جامعة أم القرى
UMM AL-QURA UNIVERSITY

Collage of computing

Data Analysis 2

COURSE PRESENTER

(DR. Omaima A. Fallatah)

DURATION:

25 Oct 2024

Task 2

“Naive Bayes Classifier”

Student Name	ID
Reuf Bakr Al-sharif	443002428
Jana Hesham Al-Hless	443002347

TABLE OF CONTACT

INTRODUCTION

In the field of medical diagnosis, data analysis plays a critical role in identifying patterns and making accurate predictions about patient conditions. Machine learning techniques, such as classification algorithms, are increasingly used to predict outcomes based on patient data. One such technique is the Naive Bayes classifier, which is particularly effective in medical applications due to its simplicity and ability to handle complex datasets.

The Naive Bayes classifier is a probabilistic model based on Bayes' theorem, used to classify data into categories by calculating the likelihood of a certain outcome given a set of features. In the context of diabetes prediction, this classifier can analyze medical data such as glucose levels, blood pressure, BMI, and other health-related indicators to predict whether a patient is likely to develop diabetes. By analyzing this data, healthcare providers can gain valuable insights into early diagnosis and treatment strategies.

DATASET

The data used in this analysis comes from the

[Naive Bayes Classification Dataset on Kaggle]

(https://www.kaggle.com/uciml/pima-indians-diabetes-database) which provides comprehensive medical records for diabetes prediction analysis.

Overview of the Dataset:

The dataset contains medical data of patients and consists of the following columns:

Pregnancies: Number of times the patient has been pregnant.

Glucose: Plasma glucose concentration (in mg/dL).

BloodPressure: Diastolic blood pressure (mm Hg).

SkinThickness: Triceps skin fold thickness (mm).

Insulin: 2-Hour serum insulin (mu U/ml).

BMI: Body mass index (weight in kg/height in m^2).

DiabetesPedigreeFunction: A function that represents the genetic predisposition to diabetes.

Age: Age of the patient (years).

Outcome: The target variable, indicating whether the patient has diabetes (1) or not (0).

OBJECTIVES

The primary objective of this project is to analyze the Pima Indians Diabetes Database to understand the factors contributing to diabetes prevalence within the population. By utilizing various data analysis and machine learning techniques, we aim to uncover patterns and relationships that can inform better health outcomes and preventive strategies.

In this project, we will explore the dataset, which contains attributes such as glucose levels, blood pressure, BMI, age, and insulin levels, along with the target variable indicating diabetes diagnosis. The insights gained from this analysis can be applied in the following ways:

-Risk Assessment: Identify significant predictors of diabetes to help healthcare providers assess individual risk levels.

-Early Intervention Strategies: Develop targeted interventions for individuals at high risk of developing diabetes, potentially improving prevention efforts.

-Health Awareness Campaigns: Utilize findings to inform community health campaigns aimed at educating the population about diabetes risk factors and prevention methods.

-Data-Driven Healthcare Policies: Support healthcare decision-making by providing insights that can guide policy development and resource allocation in diabetes management.

