Breast Cancer Diagnostic Analysis (WDBC)

Medical Data Analysis Project

January 15, 2026

1 Dataset

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset contains 569 samples with 30 numeric features derived from fine needle aspirate images. The target label indicates malignant (M) or benign (B) tumors.

2 Preprocessing

We removed the identifier column, encoded diagnosis as binary (M=1, B=0), and used stratified train/test splits. Summary statistics are shown below.

2.1 Data Overview

- Rows: 569
- Columns: 32
- Missing values: 0
- Duplicate rows: 0

3 Modeling and Evaluation

We trained a logistic regression model with standardization and a random forest classifier. Performance metrics are reported in Table 1.

model	accuracy	precision	recall	f1	roc_auc
LogisticRegression	0.974	0.976	0.952	0.964	0.995
RandomForest	0.974	1.000	0.929	0.963	0.997

Table 1: Model evaluation metrics on the test set.

4 Visual Exploration

Figures 1–6 summarize key findings.

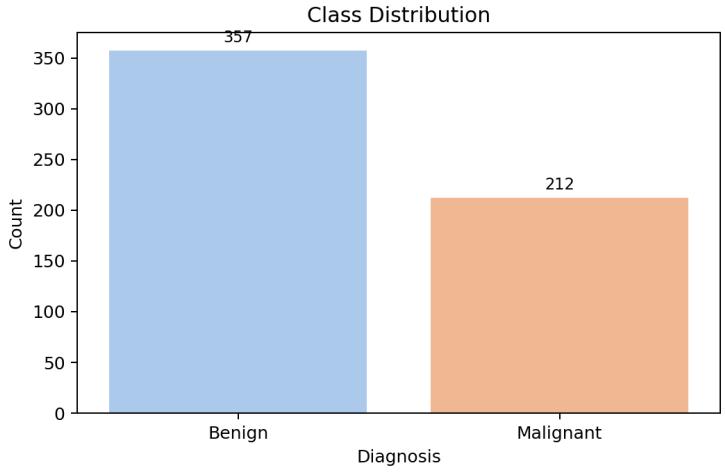

Figure 1: Class distribution.

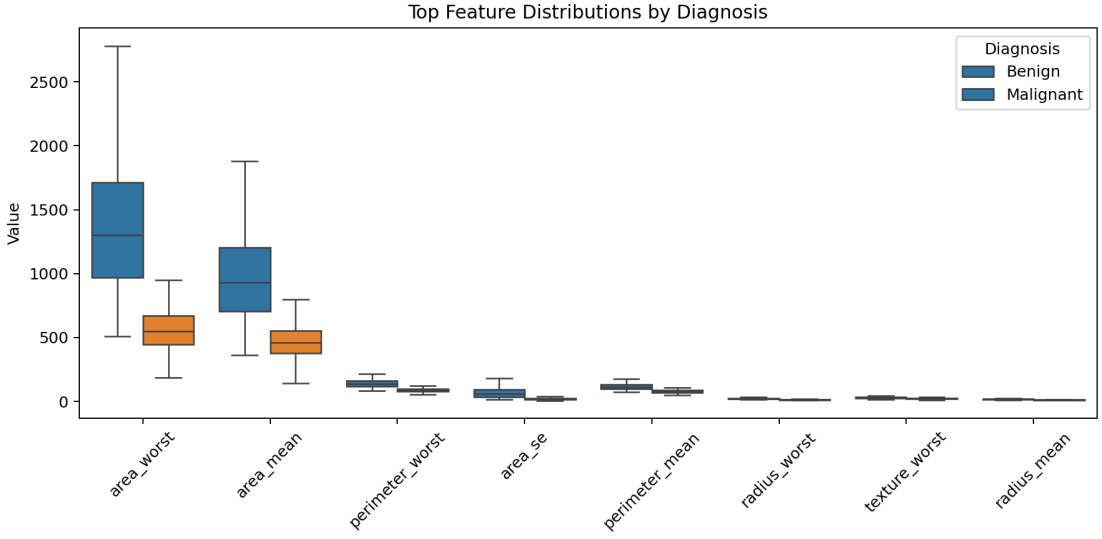

Figure 2: Top feature distributions.

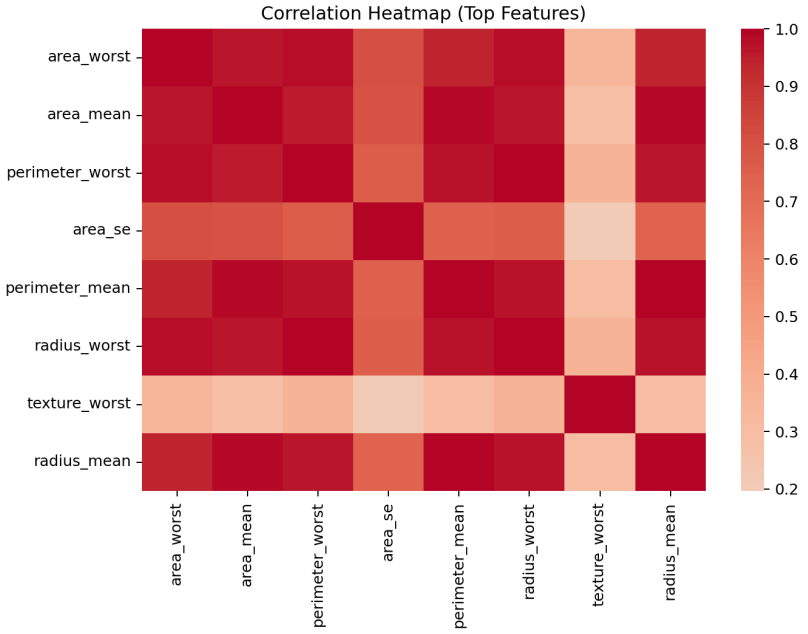

Figure 3: Correlation heatmap of top features.

