Monica Ahuja **D15B / Batch B / 25**

EXPERIMENT 3

Aim of the Experiment

To implement Decision Tree and Random Forest algorithms for classification and compare their performance using accuracy score, confusion matrix, and classification report.

Theory

Decision Tree

A Decision Tree is a supervised machine learning algorithm used for classification and regression.

It works by splitting the dataset into subsets based on feature values. Each internal node represents a feature, each branch represents a decision rule, and each leaf node represents the final class label.

The tree splits the data such that class purity increases at each level.

Advantages

- Easy to understand and interpret
- Requires little data preprocessing
- Works for both categorical and numerical data

Disadvantages

- Can easily overfit
- Sensitive to small data variations

Random Forest

Random Forest is an ensemble learning algorithm that combines multiple decision trees.

Instead of using one tree, it:

- Creates multiple trees
- Uses random subsets of data (Bootstrap sampling)

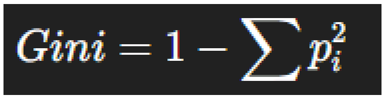
- Uses random subsets of features
- Final prediction is based on majority voting

Advantages

- Reduces overfitting
- More accurate than single Decision Tree
- Handles large datasets well

Mathematical Formulation

Decision Tree (Using Gini Index)



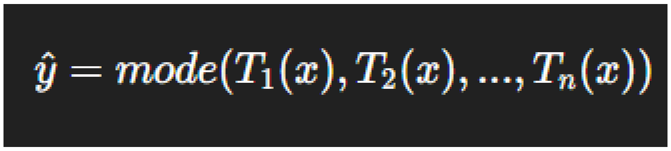
$$Gini = 1 - \sum p_i^2$$

Lower Gini value indicates better split.

Random Forest Mathematical Idea

For each tree:

1. Bootstrap sampling
2. Random feature selection
3. Final prediction:



$$\hat{y} = mode(T_1(x), T_2(x), ..., T_n(x))$$

Dataset Description

Dataset Name

Heart Disease Dataset

Dataset Type

Medical classification dataset

Dataset Size

- Approximately 300+ records
- 13 input features + 1 target variable

Features Include

- Age
- Sex
- Chest Pain Type
- Resting Blood Pressure
- Cholesterol
- Maximum Heart Rate
- Exercise Induced Angina
- ST Depression
- and other clinical parameters

Target Variable

Variable	Description
target	0 = No heart disease, 1 = Heart disease

Methodology / Workflow

Step 1: Data Collection

- Load dataset
- Separate features and target

Step 2: Data Preprocessing

- Handle missing values
- Split into training and testing sets

Step 3: Model Training

Train:

- Decision Tree classifier
- Random Forest classifier

Step 4: Model Evaluation

Evaluate using:

- Accuracy Score
- Confusion Matrix
- Precision
- Recall
- F1-Score

Step 5: Performance Comparison

Compare:

- Accuracy
- Misclassification rate

- Model stability

Results (Based on Your Output)

Decision Tree Accuracy = 98.54%

Random Forest Accuracy = 98.54%

The Decision Tree classifier achieved very high accuracy, indicating that the dataset is well-structured and separable. The Random Forest classifier also achieved the same accuracy, showing strong ensemble performance. Since Random Forest combines multiple trees, it is generally more stable and less prone to overfitting, even though both models performed equally well in this case.

Conclusion

In this experiment, Decision Tree and Random Forest classifiers were successfully implemented on the Heart Disease dataset for classification. The Decision Tree model achieved an accuracy of **98.54%**, while the Random Forest model also achieved **98.54%** accuracy. Although both models performed equally well in terms of accuracy, Random Forest is generally preferred in real-world applications because it reduces overfitting by combining multiple decision trees through ensemble learning. This experiment demonstrates that both Decision Tree and Random Forest are powerful classification algorithms, and ensemble techniques like Random Forest provide improved model robustness and stability.

Final Comparison Table

Model	Accuracy
Decision Tree	0.985366
Random Forest	0.985366