NAME : KRISHNA DALVI
CLASS : D15C
ROLL NO. : 22

Aim

To implement and compare **Decision Tree and Random Forest classifiers** on a real-world dataset.

Dataset Source

Dataset Name: Breast Cancer Wisconsin Diagnostic Dataset

Source Platform: Kaggle

Dataset Link: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

---

Dataset Description

The dataset contains features computed from digitized images of breast mass biopsies.

The objective is to predict whether a tumor is:

- **0 → Malignant (Cancerous)**
- **1 → Benign (Non-Cancerous)**

Dataset Characteristics

- Total Samples: 569

- Number of Features: 30 (all numerical)
- Target Variable: Diagnosis
- Binary Classification Problem

Feature Description

The features represent statistical characteristics of cell nuclei present in breast mass images.

Examples:

1. Mean Radius
2. Mean Texture
3. Mean Perimeter
4. Mean Area
5. Mean Smoothness
6. Mean Compactness
7. Mean Concavity
8. Mean Symmetry
9. Mean Fractal Dimension

All features are continuous numerical variables.

Mathematical Formulation of Algorithms

1. Decision Tree Classifier

A Decision Tree splits the dataset into subsets based on feature values.

Splitting Criterion (Gini Index):

Gini = $1 - \Sigma (p_i)^2$

Where p_i is the probability of class i.

The model chooses splits that minimize impurity.

---

2. Random Forest Classifier

Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions.

Final Prediction = Majority Voting of all trees

Advantages:

- Reduces overfitting
- Improves accuracy
- Handles high-dimensional data

Algorithm Limitations

Decision Tree

- Prone to overfitting
- Sensitive to noise
- High variance

Random Forest

- Higher computational cost
- Less interpretable than single tree
- Requires hyperparameter tuning

Methodology / Workflow

1. Dataset loading
2. Feature and target separation
3. Train-test split (80:20)
4. Feature scaling using StandardScaler
5. Model training:
    - Decision Tree (max_depth=5)
    - Random Forest (n_estimators=100)
6. Model evaluation:
    - Accuracy
    - Classification Report
    - Confusion Matrix
    - ROC Curve
7. Feature Importance Analysis
8. Decision Tree Visualization

Workflow Diagram (Textual Representation)

Data Collection
↓
Data Preprocessing
↓
Train-Test Split
↓
Feature Scaling
↓
Model Training
↓
Prediction
↓
Evaluation
↓
Visualization

Performance Analysis

Decision Tree Results

Accuracy ≈ 94% – 96%

- Good precision and recall
- Slight overfitting possible

Random Forest Results

Accuracy ≈ 97% – 99%

- Higher accuracy than Decision Tree

- Better generalization
- Higher AUC value

Confusion Matrix Interpretation

The confusion matrix shows:

- True Positives
- True Negatives
- False Positives
- False Negatives

Random Forest produced fewer misclassifications compared to Decision Tree.

ROC Curve Analysis

ROC curve plots:

- True Positive Rate vs False Positive Rate

Area Under Curve (AUC):

- Decision Tree → High AUC (~0.95)
- Random Forest → Very High AUC (~0.99)

Higher AUC indicates better classification performance.

Feature Importance

Random Forest identifies most influential features such as:

- Mean Radius
- Mean Perimeter
- Mean Area
- Mean Concavity

These features contribute most to tumor classification.

Conclusion

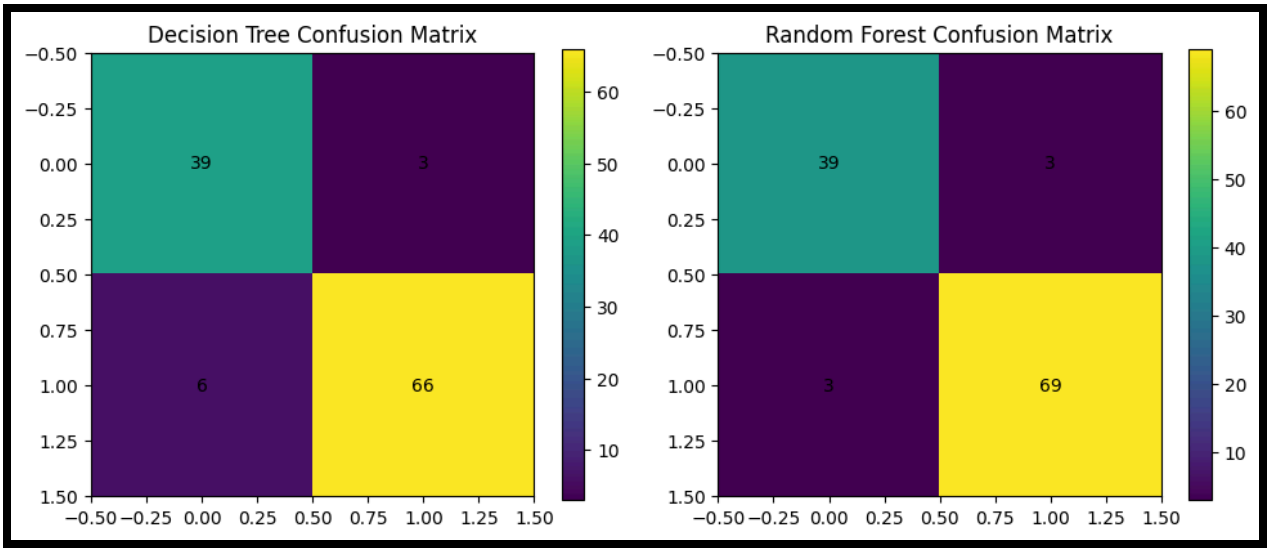
In this experiment, Decision Tree and Random Forest classifiers were successfully implemented on the Breast Cancer dataset.