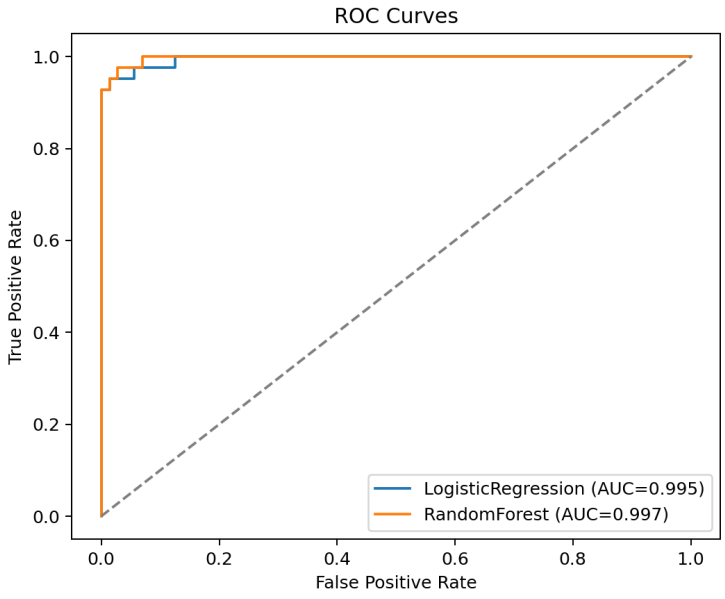

Figure 4: ROC curves.

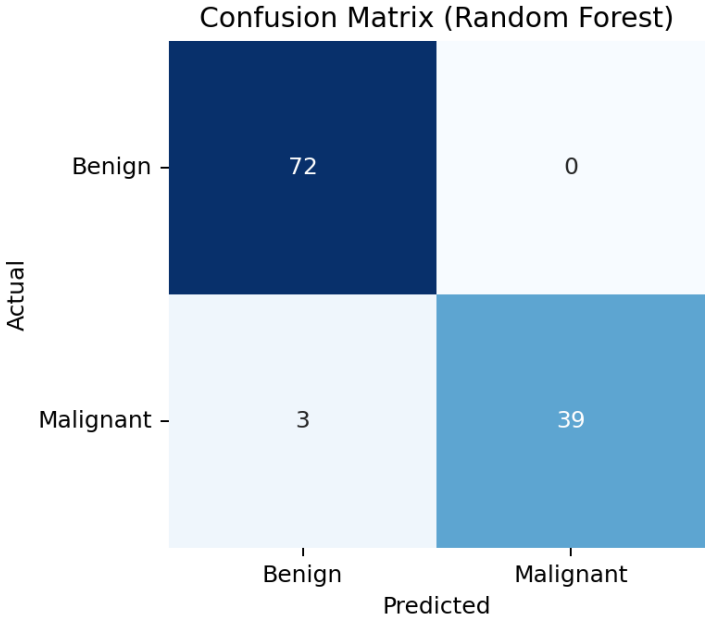

Figure 5: Random forest confusion matrix.

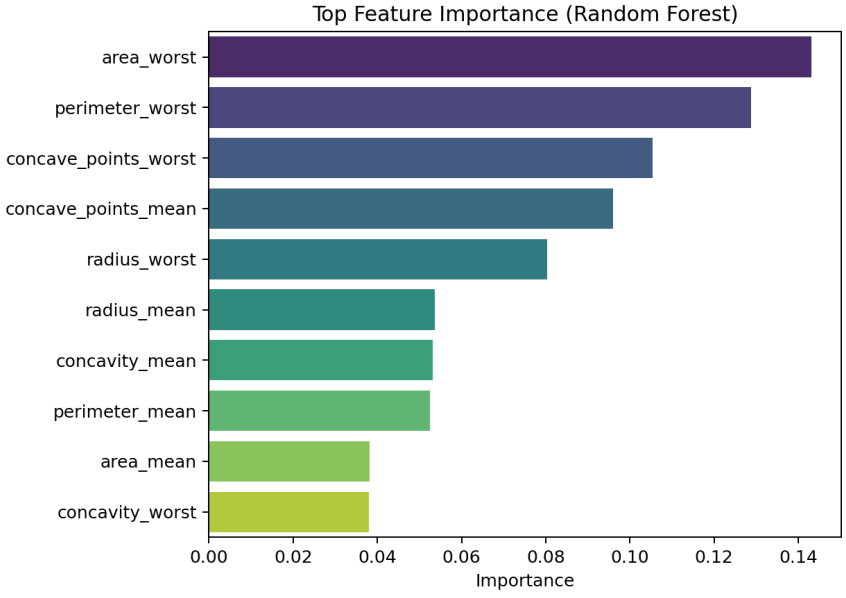

Figure 6: Top feature importance from random forest.