Mini Project - 2

Lasya Sri Nidumukkala

I. Problem Statement and Description

The heart disease dataset comprises various medical and health-related features aimed at predicting the presence of heart disease in patients. The dataset includes attributes such as age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, rest electrocardiographic (ECG) results, maximum heart rate, exercise-induced angina, and others. These features are crucial in understanding and diagnosing heart disease, with the target variable indicating the presence or absence of the condition.

II. Brief Description of the Three Methods Used

Logistic Regression: A statistical model used to predict a binary outcome based on a set of independent variables. It's particularly useful in medical fields for predicting the probability of a certain condition, such as heart disease, based on various risk factors.

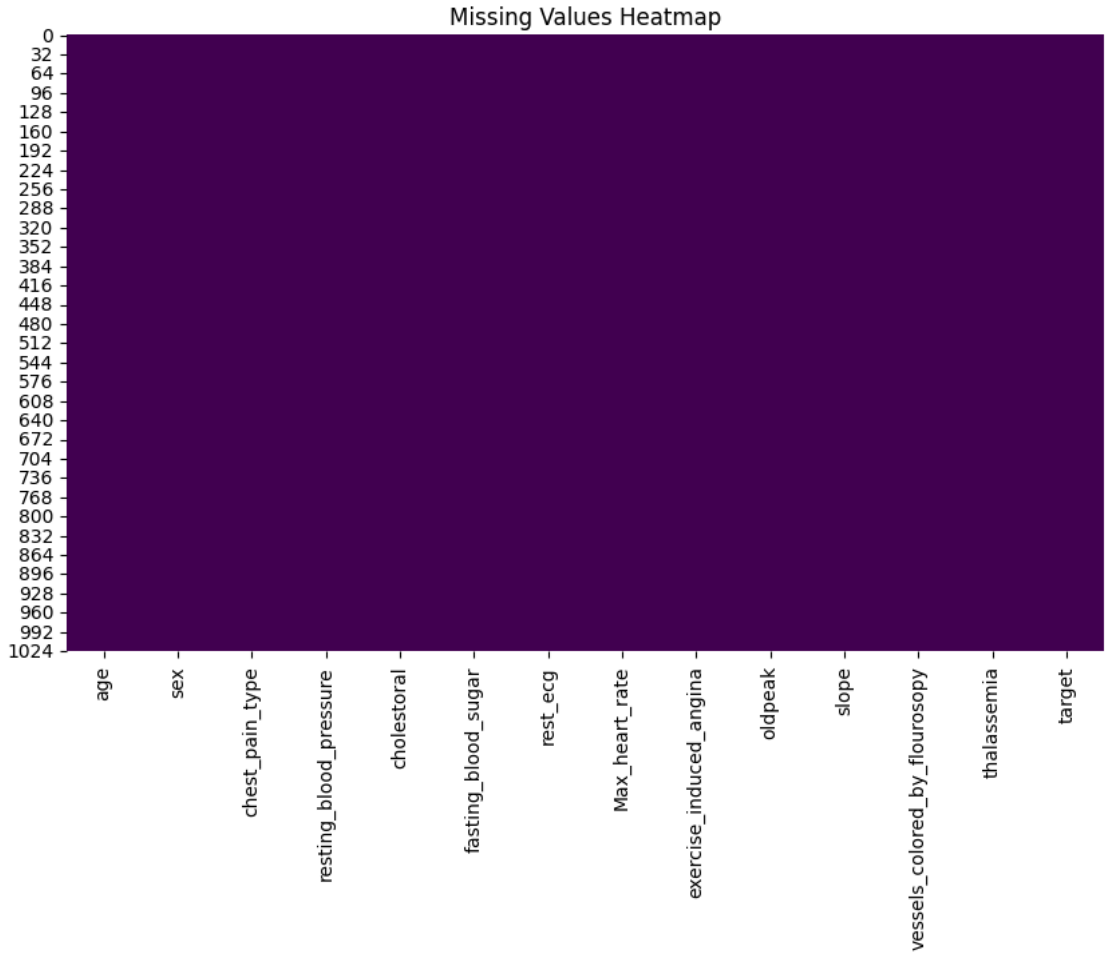
Random Forest Classifier: An ensemble learning method that operates by constructing multiple decision trees during training and outputting the mode of the classes (classification) of the individual trees. Random forests correct for decision trees' habit of overfitting to their training set, making them more reliable for complex datasets like medical records.

Gradient Boosting Classifier: Another ensemble technique that builds one tree at a time, where each new tree helps to correct errors made by previously trained trees. It's known for its effectiveness in handling varied types of data and its ability to improve prediction accuracy.

