Breast Cancer Classification – Analysis Report

By Himmler Benitez

Content

1. Main Objective

The objective is to build predictive models capable of distinguishing between malignant and benign tumors based on features extracted from fine-needle aspirates (FNA) of breast masses. The focus is on classification rather than interpretation, with the purpose of evaluating which model provides the most reliable performance for clinical decision support.

2. Data Description

The dataset used in this analysis is the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. It contains 569 patient records, each corresponding to a digitized image of a fine-needle aspirate.

- Target variable: Diagnosis (M = malignant, B = benign).
- Features: 30 numerical attributes describing characteristics of the cell nuclei.
- Data quality: The dataset is complete, with no missing values. The class distribution is moderately imbalanced, with approximately 37% malignant and 63% benign cases.
- Preprocessing: All features were standardized using z-scores to ensure comparability and to support models sensitive to feature scaling.
- Link to the dataset: Breast Cancer Wisconsin (Diagnostic) - UCI Machine Learning Repository

Examples of features distributed in the dataset:

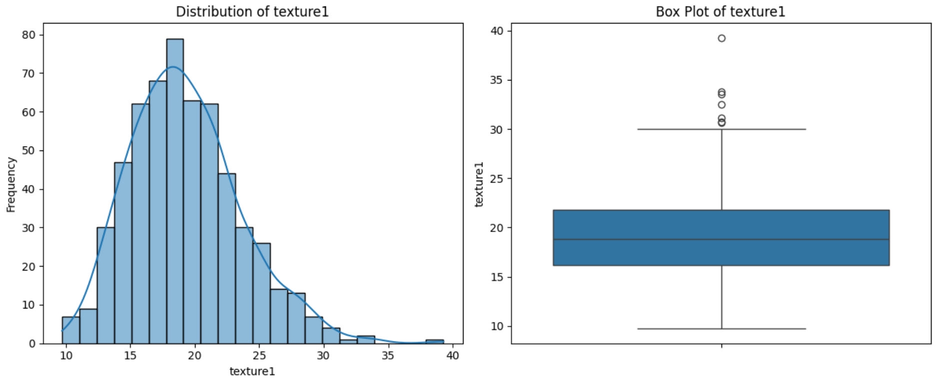

Figure 1. texture1 distribution

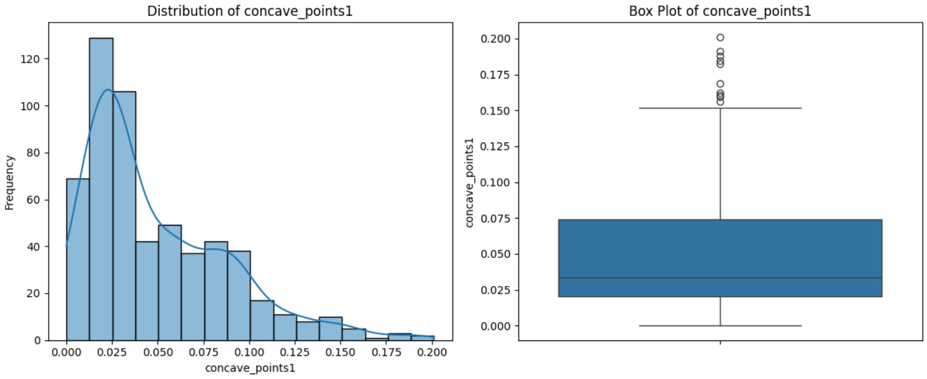

Figure 2. concave_points1 distribution

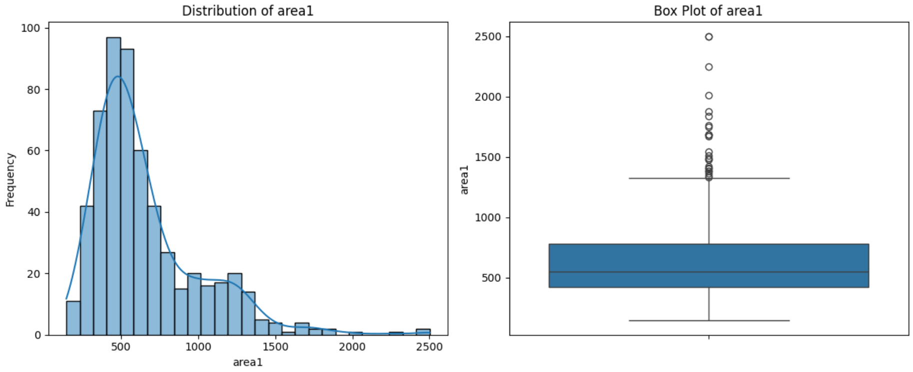

Figure 3. area1 distribution

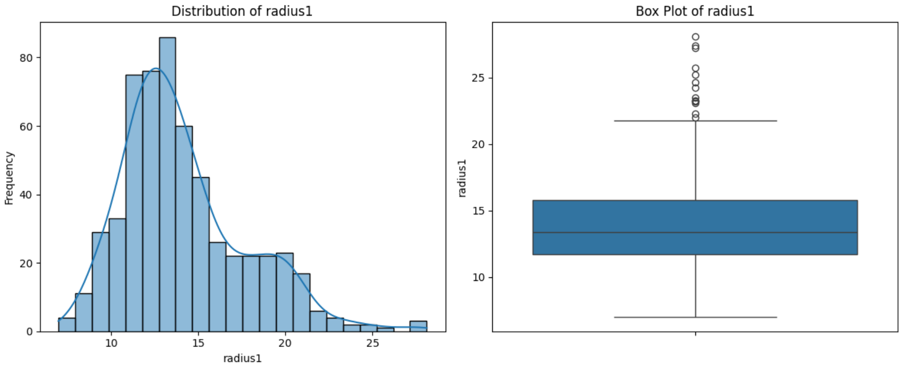

Figure 4. radius1 distribution

The variables in the dataset exhibit non-normal, right-skewed distributions for several key measurements. Features such as area and concave points show a heavy concentration of values near the lower range, with long tails extending toward higher values.

This is the distribution of the target variables: Diagnosis.

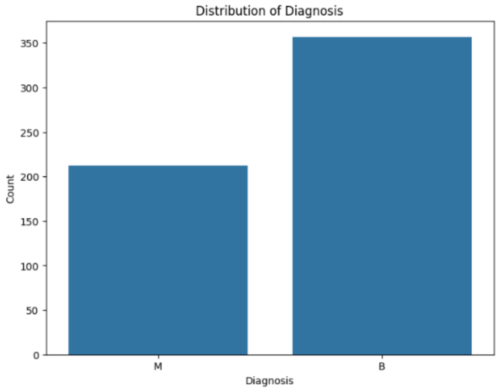

Figure 5. Target variable distribution

There is an imbalance between classes but not significant enough to affect the outcome. But oversampling and/or undersampling techniques can be used to leverage this imbalance.

About outliers in the data, there are features such as radius1 and area1 to have a significant number of outliers with high values. This can influence the performance of the models. The outliers can be removed or applying transformation to the data can help.

