PART C: Heart Disease Prediction

1. Introduction

Cardiovascular diseases are a leading cause of death worldwide. Early detection of heart disease is critical for improving the quality of life for patients. This project involves building a machine learning model to predict the likelihood of heart disease based on patient data. The dataset includes features such as age, blood pressure, cholesterol, and lifestyle factors.

2. Data Preprocessing

2.1 Handling Missing Values

Missing values in the dataset were handled by replacing them with the median of each column. Additionally, any infinite values were replaced with NaN and subsequently filled with the median.

2.2 Encoding Categorical Variables

Categorical variables like 'gender' were mapped to numerical values. Specifically, '1' was mapped to '0' (male) and '2' to '1' (female).

2.3 Scaling Features

Numerical features were standardized using the StandardScaler to ensure uniform scaling across all features.

3. Model Development

3.1 Logistic Regression

A logistic regression model was trained as a baseline for comparison. The model achieved a good balance between precision and recall.

3.2 Decision Tree

A decision tree model was trained, but it showed lower accuracy and F1-score compared to other models, indicating potential overfitting.

3.3 Random Forest

A random forest classifier was trained, which achieved the highest F1-score among all models. Feature importance analysis revealed that 'age', 'ap_hi', and 'weight' were the most significant predictors.

3.4 Support Vector Machine (SVM)

An SVM model with a linear kernel was trained. It showed a competitive performance in accuracy and F1-score.

4. Model Evaluation

All models were evaluated using accuracy, precision, recall, and F1-score. The performance metrics are summarized below:

Model Performance Summary:

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.712	0.730	0.672	0.699
Decision Tree	0.630	0.630	0.629	0.629
Random Forest	0.714	0.719	0.702	0.711
SVM	0.719	0.764	0.632	0.692

5. Feature Importance

Feature importance analysis based on the Random Forest model showed that 'age', 'ap_hi', 'weight', 'height', and 'ap_lo' were the most significant predictors of heart disease.

6. Conclusion

Based on the evaluation, the Random Forest model was chosen as the final model due to its superior performance across multiple metrics. The model balances accuracy with interpretability and highlights key predictors of heart disease.