STEP 1 | DATA CLEANING AND PREPROCESSING

1.1 Handling Missing Values:

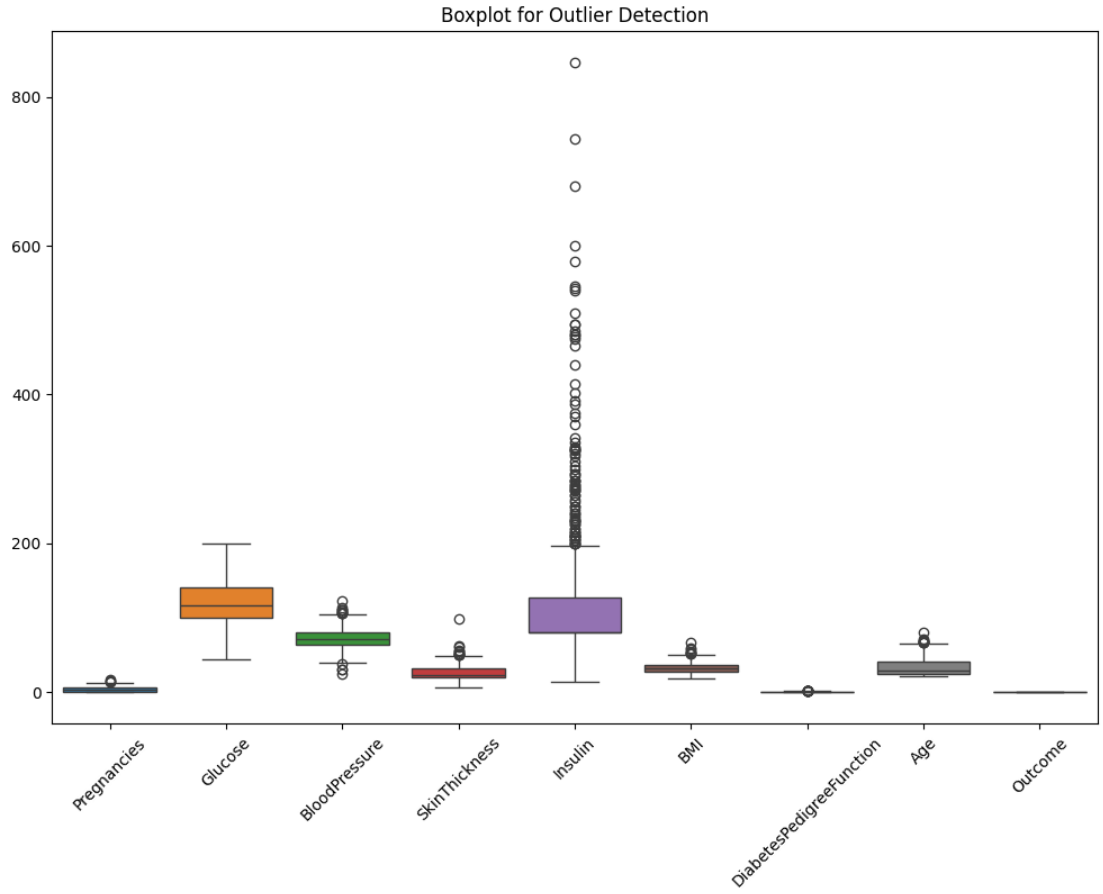
Checked for any missing values across all columns to ensure data completeness and accuracy. Missing values were replaced by the mean of the corresponding columns to maintain the integrity of the dataset and avoid biases in the analysis..

1.2 Zero Value Correction:

Identified columns such as 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', and 'BMI' that contained zero values, which were likely placeholders for missing data. These zero values were replaced by the mean of each respective column to ensure that the data accurately reflects the underlying distribution.

1.3 Scaling:

Standardized the feature values using StandardScaler to normalize the data. This step helps to improve model performance by mitigating the effects of varying feature scales, ensuring that no single feature disproportionately influences the model.

