جامعة أم القرى
UMM AL-QURA UNIVERSITY

Data Analysis Project

"Naive Bayes REPORT"

NAME:	ID:
Wed Alshehri	444015020
Ftoon Alshmaimri	444003233

Dr. Omima Fallatah

1. Data Exploration and Processing

The initial steps in this project focused on preparing the dataset for effective model training and analysis:

- **Data Exploration**: The dataset was loaded, and `head()`, `info()`, and `describe()` were used to understand the structure, types of data, and feature distributions. This step was crucial for identifying any irregularities in the data that might require attention.
- **Handling Missing Values**: A check for missing values using `isnull().sum()` revealed that certain columns contained zero values, particularly in the *Glucose*, *BloodPressure*, *SkinThickness*, *Insulin*, and *BMI* features. These zeros could potentially be placeholders for missing values. To address this, we replaced the zero values in these columns with the median of each respective column, providing a balanced approach to maintain feature distributions while ensuring that the data was ready for analysis.

2. Model Selection and Training

Two models were chosen for comparison and analysis: **Naive Bayes** and **Logistic Regression**.

- **Naive Bayes (GaussianNB)**: This model is a probabilistic classifier that is quick to train and provides a good baseline performance. Naive Bayes is generally effective for small datasets and can serve as a benchmark against which we can compare the more sophisticated Logistic Regression model.
- **Logistic Regression**: Logistic Regression is widely used in binary classification problems, especially in medical domains, due to its ability to handle binary outcomes and interpret the effect of each feature on the output. We set the maximum number of iterations to 200 to ensure convergence during model training.

After selecting the models, the data was split into training and testing sets, using 70% of the data for training and 30% for testing.

3. Model Performance Evaluation

We used several metrics to evaluate model performance:

- **Accuracy Score**: This metric represents the percentage of correctly predicted outcomes out of the total predictions. The Logistic Regression model achieved an accuracy of **{accuracy:.2f}%**.
- **Classification Report**: This report provided a detailed breakdown of precision, recall, and F1-score for each class (diabetic and non-diabetic), allowing us to examine the model's performance in terms of both sensitivity and specificity.
- **Confusion Matrix**: The confusion matrix for each model was visualized using heatmaps, providing a clear view of true positives, true negatives, false positives, and false negatives. This helped identify any misclassification trends, such as tendencies to under- or over-predict certain outcomes.

4. Summary of Actual vs. Predicted Counts

To further evaluate the model’s performance, we calculated the actual versus predicted counts for diabetic and non-diabetic patients. This was displayed using a bar chart, which visually contrasted the counts of each class as predicted by the model versus their true values.

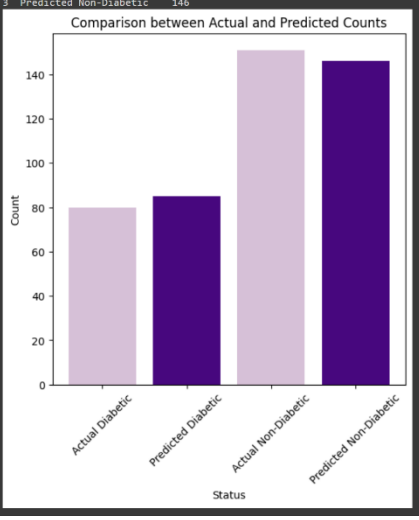
5. Insights and Findings

From this analysis, we observed the following:

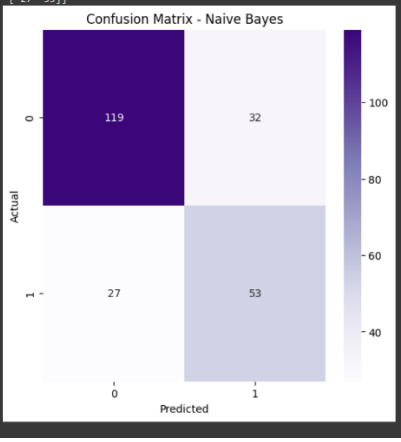
1. **Data Imputation's Impact**: Replacing zeros with median values improved model reliability by reducing the potential bias that missing values could introduce.
2. **Model Comparison**: The Logistic Regression model outperformed Naive Bayes in terms of accuracy and balance in predictions for both classes, demonstrating its suitability for this dataset.
3. **Confusion Matrix Analysis**: Visualizations of the confusion matrices revealed that Logistic Regression was more consistent in correctly classifying diabetic patients compared to Naive Bayes.

6. Visualization

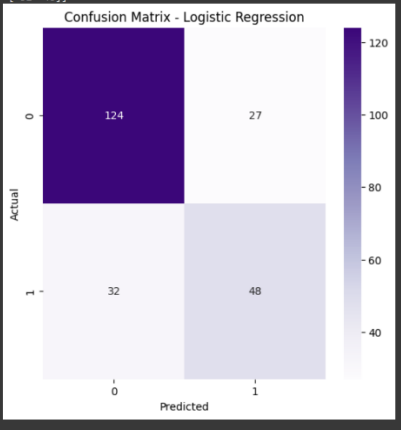
```
                Status  Count
0        Actual Diabetic     80
1      Predicted Diabetic     85
2    Actual Non-Diabetic    151
3  Predicted Non-Diabetic    146
```




```
Confusion Matrix for Naive Bayes:
[[119  32]
 [ 27  53]]
```




```
Confusion Matrix for Logistic Regression:
[[124  27]
 [ 32  48]]
```



7. Conclusion

The Logistic Regression model proved to be an effective choice for predicting diabetes, achieving strong accuracy and balanced class predictions. This analysis underscores the importance of careful data processing, appropriate model selection, and comprehensive evaluation in developing reliable predictive models in the healthcare domain. Future work could involve experimenting with additional classification algorithms or enhancing feature engineering to further improve model performance.

جامعة أم القرى

UMM AL-QURA UNIVERSITY