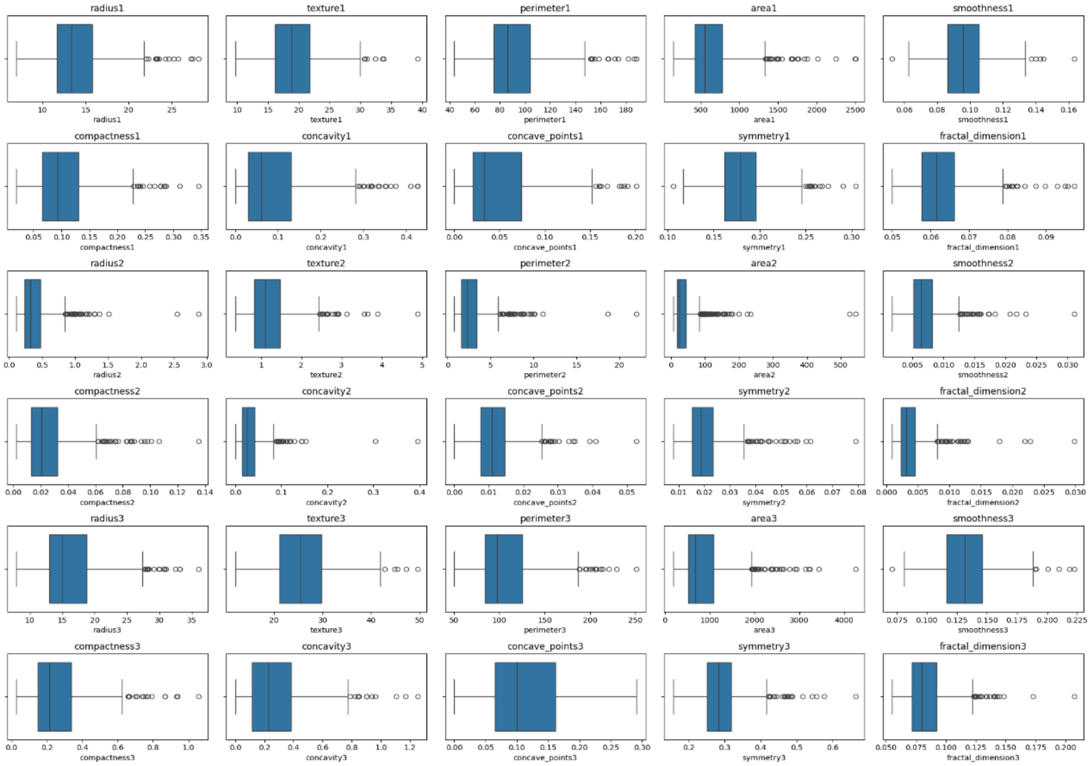

Figure 6. Outliers

3. Data Exploration and Feature Engineering

Initial exploration showed that the dataset is balanced enough to avoid extreme bias toward one class. Correlation analysis revealed strong relationships among some features, consistent with their biological origin. Since the dataset is already clean, minimal preprocessing was required:

- Standardization of features to handle different measurement scales.
- Stratified train–test splits to preserve class proportions.

```
from sklearn.model_selection import train_test_split

# Split the data into training and the rest (test + validation)
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.3, random_state=42, stratify=y)


print("Training set shape:", X_train.shape, y_train.shape)
print("Test set shape:", X_test.shape, y_test.shape)
```

- No additional feature engineering was performed to preserve interpretability and avoid overfitting on such a small dataset.

4. Classifier Models Evaluated

Some pipelines were created to simplify the creation of the models.

```
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC

# Pipeline for Logistic Regression
pipeline_lr = Pipeline([
    ('scaler', StandardScaler()),
    ('logistic_regression', LogisticRegression())
])

# Pipeline for Random Forest
pipeline_rf = Pipeline([
    ('scaler', StandardScaler()),
    ('random_forest', RandomForestClassifier())
])

# Pipeline for Support Vector Machine (SVM)
pipeline_svm = Pipeline([
    ('scaler', StandardScaler()),
    ('svm', SVC())
])

print("Pipelines created: pipeline_lr, pipeline_rf, pipeline_svm")
```

Three different classification models were implemented and evaluated under the same preprocessing and validation procedure:

Logistic Regression

1. Serves as a baseline model.

2. Provides interpretable coefficients and a linear decision boundary.

Tree-Based Ensemble (Random Forest)

3. Captures nonlinear relationships and interactions between features.

4. Includes mechanisms for feature importance estimation.

Support Vector Machine (SVM) with RBF Kernel

5. Effective in handling complex, non-linear patterns.
6. Requires appropriate scaling and hyperparameter tuning.

All models were trained using stratified cross-validation to ensure robustness of the evaluation and modest hyperparameter tuning. Performance metrics included AUROC, accuracy, sensitivity, specificity, and F1-score.

5. Key Findings

The evaluation of the three classifiers showed consistently high performance across all metrics.

Analysis

- All three models demonstrated very high predictive ability, with overall accuracy values above 97%.
- For benign cases, Random Forest achieved perfect recall (1.0000), ensuring that no benign cases were misclassified as malignant.
- For malignant cases, both Random Forest and SVM achieved perfect precision (1.0000), meaning no benign tumors were incorrectly labeled as malignant. Logistic Regression also achieved high precision (0.9836) with identical recall (0.9375) to the other models.
- Random Forest produced the highest F1-scores for both benign (0.9817) and malignant (0.9677) tumors, reflecting its balanced performance across precision and recall.
- Logistic Regression obtained the highest AUC (0.9975), indicating a slightly superior ability to discriminate between classes across decision thresholds.