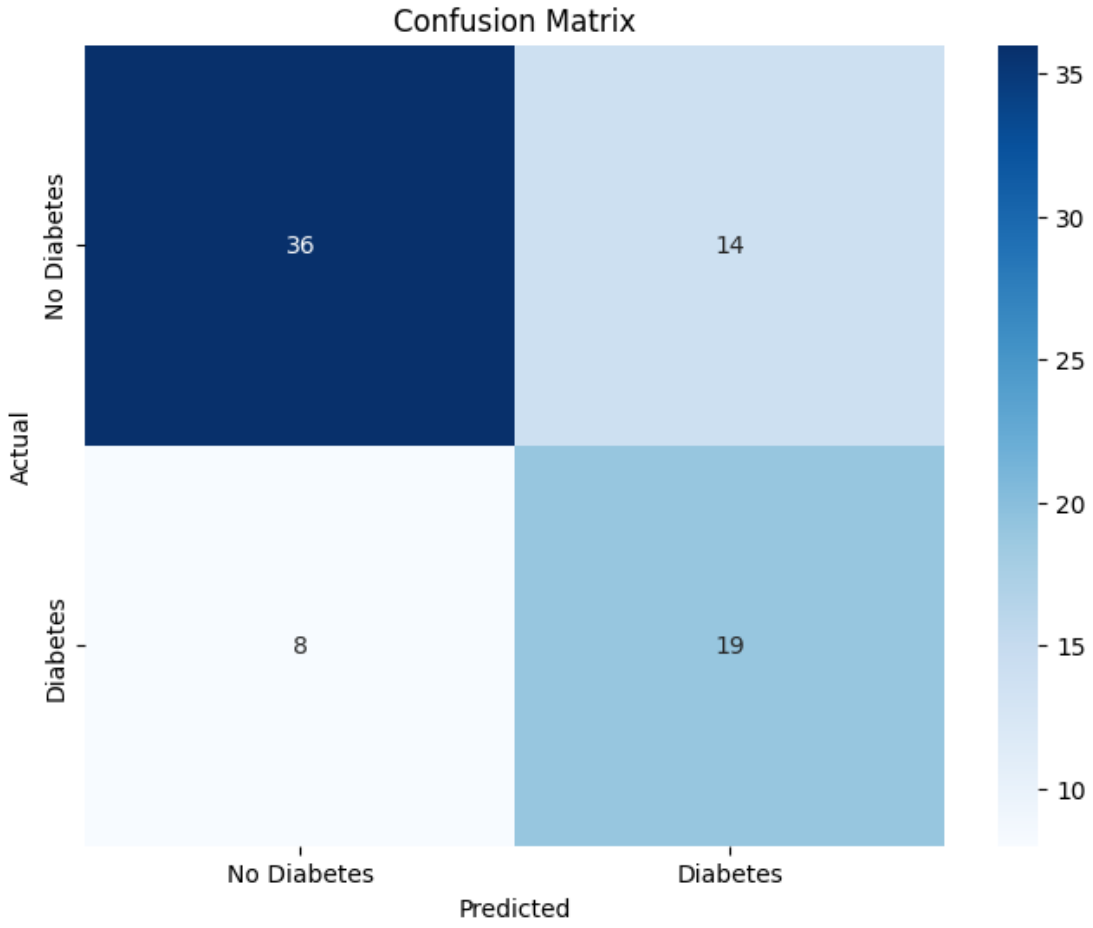


- Description: The boxplot displays the distribution of various features, such as glucose, insulin, and others. It highlights several outliers, particularly in the insulin feature, with values exceeding 600.

- Insights:

1. Outliers can significantly impact model performance and should be addressed through techniques like normalization, transformation, or removal.
2. Identifying and handling outliers is an important data preprocessing step before model training.