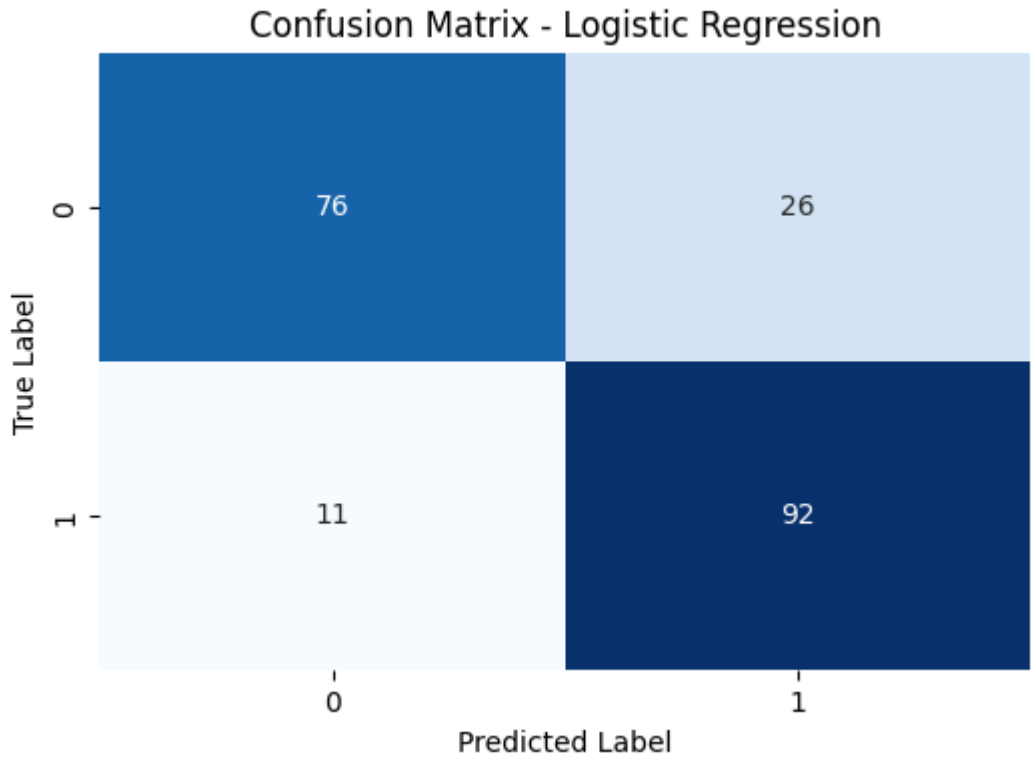
III. Experimental Results

Data Processing :

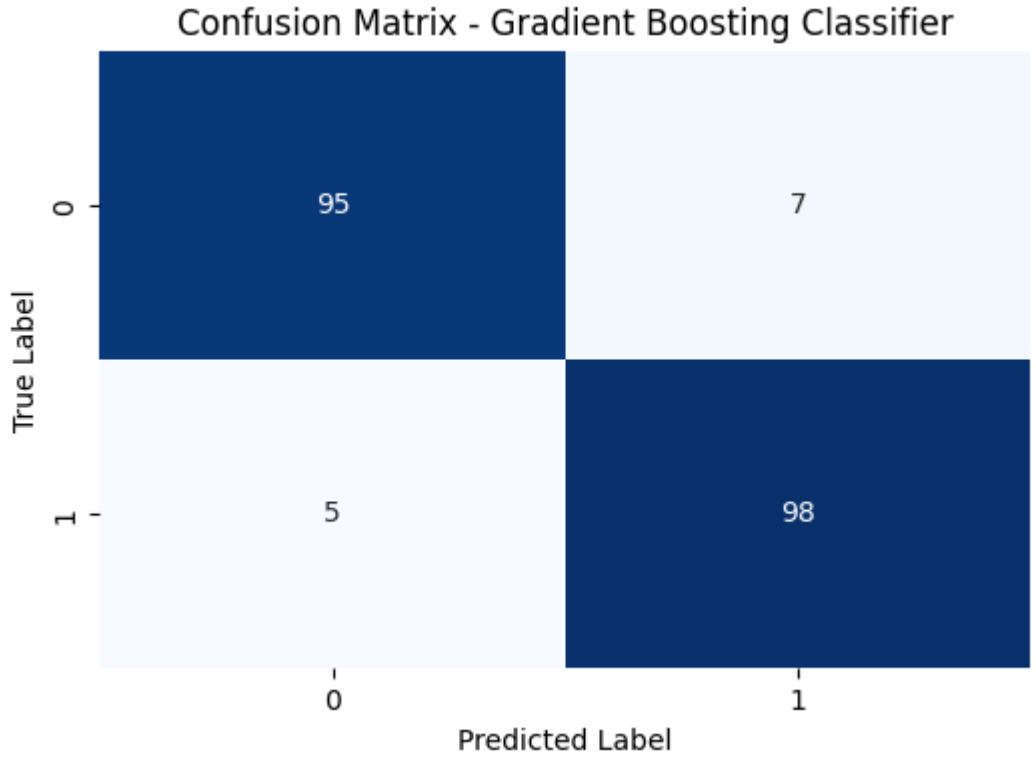
```
Missing Values:
age                                 0
sex                                 0
chest_pain_type                     0
resting_blood_pressure              0
cholestoral                         0
fasting_blood_sugar                 0
rest_ecg                            0
Max_heart_rate                      0
exercise_induced_angina             0
oldpeak                             0
slope                               0
vessels_colored_by_flourosopy       0
thalassemia                         0
target                              0
dtype: int64
```



