Naive Bayes Model Comparison Report

Dataset:

Pima Indians Diabetes Dataset was used to compare the performance of two Naive Bayes models: Gaussian Naive Bayes and Complement Naive Bayes.

Data Preparation:

- The dataset was cleaned by removing or replacing unrealistic zero values.
- Outliers were examined using boxplots, and necessary corrections were applied.
- For Gaussian Naive Bayes, StandardScaler was applied to standardize the feature values.
- For Complement Naive Bayes, MinMaxScaler was used to ensure all feature values were positive, as required by the model.

Gaussian Naive Bayes Results:

- Accuracy: 77.7%
- Class 0 (Non-diabetic) Metrics:
    - Precision: 86%
    - Recall: 81%
    - F1-Score: 84%
- Class 1 (Diabetic) Metrics:
    - Precision: 61%
    - Recall: 69%
    - F1-Score: 65%
- Confusion Matrix:
    - True Negatives (Non-diabetic correctly predicted): 88
    - False Positives (Non-diabetic predicted as diabetic): 20
    - False Negatives (Diabetic predicted as non-diabetic): 14
    - True Positives (Diabetic correctly predicted): 31

Complement Naive Bayes Results:

- Accuracy: 67.9%
- Class 0 (Non-diabetic) Metrics:
    - Precision: 79%
    - Recall: 75%
    - F1-Score: 77%
- Class 1 (Diabetic) Metrics:
    - Precision: 46%
    - Recall: 51%
    - F1-Score: 48%
- Confusion Matrix:
    - True Negatives: 81
    - False Positives: 27
    - False Negatives: 22
    - True Positives: 23

Analysis and Conclusion:

- The Gaussian Naive Bayes model demonstrated better overall performance.
- It achieved higher accuracy, F1-Score, and recall, especially in detecting diabetic cases.
- Complement Naive Bayes struggled with the dataset, showing lower accuracy and weaker detection of positive cases.
- This is expected, as Complement Naive Bayes is more suited for text classification and count-based data.
- Based on these results, Gaussian Naive Bayes is the recommended model for predicting diabetes using this dataset.

End of Report.