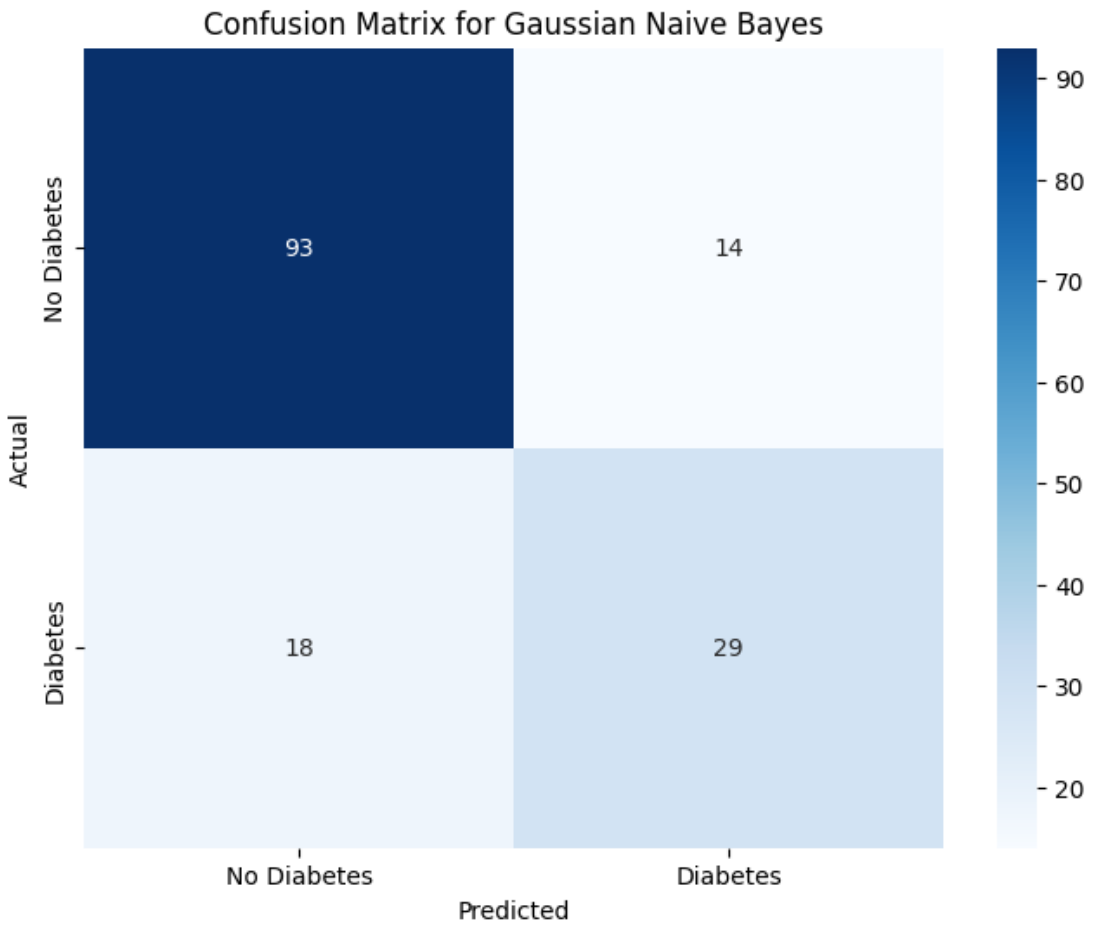
STEP 2 | Predictive Modeling



- Description: This confusion matrix represents the performance of a different classification model. It shows that 36 instances of "No Diabetes" were correctly classified, 14 were misclassified as "Diabetes," 8 instances of "Diabetes" were misclassified as "No Diabetes," and 19 were correctly classified as "Diabetes."

- Insights:

