Naive Bayes classifiers

**Aim**

To develop a classification model to predict diabetes in individuals based on various health metrics using Naive Bayes classifiers, thereby providing insights into the factors influencing diabetes risk and improving early detection.

**Problem Statement**

Diabetes is a growing health concern worldwide, affecting millions of individuals and leading to serious health complications. The Pima Indians Diabetes dataset contains various health-related attributes that can be used to predict whether an individual is likely to have diabetes. The objective is to analyze these attributes and build a predictive model using different types of Naive Bayes classifiers, including Gaussian, Multinomial, and Bernoulli Naive Bayes, to identify the most effective method for predicting diabetes outcomes.

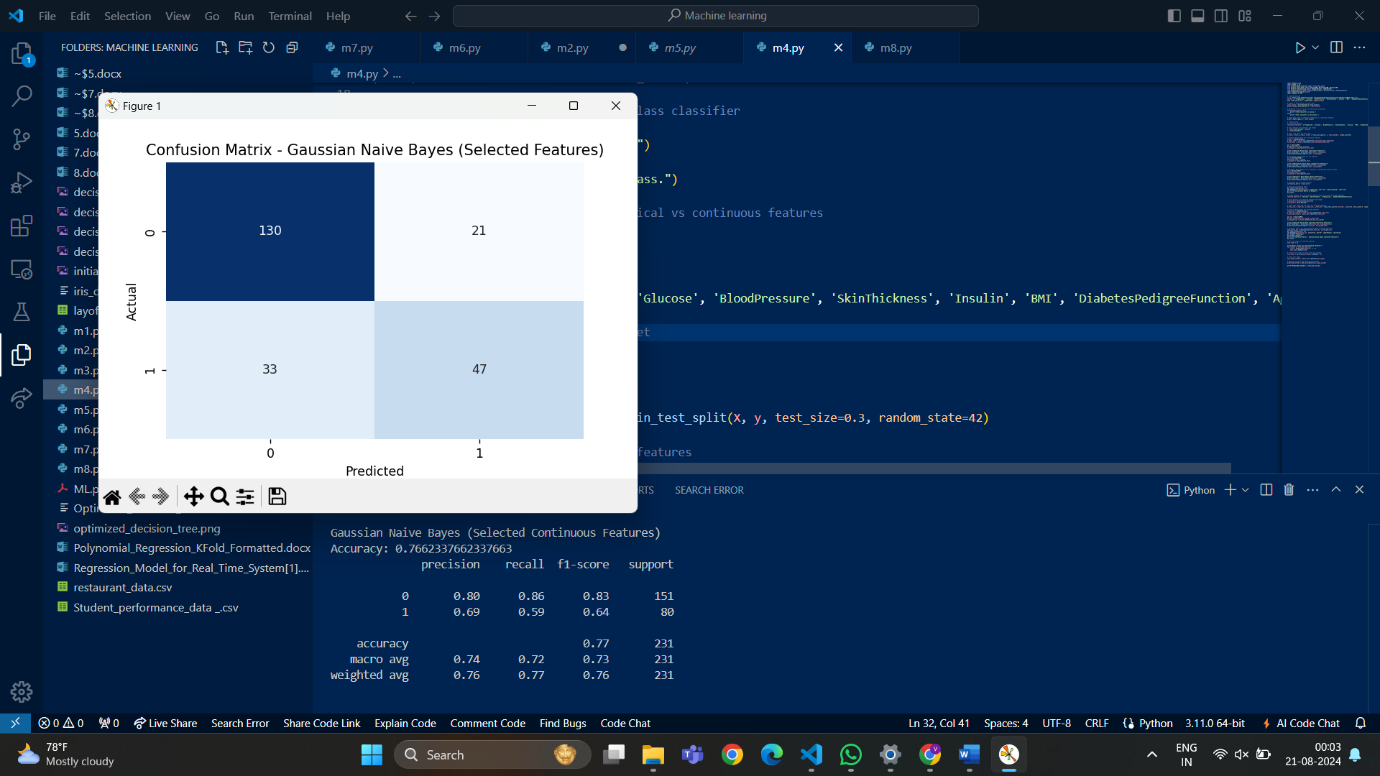
**PROCEDUERE**

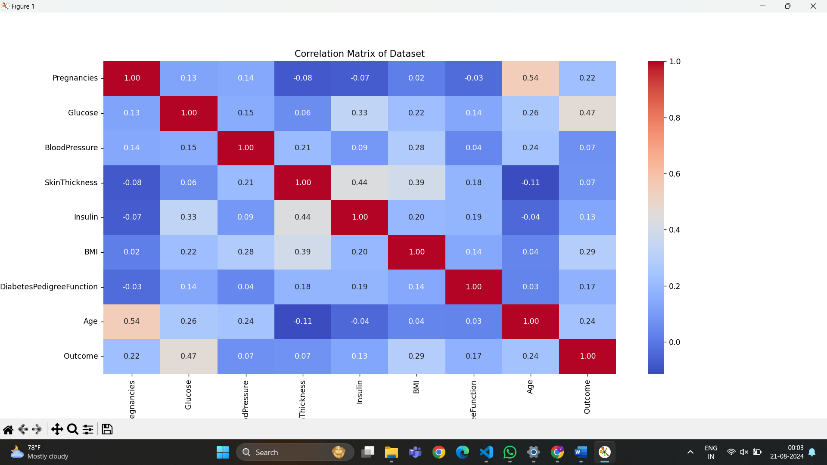
**Data Collection:**

* + The dataset is retrieved from the UCI Machine Learning Repository and contains data on the health metrics of Pima Indian women.

1. **Data Preprocessing:**
   * Load the dataset and assign appropriate column names.
   * Analyze the distribution of the target variable (Outcome) to determine if it is a binary or multiclass problem.
   * Check the data types to identify continuous and categorical features.
   * Handle any missing or invalid values (if present).
2. **Feature Selection:**
   * Identify continuous features that contribute to predicting diabetes.
   * Calculate the correlation matrix to assess relationships between features and the target variable.
   * Select a subset of features with the highest correlation to the outcome for further analysis.
3. **Data Splitting:**
   * Split the dataset into training and testing sets (70% train, 30% test) to evaluate model performance.
4. **Model Development:**
   * **Gaussian Naive Bayes:**
     + Standardize continuous features using StandardScaler.
     + Train a Gaussian Naive Bayes model and evaluate its performance using accuracy, classification report, and confusion matrix.
   * **Multinomial Naive Bayes:**
     + Train a Multinomial Naive Bayes model using all features and evaluate its performance.
   * **Bernoulli Naive Bayes:**
     + Train a Bernoulli Naive Bayes model treating features as binary and evaluate its performance.
5. **Model Evaluation:**
   * Compare the performance of different classifiers based on accuracy, precision, recall, and F1-score.
