaima Fallatah)

|  |  |
| --- | --- |
| Name | Id |
| Yara abodyeh | 444015039 |
| Sashed hasan | 444006091 |

**Introduction**

Diabetes is one of the most common chronic diseases affecting millions of people worldwide. With the increasing prevalence of diabetes, it has become essential to utilize data analysis techniques to understand the factors associated with the disease and develop effective strategies for diagnosis and treatment.

In this report, we will analyze diabetes patient data using the **Gaussian Naive Bayes** model. The project aims to explore the relationship between a set of health features and the outcomes of diabetes, helping to improve predictions and health interventions.

**Objectives**

Data Exploration: To understand the structure and characteristics of the diabetes dataset, including the distribution and relationships among features.

Feature Analysis: To identify key features that significantly affect the likelihood of diabetes based on historical data.

Model Development: To implement the Gaussian Naive Bayes model for predicting diabetes outcomes using the identified features.

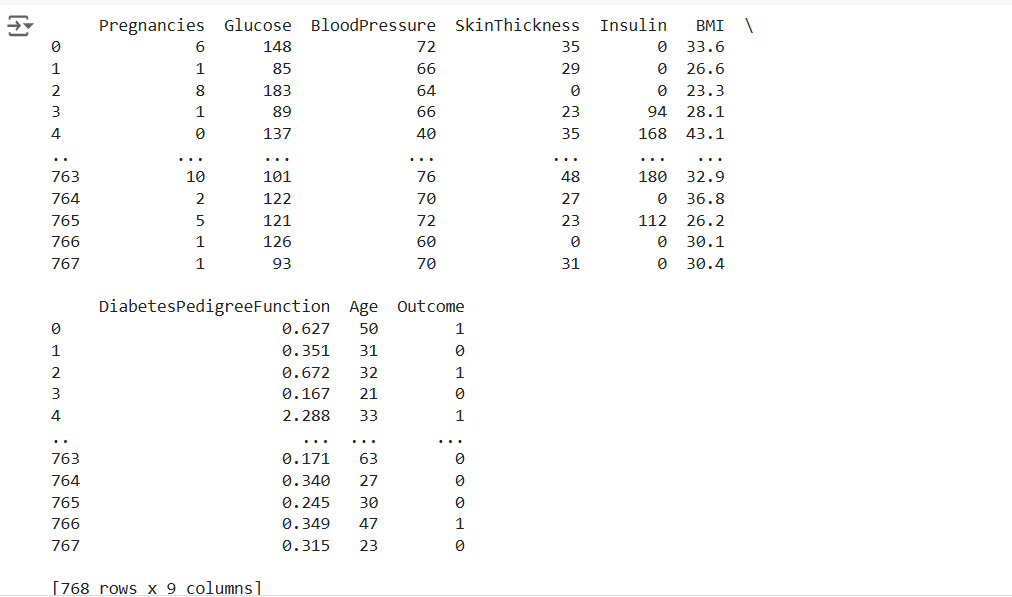
Performance Evaluation: To assess the model's accuracy and effectiveness using metrics such as accuracy score and F1 score.

Insights Generation: To provide insights and recommendations based on the analysis that can aid healthcare professionals in diagnosing and managing diabetes.

**1. Columns in the Diabetes Dataset**

The dataset contains the following columns:

* **Pregnancies**: Number of pregnancies.
* **Glucose**: Blood glucose level.
* **BloodPressure**: Blood pressure.
* **SkinThickness**: Skin thickness (measurement in a specific area).
* **Insulin**: Insulin level in the blood.
* **BMI**: Body Mass Index.
* **DiabetesPedigreeFunction**: Genetic function related to diabetes.
* **Age**: Age of the patient.
* **Outcome**: Result (0 means no diabetes, 1 means diabetes present).



**2. Naive Bayes Model**

**Model Name**:  
**Gaussian Naive Bayes**

A blue and white rectangular object with black text

Description automatically generated

**Why It Was Used**:

* **Simplicity**: The Naive Bayes model is simple and easy to understand, making it a good option for beginners in data analysis.
* **Performance Efficiency**: This model shows good performance in many classification applications, especially when the features of the data are nearly independent.
* **Speed**: The model is fast in training and prediction, making it suitable for large datasets.
* **Handling Imbalanced Data**: It can perform well with imbalanced datasets, making it a suitable choice for analyzing diabetes data.
* **Probability Estimation**: It relies on Bayes' theorem, allowing for the estimation of the probability of belonging to a specific class based on the input features.

**3. Performance Report**

* **Model Accuracy**: 76.62%
* **F1 Score**: 76.45%

A screenshot of a computer

Description automatically generated

**4. Results**

The model was evaluated using a confusion matrix and a classification report.  
The model demonstrated good performance in predicting the presence of diabetes based on accuracy and F1 score.