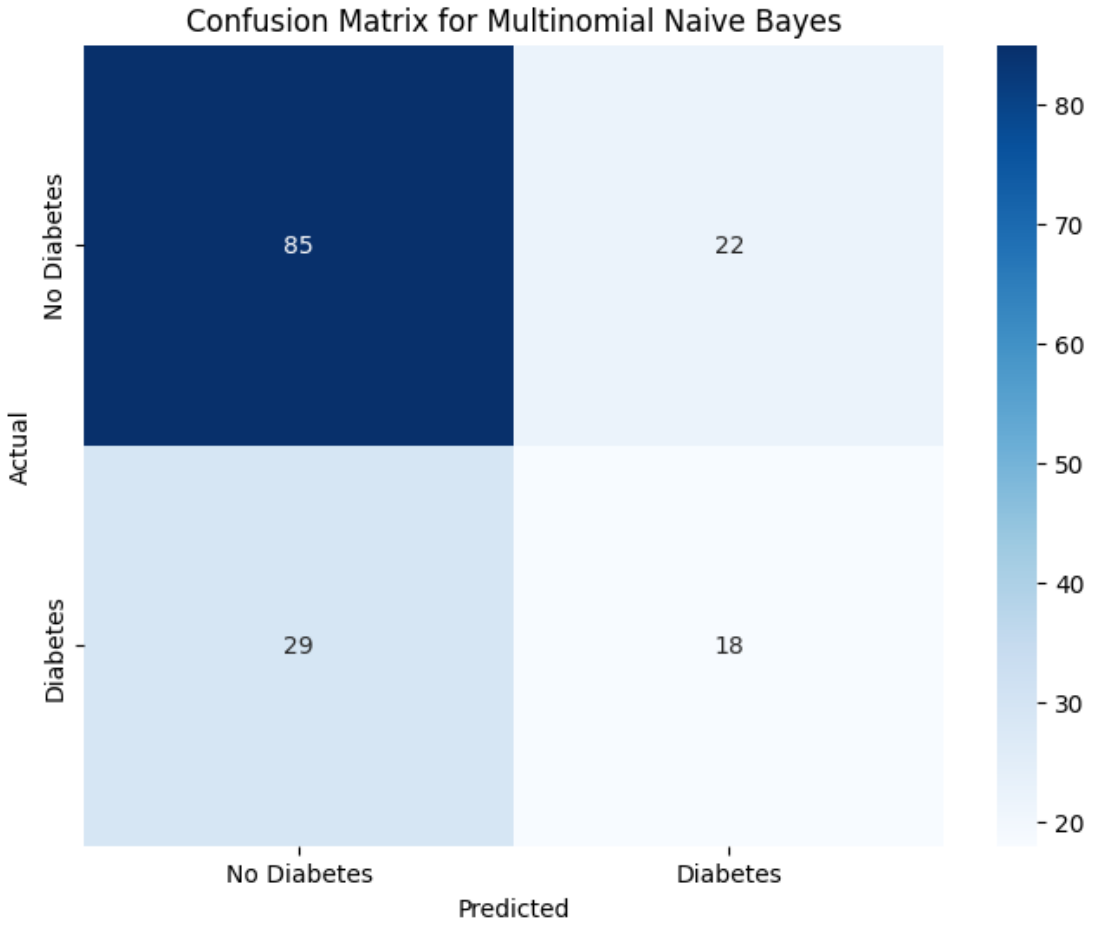
1. Higher accuracy in predicting 'No Diabetes' (81.8%) than 'Diabetes' (57.6%).
2. The model misclassified 14 out of 33 'Diabetes' cases, indicating difficulty in accurate identification.
3. The misclassification rate suggests a need for further optimization through feature engineering, hyperparameter tuning, or alternative algorithms.



-Description: This matrix shows the Gaussian Naive Bayes classifier's performance in predicting diabetes status, with 93 true negatives, 18 false positives, 29 true positives, and 14 false negatives.

-Insights:

1. Higher accuracy for 'No Diabetes' (83.8%) than 'Diabetes.(%67.4) '
2. Misclassifications suggest potential for improvement through feature engineering and tuning.

