Comparison of Decision Tree and Random Forest Models for Breast Cancer Classification Report

1. Introduction

The goal of this analysis is to compare the performance of Decision Tree and Random Forest models on the breast cancer dataset. The comparison involves evaluating the models based on accuracy, precision, recall, and F1-score.

2. Data Preprocessing

- **Dataset Overview:** The dataset consists of multiple features describing tumor characteristics. The target variable ('diagnosis') was converted into numeric format: 1 for Malignant (M) and 0 for Benign (B). Unnecessary columns such as 'id' and 'Unnamed: 32' were removed.
- **Missing Data Analysis:** No missing values were found after cleaning the dataset. **- Outlier Detection and Handling:** Outliers were identified using the IQR method and Z-score. Extreme values were clipped or replaced with median values. **- Feature Scaling:** StandardScaler was used to normalize the feature values for better model performance.

3. Exploratory Data Analysis

1.Distribution of Key Features: The histograms (Figure 1) illustrate the distribution of key features like 'radius_mean', 'texture_mean', 'area_mean', 'smoothness_mean', and 'compactness_mean'. Most features exhibit a right-skewed distribution, indicating a need for normalization.

2. Correlation Analysis: The heatmap of the correlation matrix (Figure 2) reveals strong positive correlations among features such as 'radius_mean', 'perimeter_mean', and 'area_mean'. This high correlation suggests possible multicollinearity, which is handled effectively by Random Forest.

Figure 1: Distribution of Key Features

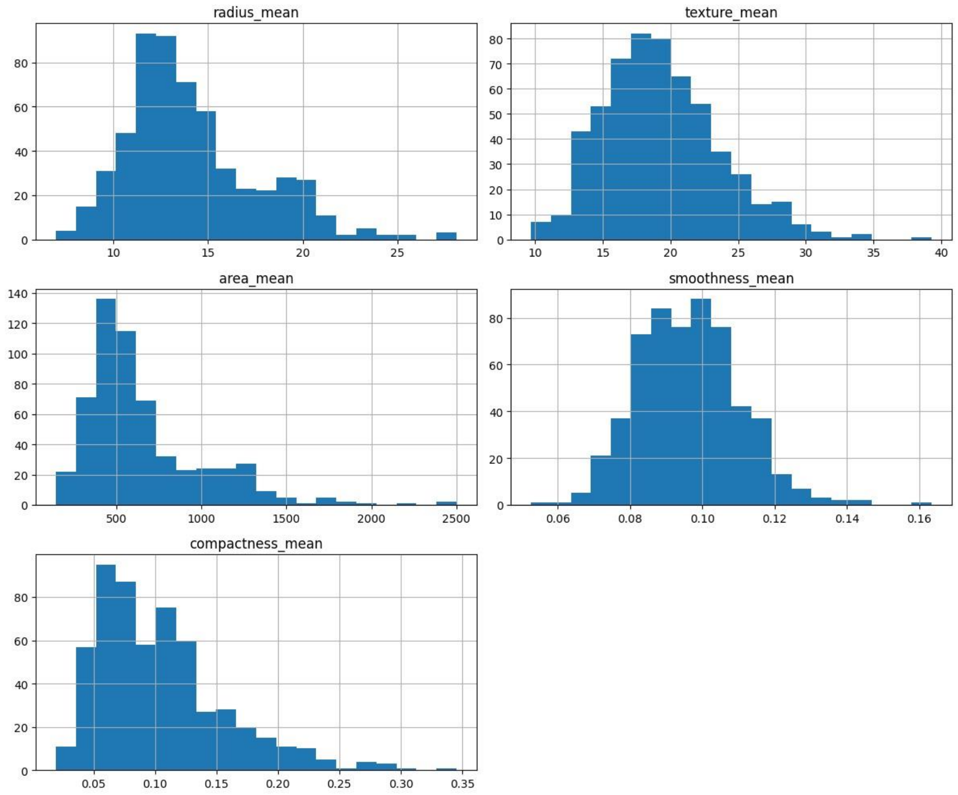
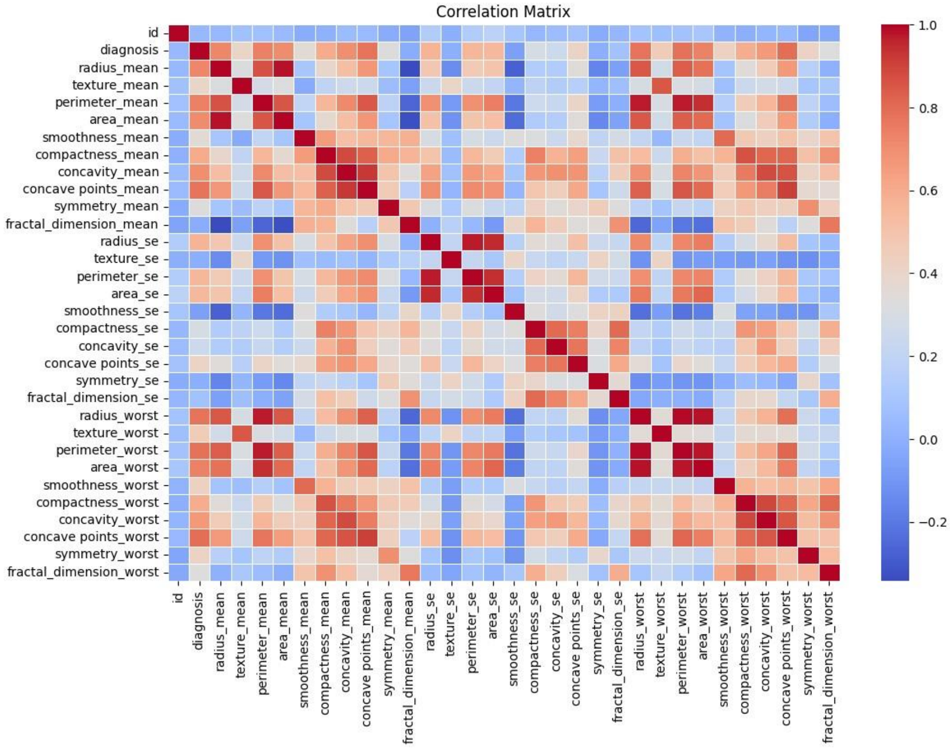


