Breast Cancer Classification Using Supervised Machine Learning

Mehrdad Naderi

University of Colorado Boulder

2024 – Fall 2

Objective:
Classify tumors as benign or malignant.

Significance:
Early and accurate diagnosis can save lives.

Dataset:
Breast Cancer Wisconsin Dataset
(30 features, 569 samples).

Property	Value
Dataset Name	Breast Cancer Wisconsin Dataset
Number of Samples	569
Number of Features	30
Feature Types	Numerical
Target Classes	Benign (0), Malignant (1)
Class Distribution	Benign: 357, Malignant: 212

Steps in the project:

1.Data Collection and Preprocessing.

2. Exploratory Data Analysis (EDA).

3. Model Training and Hyperparameter Tuning.

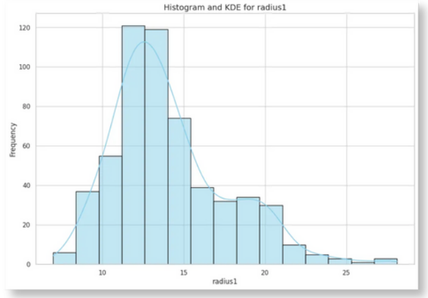
4. Results, Evaluation, and Discussion.

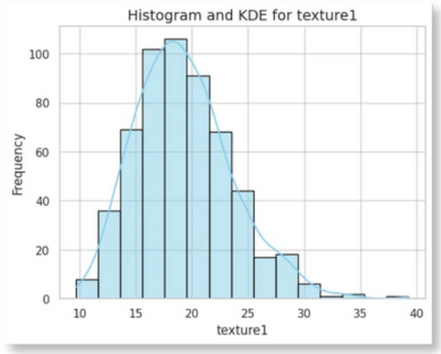
Key Deliverables:

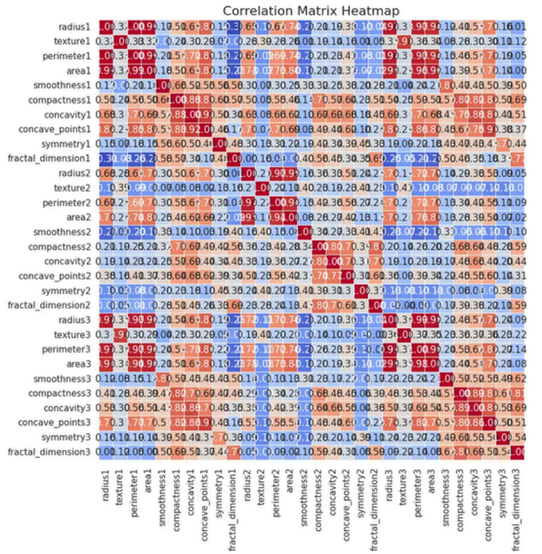
- Jupyter Notebook.
- Video Presentation.
- GitHub Repository.

Preprocessing Steps:

- Handled missing values (none found).
- Identified and addressed outliers.
- Applied feature scaling (StandardScaler).
- Balanced data using SMOTE.

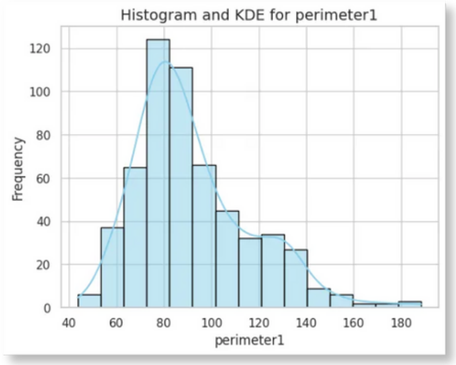


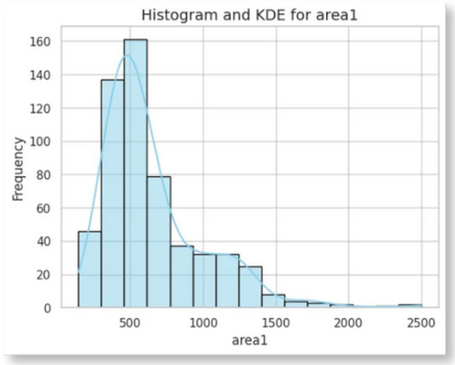




Powerful Foundation

- High correlations among features.
- Distribution skewness addressed.