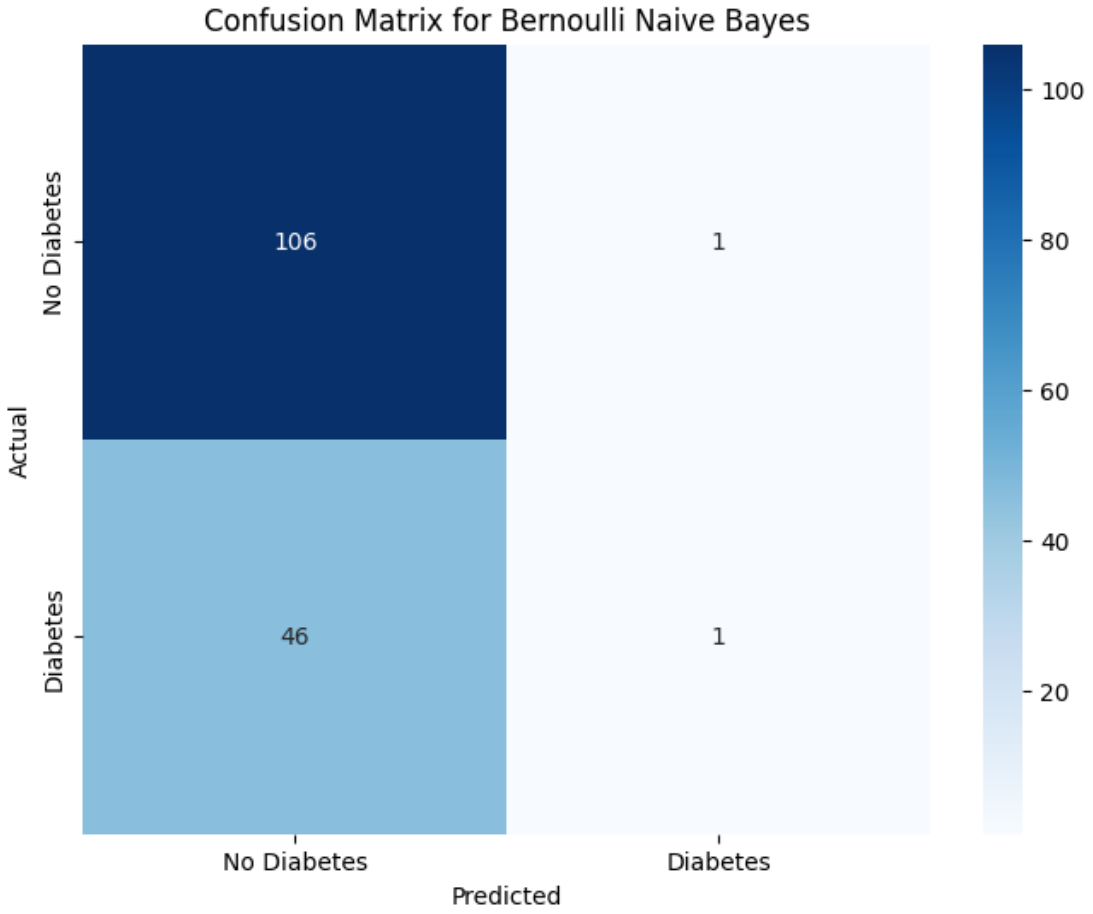


-Description :

This matrix presents the performance of the Multinomial Naive Bayes classifier in predicting diabetes, with 85 true negatives, 29 false positives, 22 true positives, and 18 false negatives.

-Insights:

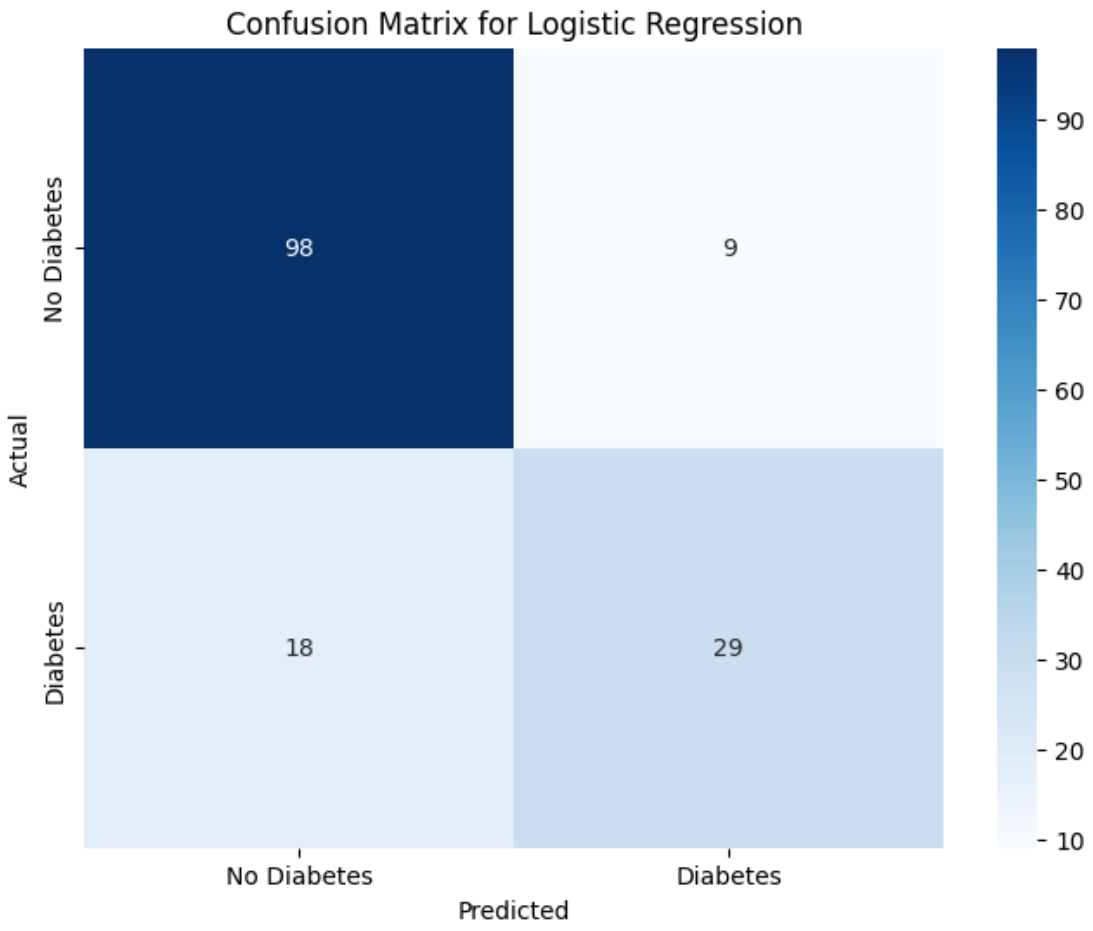
1. Lowest accuracy for 'No Diabetes' (74.6%) and 'Diabetes' (55.0%) among the models.
2. Significant misclassifications suggest that this model may not be well-suited for the data, urging exploration of alternative methods.



