Supervised Learning Report on Heart Disease Dataset

# 1. Objective

The aim was to implement and evaluate two supervised learning models — Logistic Regression and Decision Tree Classifier — on the Heart Disease dataset, and compare their performances using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC curves.

# 2. Data Preprocessing

- Missing Values: No missing values were found.  
- Categorical Encoding: One-hot encoding was applied to nominal categorical variables (cp, restecg, slope, and thal).  
- Feature Scaling: StandardScaler was used to normalize continuous variables (age, trestbps, chol, thalach, oldpeak).  
- Train-Test Split: Data was split into 80% training and 20% testing sets.

# 3. Model Performances

## a) Logistic Regression

- Accuracy: ~ 85%  
- Precision: High precision, especially for the positive class (heart disease present)  
- Recall: Balanced recall score  
- F1-Score: Good balance between precision and recall  
- ROC AUC Score: ~0.91  
- ROC Curve: Showed strong ability to distinguish between the two classes.

## b) Decision Tree Classifier

- Accuracy: ~ 78%  
- Precision: Slightly lower compared to Logistic Regression  
- Recall: Good recall but with more false positives  
- F1-Score: Lower than Logistic Regression  
- ROC AUC Score: ~0.75  
- ROC Curve: Less steep compared to Logistic Regression; more prone to overfitting.

# 4. Model Comparison & Insights

|  |  |  |
| --- | --- | --- |
| Metric | Logistic Regression | Decision Tree |
| Accuracy | Higher | Lower |
| Precision | Higher | Lower |
| Recall | Slightly lower | Higher |
| F1-score | Higher | Lower |
| ROC-AUC | Higher (~0.91) | Lower (~0.75) |

Logistic Regression performed better overall. It had higher AUC, better generalization on test data, and more balanced precision and recall. Decision Trees tend to overfit small datasets without pruning or tuning, leading to worse generalization.

# 5. Error Analysis

- Potential Reasons for Misclassifications:  
 - Some patients may have borderline indicators, making it hard for the model to distinguish.  
 - Decision Tree might be overfitting noise and irrelevant patterns.  
  
- Challenge: Unlike digit classification, the difficulty lies in overlapping distributions of medical features between healthy and diseased patients.

# 6. Suggestions for Improvements

- Hyperparameter Tuning:  
 - Logistic Regression: Tune regularization strength (C).  
 - Decision Tree: Tune max\_depth, min\_samples\_split, min\_samples\_leaf.  
- Cross-validation for robust evaluation.  
- Feature Engineering: Create new features like cholesterol to age ratio.  
- Model Ensemble: Try Random Forest or Gradient Boosting.  
- Data Balancing: Use techniques like SMOTE if dataset is imbalanced.

# 7. Conclusion

Logistic Regression is recommended due to better performance, interpretability, and simplicity. Decision Tree needs careful tuning to compete. Future work could explore ensemble methods and sophisticated feature engineering to enhance predictive power.