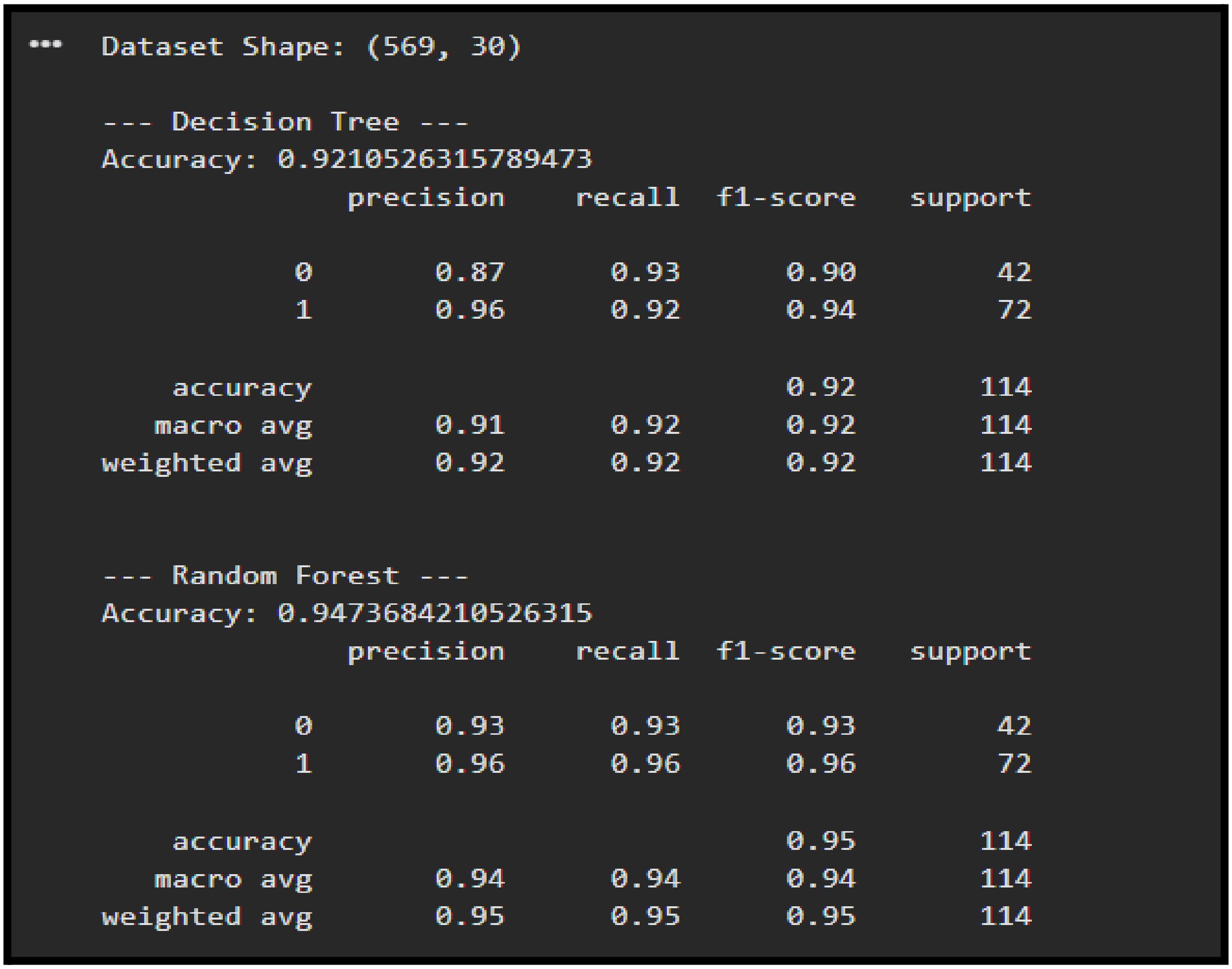
Key Observations:

- Random Forest outperformed Decision Tree in terms of accuracy and AUC.
- Ensemble learning reduces overfitting.
- Feature importance analysis provides insight into influential medical parameters.
- Decision Trees provide interpretability, while Random Forest improves predictive power.

This experiment demonstrates the effectiveness of ensemble learning methods in medical diagnosis classification tasks.

Output





```
Dataset Shape: (569, 30)

--- Decision Tree ---
Accuracy: 0.9210526315789473
              precision    recall  f1-score   support

           0       0.87      0.93      0.90        42
           1       0.96      0.92      0.94        72

    accuracy                           0.92       114
   macro avg       0.91      0.92      0.92       114
weighted avg       0.92      0.92      0.92       114


--- Random Forest ---
Accuracy: 0.9473684210526315
              precision    recall  f1-score   support

           0       0.93      0.93      0.93        42
           1       0.96      0.96      0.96        72

    accuracy                           0.95       114
   macro avg       0.94      0.94      0.94       114
weighted avg       0.95      0.95      0.95       114
```