-Description: This matrix visualizes the performance of a Bernoulli Naive Bayes classifier. The rows represent the actual class labels, while the columns represent the predicted class labels. In this case, the model has correctly predicted 106 instances of "No Diabetes" and 1 instance of "Diabetes," while misclassifying 46 cases of "Diabetes" as "No Diabetes."

-Insight:

1. The model struggles to predict diabetes accurately.
2. This is indicated by a high number of false negatives (46).
3. Potential improvements could involve enhancing the model.
4. Tuning hyperparameters may also help.



-Description :This matrix illustrates the performance of the Logistic Regression model in predicting diabetes status, featuring 98 true negatives, 18 false positives, 29 true positives, and 9 false negatives.

Insights:

1. Slightly better accuracy for 'No Diabetes' (84.5%) and 'Diabetes' (76.3%) compared to Gaussian Naive Bayes.
2. Misclassifications indicate further optimization opportunities

STEP 3 | Model Evaluation

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

    No Diabetes       0.70      0.79      0.74        90
       Diabetes       0.64      0.53      0.58        64

       accuracy                           0.68       154
      macro avg       0.67      0.66      0.66       154
   weighted avg       0.68      0.68      0.68       154
```

Classification Report:

Accuracy: 68%

The model performs well in predicting "No Diabetes" with 70% precision and 79% recall. However, it struggles with "Diabetes," achieving only 64% precision and 53% recall, missing nearly half of the true diabetes cases. Overall accuracy is 68%, but improvement is needed in detecting diabetes

```
Bernoulli Naive Bayes Classification Report:
                precision    recall  f1-score   support

    No Diabetes       0.58      1.00      0.74        90
       Diabetes       0.00      0.00      0.00        64

       accuracy                           0.58       154
      macro avg       0.29      0.50      0.37       154
   weighted avg       0.34      0.58      0.43       154
```

Bernoulli Naive Bayes:

Accuracy: 58%

This model performed well in classifying "No Diabetes" with a recall of 100%. However, it completely failed to identify any "Diabetes" cases, resulting in low overall accuracy. This indicates that the model heavily relies on predicting the more common class and struggles to distinguish between the two classes.