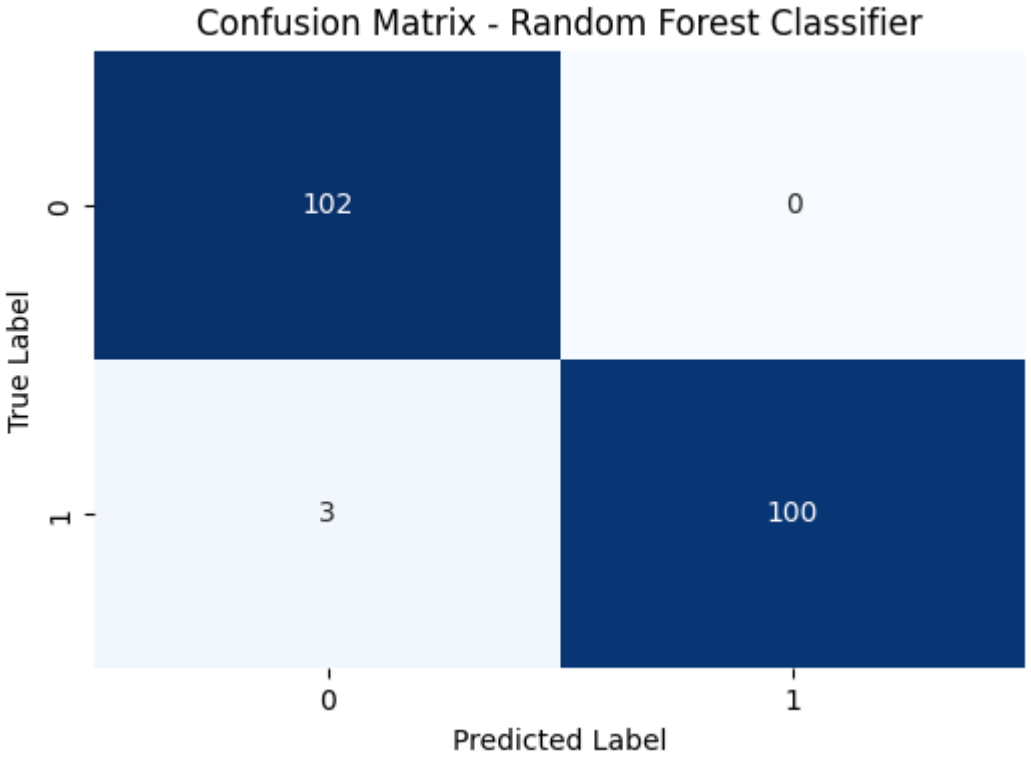
```
Training Logistic Regression...
Logistic Regression Accuracy: 0.8195
Classification Report for Logistic Regression:
              precision    recall  f1-score   support

           0       0.87      0.75      0.80       102
           1       0.78      0.89      0.83       103

    accuracy                           0.82       205
   macro avg       0.83      0.82      0.82       205
weighted avg       0.83      0.82      0.82       205
```



```
Training Gradient Boosting Classifier...
Gradient Boosting Classifier Accuracy: 0.9415
Classification Report for Gradient Boosting Classifier:
              precision    recall  f1-score   support

           0       0.95      0.93      0.94       102
           1       0.93      0.95      0.94       103

    accuracy                           0.94       205
   macro avg       0.94      0.94      0.94       205
weighted avg       0.94      0.94      0.94       205
```



```
Training Random Forest Classifier...
Random Forest Classifier Accuracy: 0.9854
Classification Report for Random Forest Classifier:
              precision    recall  f1-score   support

           0       0.97      1.00      0.99       102
           1       1.00      0.97      0.99       103

    accuracy                           0.99       205
   macro avg       0.99      0.99      0.99       205
weighted avg       0.99      0.99      0.99       205
```



- Logistic Regression showed an accuracy of 81.95%, with precision, recall, and f1-score metrics provided for both classes (0 and 1), indicating how well the model can identify and classify the presence of heart disease.
- Random Forest Classifier achieved a high accuracy of 98.54%, with near-perfect precision and recall, suggesting it was highly effective in classifying the cases correctly.

- Gradient Boosting Classifier reported an accuracy of 94.15%, also demonstrating strong performance across all metrics.

These results were based on a test set of 205 instances, providing a comprehensive view of each model's predictive capabilities.

IV. Discussion of Results

The comparison of the three models reveals significant differences in performance, with the Random Forest Classifier outperforming the others in terms of accuracy. This might be due to its ability to handle the complex interactions and non-linear relationships between the variables in the dataset. The Logistic Regression model, while generally robust and interpretable, showed lower accuracy, possibly due to the linear nature of its decision boundary. The Gradient Boosting Classifier's performance indicates its effectiveness in sequential improvement on the misclassifications of the previous trees.

The choice of model could depend on various factors, including the need for interpretability, computational efficiency, and the specific nuances of the dataset. The results highlight the importance of selecting the right algorithm based on the dataset characteristics and the problem at hand.