Tree-Based Models for Classification on Heart Disease Dataset

Prepared by: Neeraj Gupta

Date: 11 August 2025

Tools: Python, Scikit-learn, Graphviz, Matplotlib

Abstract

This project demonstrates the application of tree-based machine learning models for classifying heart disease presence in patients using the Heart Disease dataset. Models implemented include Decision Tree, Random Forest, and Gradient Boosting. The workflow covers data preprocessing, overfitting analysis, hyperparameter tuning, feature importance interpretation, and model evaluation through cross-validation. Results indicate that ensemble methods outperform single decision trees in terms of accuracy and generalization.

Introduction

Tree-based models are popular supervised learning methods that split the dataset into branches based on feature values, leading to decision outcomes. Decision Trees are highly interpretable but prone to overfitting. Ensemble techniques such as Random Forest and Gradient Boosting combine multiple trees to improve accuracy and reduce variance.

Dataset Overview

The Heart Disease dataset contains patient medical attributes to predict the likelihood of heart disease. Features include age, sex, chest pain type (cp), resting blood pressure (trestbps), cholesterol (chol), maximum heart rate achieved (thalach), exercise-induced angina (exang), and others. The target variable is binary: 0 for no disease, 1 for disease.

Methodology

1. Decision Tree Classifier: Trained and visualized using Graphviz. Analyzed overfitting and controlled tree depth. 2. Random Forest Classifier: Ensemble of trees trained with 100 estimators. Compared accuracy with Decision Tree. 3. Gradient Boosting: Applied

GradientBoostingClassifier for better bias-variance trade-off. 4. Feature Importances: Extracted from Random Forest to identify key predictors. 5. Cross-Validation: Used 5-fold CV to obtain reliable accuracy estimates.

Results

Model	Train Accuracy	Test Accuracy	Mean CV Accuracy
Decision Tree (max_depth=4)	0.85	0.83	0.82
Random Forest (100 trees)	0.91	0.88	0.87
Gradient Boosting	0.90	0.89	0.88

Conclusion

The study demonstrates that ensemble models like Random Forest and Gradient Boosting outperform single Decision Trees in terms of test and cross-validation accuracy. Feature importance analysis revealed chest pain type (cp) and maximum heart rate (thalach) as key predictors. Future improvements could involve trying XGBoost or LightGBM, tuning hyperparameters further, and exploring feature engineering techniques.

References

1. Scikit-learn Documentation: https://scikit-learn.org/stable/ 2. Heart Disease Dataset (UCI Repository) 3. Graphviz Documentation: https://graphviz.org/