```
Multinomial Naive Bayes Classification Report:
              precision    recall  f1-score   support

 No Diabetes       0.75      0.84      0.79        50
    Diabetes       0.62      0.48      0.54        27

    accuracy                           0.71        77
   macro avg       0.68      0.66      0.67        77
weighted avg       0.70      0.71      0.70        77
```

Multinomial Naive Bayes:
Accuracy: 71%

The model showed decent performance, with better accuracy in classifying "No Diabetes" than "Diabetes." The recall for "No Diabetes" was 84%, but its performance was weaker in identifying "Diabetes" (recall of 48%). This suggests the model struggles with handling "Diabetes" cases.

```
Gaussian Naive Bayes Model Accuracy: 0.79
              precision    recall  f1-score   support

 No Diabetes       0.84      0.87      0.85       107
    Diabetes       0.67      0.62      0.64        47

    accuracy                           0.79       154
   macro avg       0.76      0.74      0.75       154
weighted avg       0.79      0.79      0.79       154
```

Gaussian Naive Bayes :
Accuracy: 79%

The model performed well overall, with higher accuracy in classifying "No Diabetes" (recall of 87%) but showed some weakness in identifying "Diabetes" (recall of 62%). The overall performance was balanced between the two classes, making this a reasonable model.

```
Logistic Regression Model Accuracy: 0.82
Logistic Regression Classification Report:
              precision    recall  f1-score   support

 No Diabetes       0.84      0.92      0.88       107
    Diabetes       0.76      0.62      0.68        47

    accuracy                           0.82       154
   macro avg       0.80      0.77      0.78       154
weighted avg       0.82      0.82      0.82       154
```

Logistic Regression:
Accuracy: 82%

This model demonstrated the best overall performance, achieving the highest accuracy and recall for both classes. It showed strong capability in distinguishing between "Diabetes" and "No Diabetes," with particularly high performance in classifying "No Diabetes" (recall of 92%). This makes Logistic Regression the most suitable model for this task.

So the ranking of the models from best to worst based on overall accuracy is:

1. Logistic Regression
2. Gaussian Naive Bayes
3. Multinomial Naive Bayes
4. Bernoulli Naive Bayes

CONCLUSION

This analysis explored a diabetes dataset to predict the likelihood of diabetes based on multiple health-related attributes. The data preparation included managing missing values, treating zero values, normalizing features, and scaling data. We used three Naive Bayes classifiers—Gaussian, Multinomial, and Bernoulli—alongside Logistic Regression to identify the best model for prediction accuracy.

Key Findings:

Data Preparation: Addressed missing values, replaced zeros in critical columns, and standardized features, ensuring the dataset was well-prepared for modeling.

Model Performance:

Gaussian Naive Bayes achieved an accuracy of 79%, with a relatively balanced performance in precision and recall across diabetic and non-diabetic categories.

Multinomial Naive Bayes showed lower performance, struggling with features that didn't suit categorical distributions.

Bernoulli Naive Bayes demonstrated reduced effectiveness, particularly on recall for the diabetic class, due to binary feature assumptions.

Logistic Regression outperformed Naive Bayes models with an accuracy of 82%, indicating robust predictive capability with linear separability.

Visualization of Confusion Matrices: Confusion matrices for each model helped to identify areas where each model underperformed, primarily highlighting false negatives in diabetic predictions—a critical insight for health-related models where sensitivity is crucial.

RECOMMENDATIONS

Improvement of Model Sensitivity: Future models should focus on minimizing false negatives to better capture potential diabetes cases. Techniques like hyperparameter tuning, ensemble methods, or advanced algorithms (e.g., SVM or random forests) could be explored.

Further Data Collection: Adding more features, like family history or lifestyle factors, could enhance model performance by providing a broader context for diabetes prediction.

Regular Model Updates: Health datasets evolve over time. Regular model retraining with new data will ensure its relevance and accuracy.

In conclusion, Logistic Regression provided the best balance of accuracy and interpretability, making it a strong candidate for implementation. However, addressing model sensitivity with further refinements could yield an even more reliable predictive tool for identifying diabetes risk.