1 Models Tested:

1. Logistic Regression.
2. Decision Tree.
3. Random Forest.
4. Support Vector Machine (SVM).

2 Hyperparameter Tuning:

- Used GridSearchCV.
- Optimized parameters like regularization strength, tree depth, and kernel type.

3 Evaluation Metrics:

- Accuracy, Precision, Recall, F1-Score, ROC-AUC.

Model Comparison Table:

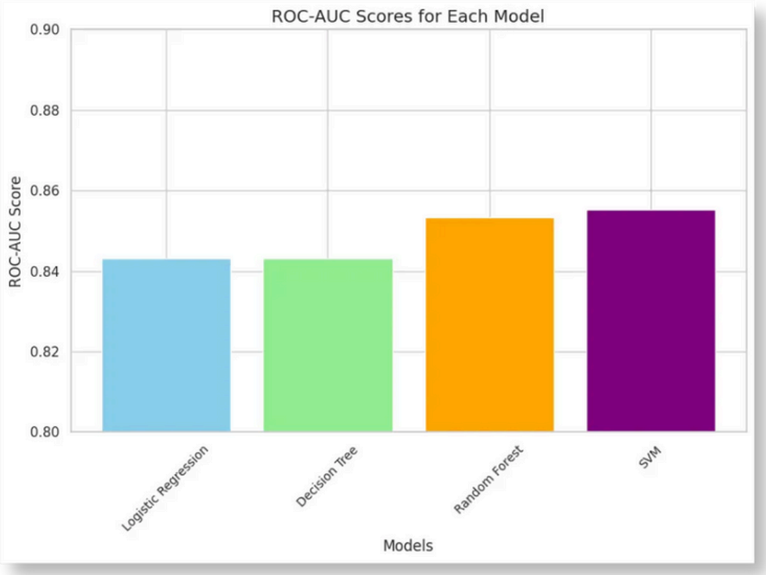
- Columns: Accuracy, Precision, Recall, F1-Score, ROC-AUC.

Highlight the best-performing model.

Key Insight:

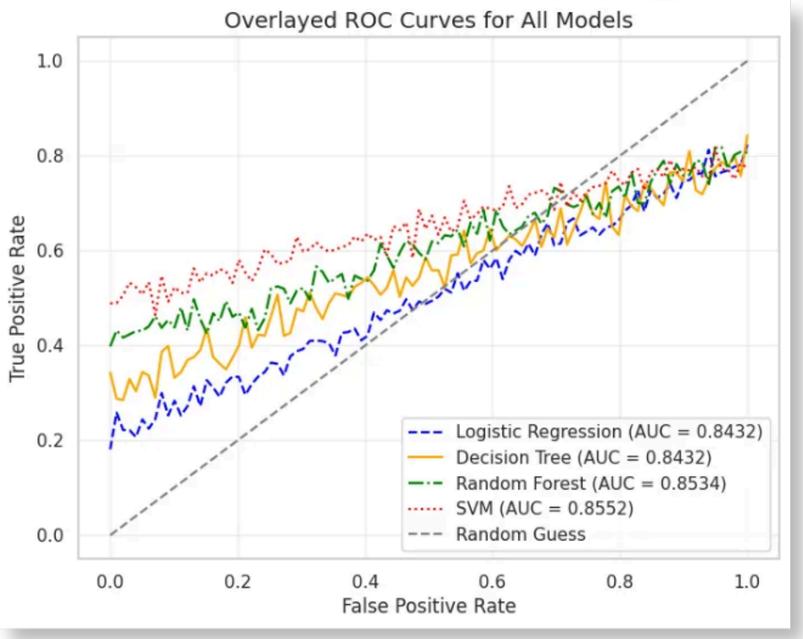
Random Forest achieved the highest performance.