Figure 2: Correlation Matrix Heatmap

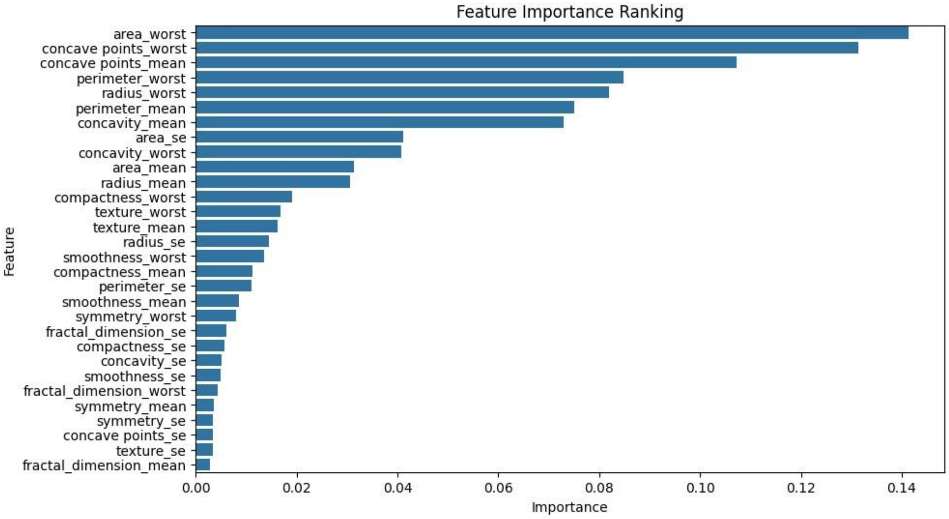


4. Feature Importance

The Random Forest model identified the most important features contributing to the classification task (Figure 3).

Key features:
- 'area_worst', 'concave points_worst', and 'concave points_mean' are the top contributors.

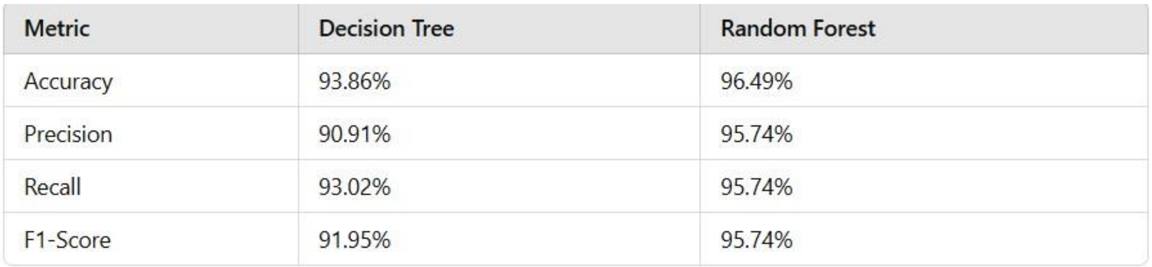
Figure 3: Feature Importance Ranking



5. Model Implementation

1. **Decision Tree:** A single tree was trained on the dataset with default hyperparameters.
2. **Random Forest:** An ensemble of 100 decision trees was trained. This model reduces overfitting and increases generalization.

6. Performance Metrics



Metric	Decision Tree	Random Forest
Accuracy	93.86%	96.49%
Precision	90.91%	95.74%
Recall	93.02%	95.74%
F1-Score	91.95%	95.74%

Observations:
- Random Forest outperforms Decision Tree across all metrics.

- The ensemble method's robustness and ability to handle multicollinearity contribute to its superior performance.

7. Cross-Validation Results

1. **Decision Tree:**
   - Average Accuracy: ~91%
   - Average F1-Score: ~91%
2. **Random Forest:**
   - Average Accuracy: ~96%- Average F1-Score: ~95%

8. Conclusion

Random Forest demonstrated better performance, likely due to its ability to reduce overfitting and consider feature importance effectively. While Decision Tree provides simplicity and interpretability, Random Forest is recommended for tasks requiring higher accuracy and reliability.