Breast Cancer Tumor Classification Using Machine Learning

Alice N. Muia

Moringa School

July 23, 2025

Overview

- Objective: Predict whether a breast tumor is benign or malignant using diagnostic data.
- Tools: Python, Scikit-learn, Jupyter Notebook.
- Methods: Logistic Regression, Decision Tree, Random Forest.

Business Understanding

- Early detection of malignant tumors can save lives.
- Hospitals and health organizations need reliable, data-driven tools to assist in diagnosis.
- Machine learning provides efficient, scalable solutions for medical diagnosis.

Data Understanding

- Dataset: Breast Cancer Wisconsin Diagnostic Dataset.
- 569 observations, 30 features + 1 target column (‘diagnosis‘).
- Target: ‘M‘ for malignant, ‘B‘ for benign.

Data Preparation

- Removed ‘id‘ column and handled duplicates.
- Converted target variable to numeric: Malignant = 1, Benign = 0.
- Applied feature scaling using StandardScaler.

Addressing Class Imbalance

- The dataset was slightly imbalanced.
- Used ‘class- weight=’balanced’‘ in Logistic Regression.
- This ensures minority classes are not ignored by the model.

Modeling

- Model 1: Logistic Regression (with class-weight ='balanced').
- Model 2: Decision Tree Classifier.
- Model 3: Random Forest Classifier.
- Each model trained on the same scaled data and evaluated with consistent metrics.

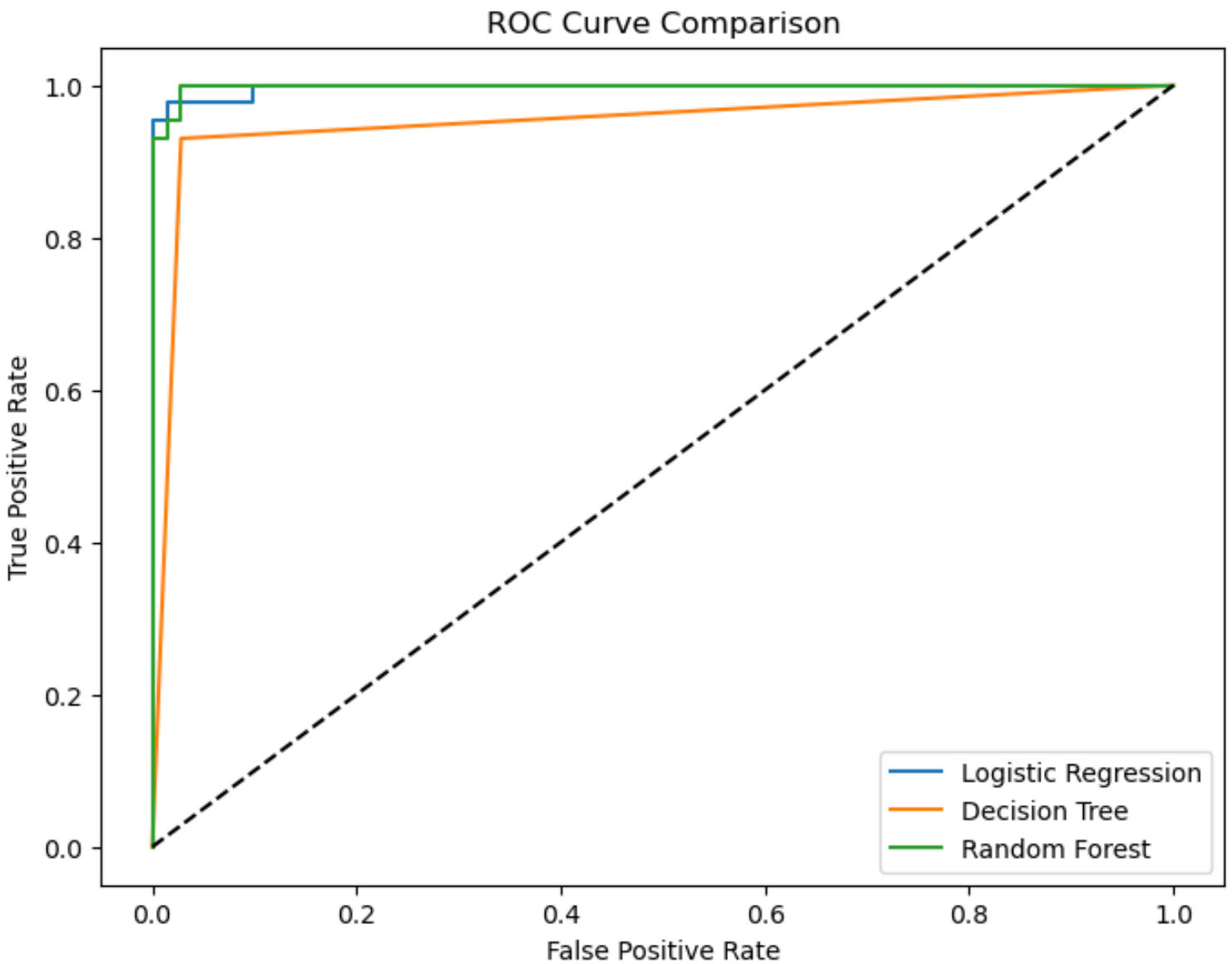
Evaluation Metrics

- Confusion Matrix
- Accuracy, Precision, Recall, F1-score
- ROC Curve and AUC Score

Model Performance Summary

- **Logistic Regression**: Accuracy = 0.98
- **Decision Tree**: Accuracy = 0.96
- **Random Forest**: Accuracy = 0.96
- The Random Forest model had the ROC curve closest to the top-left corner, indicating the best ability to distinguish between the classes.

ROC Curve


ROC Curve Comparison
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression
Decision Tree
Random Forest

Recommendations

- Health facilities should adopt machine learning models like Random Forest for tumor classification.
- Continuous model monitoring is essential to maintain accuracy with new data.
- Consider expanding the dataset and including more clinical features for improved performance.

Conclusion

- Machine learning offers effective tools for medical diagnostics.
- **Random Forest Classifier** provided the best performance in this study.
- With proper data preparation and evaluation, ML can support early and accurate tumor detection.

Thank You!