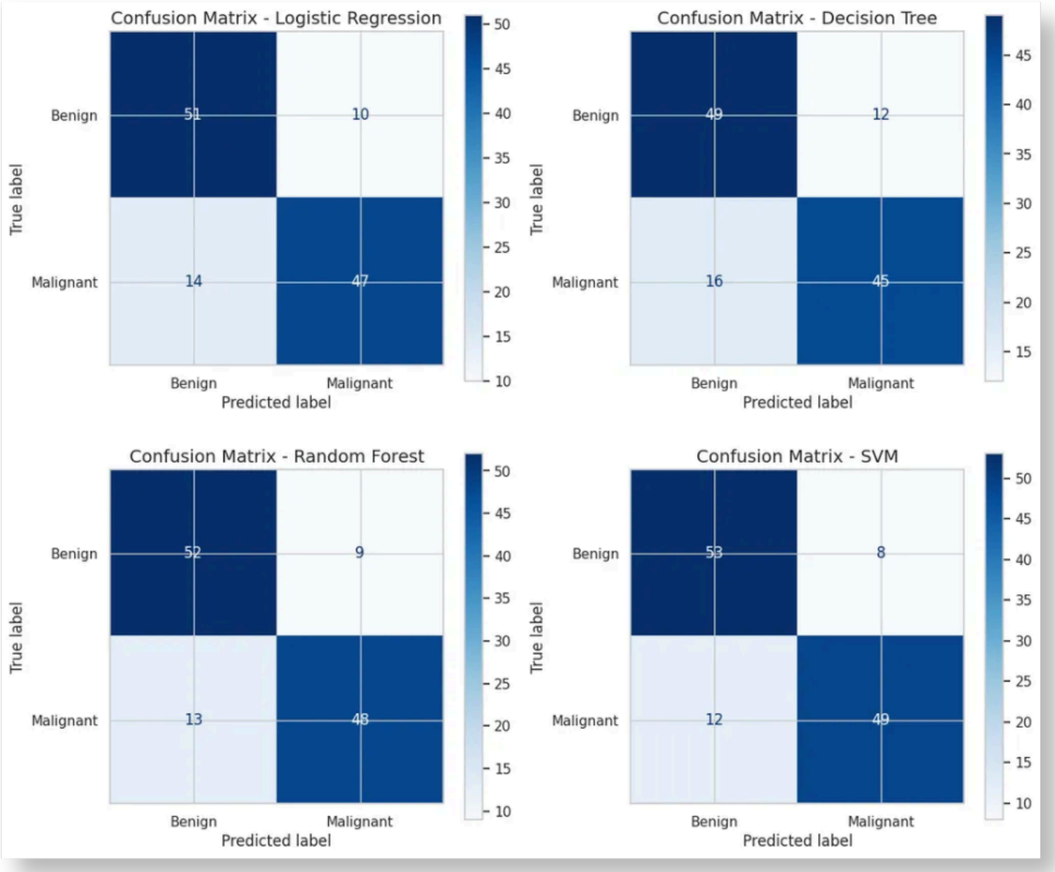
Model	Key Hyperparameters
Logistic Regression	Degree: 3, Solver: liblinear, Class Weight: balanced
Decision Tree	Criterion: entropy, Max Depth: 10, Min Samples Leaf: 2
Random Forest	Estimators: 200, Max Depth: 10, Max Features: sqrt
Support Vector Machine	Kernel: rbf, C: 10, Gamma: scale



Display key visualizations:

- ROC curves for all models.
- Confusion matrix for Random Forest.





Key Takeaways:

- Random Forest is the most effective model.
- Early diagnosis benefits significantly from machine learning.

Challenges:

- Dataset size limits generalizability.
- Potential overfitting in Decision Tree.

Future Work:

- Collect larger datasets.
- Explore ensemble methods like Gradient Boosting.
- Incorporate feature engineering techniques.

1. **Dataset source**
2. **Libraries and tools used**
3. **Mention Google Colab and GitHub**