   * Visualize the confusion matrix for the best-performing model.
6. **User Interaction:**
   * Implement a feature that allows users to input their health metrics and receive a prediction regarding their diabetes status based on the trained model.

**OUTPUT:**

****

****

**Class distribution:**

**Outcome**

**0 500**

**1 268**

**Name: count, dtype: int64**

**The classifier is binary.**

**Data types:**

**Pregnancies int64**

**Glucose int64**

**BloodPressure int64**

**SkinThickness int64**

**Insulin int64**

**BMI float64**

**DiabetesPedigreeFunction float64**

**Age int64**

**Outcome int64**

**dtype: object**

**Gaussian Naive Bayes (Continuous Features)**

**Accuracy: 0.7445887445887446**

**precision recall f1-score support**

**0 0.82 0.79 0.80 151**

**1 0.62 0.66 0.64 80**

**accuracy 0.74 231**

**macro avg 0.72 0.73 0.72 231**

**weighted avg 0.75 0.74 0.75 231**

**Multinomial Naive Bayes (Categorical Features)**

**Accuracy: 0.6190476190476191**

**precision recall f1-score support**

**0 0.72 0.69 0.70 151**

**1 0.45 0.49 0.47 80**

**accuracy 0.62 231**

**macro avg 0.59 0.59 0.59 231**

**weighted avg 0.63 0.62 0.62 231**

**Bernoulli Naive Bayes (Binary Features)**

**Accuracy: 0.6536796536796536**

**precision recall f1-score support**

**0 0.65 1.00 0.79 151**

**1 0.00 0.00 0.00 80**

**accuracy 0.65 231**

**macro avg 0.33 0.50 0.40 231**

**weighted avg 0.43 0.65 0.52 231**

**Gaussian Naive Bayes (Selected Continuous Features)**

**Accuracy: 0.7662337662337663**

**precision recall f1-score support**

**0 0.80 0.86 0.83 151**

**1 0.69 0.59 0.64 80**

**accuracy 0.77 231**

**macro avg 0.74 0.72 0.73 231**

**weighted avg 0.76 0.77 0.76 231**

**Inference**

Based on the performance of the models, the following conclusions can be drawn:

* **Gaussian Naive Bayes (Continuous Features)** achieved an accuracy of approximately 74.46%, indicating a good balance between precision and recall for both classes.
* **Multinomial Naive Bayes** performed less effectively, with an accuracy of around 61.90%, suggesting that it may not be suitable for this dataset with continuous features.
* **Bernoulli Naive Bayes** had an accuracy of approximately 65.37%, but it faced challenges with predicting one class accurately, resulting in a warning about undefined precision.
* When using only **Selected Continuous Features**, the Gaussian Naive Bayes model improved accuracy to about 76.62%, suggesting that these features contribute positively to the prediction of diabetes.