Heart Disease Prediction using Decision Trees & Random Forest
Internship Task 5

1. Abstract

This project develops a robust machine learning system for predicting heart disease using Decision Tree and Random Forest classifiers. The objective is to analyze patient health parameters, build interpretable and high-performing models, and evaluate them using advanced metrics and cross-validation techniques.

2. Problem Statement

To build an efficient and reliable predictive model capable of classifying patients based on their risk of heart disease using tree-based machine learning algorithms.

3. Dataset Description

The Heart Disease dataset contains multiple clinical attributes such as age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, ECG results, maximum heart rate, exercise-induced angina, and others. The target variable indicates the presence or absence of heart disease.

4. Methodology

- Data loading and inspection
- Feature-target separation
- Train-test split
- Decision Tree model training and visualization
- Overfitting analysis using depth control
- Random Forest model training
- Model comparison
- Feature importance analysis
- Cross-validation evaluation

5. Model Evaluation

The Decision Tree model provided high interpretability, while the Random Forest classifier achieved superior predictive performance. Overfitting analysis highlighted the importance of controlling tree depth for better generalization. Cross-validation confirmed the robustness of the Random Forest model.

6. Feature Importance Analysis

Feature importance results identified key clinical indicators such as chest pain type, maximum heart rate, ST depression (oldpeak), and number of major vessels as the most influential factors in predicting heart disease.

7. Results and Discussion

The Random Forest classifier significantly outperformed the Decision Tree model in terms of accuracy and generalization. Ensemble learning effectively reduced model variance and improved stability, making it more suitable for real-world healthcare applications.

8. Conclusion

This project demonstrates a complete and professional machine learning workflow for heart disease prediction using tree-based models. The integration of interpretability, overfitting control, ensemble learning, and cross-validation ensures high reliability, making the system suitable for clinical decision support.