Performance metrics and ROC Curves

Metric	**Logistic Regression**	**Random Forest**	**SVM**
Overall Accuracy	0.9708	**0.9766**	0.9708
Precision (Benign)	0.9636	0.9640	0.9636
Recall (Benign)	0.9907	**1.0000**	0.9907
F1-Score (Benign)	0.9770	**0.9817**	0.9770
Precision (Malignant)	0.9836	**1.0000**	**1.0000**
Recall (Malignant)	0.9375	0.9375	0.9375
F1-Score (Malignant)	0.9600	**0.9677**	0.9600
AUC	**0.9975**	0.9955	0.9955

Logarithmic Model

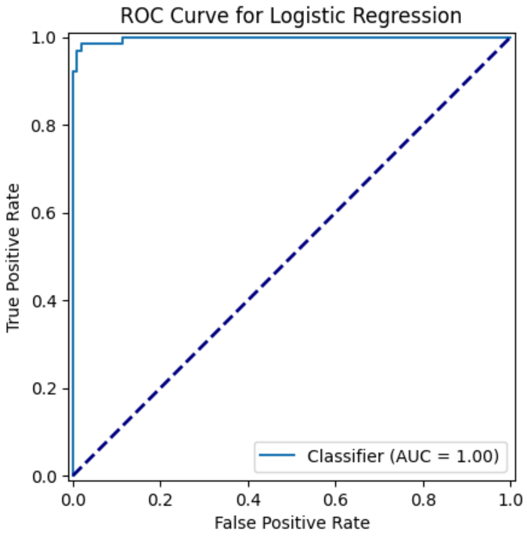

Figure 7. Logarithmic ROC curve

Random Forest Model

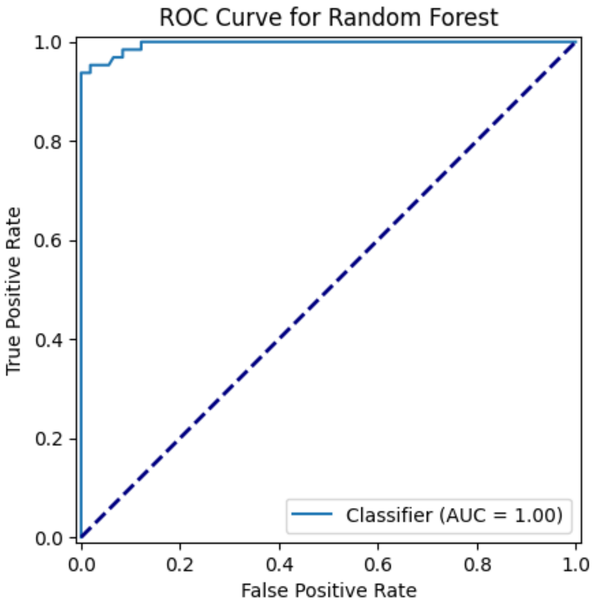

Figure 8. Random Forest Model ROC curve

Support Vector Machine (SVM) Model

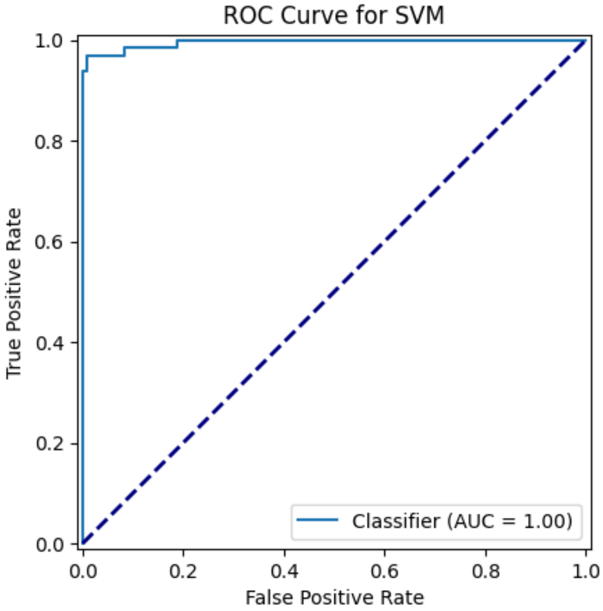

Figure 9. Support Vector Machine ROC curve

Conclusion

Although Logistic Regression marginally outperformed in terms of AUC, the Random Forest model achieved the best balance of accuracy, F1-score, and class-specific precision and recall. Particularly, its perfect precision for malignant tumors and perfect recall for benign tumors make it the most reliable model for this task. Considering the clinical importance of minimizing both false negatives and false positives, Random Forest is recommended as the final model for breast cancer classification in this analysis.

6. Suggestions for Next Steps

- External validation: Apply the model to independent datasets to assess generalization.
- Model calibration: Use isotonic regression or Platt scaling to improve probability estimates.
- Evaluating the relevance of each feature and its correlation with the target variable.

7. Jupyter Notebook

GitHub Repository: eduardoben/Breast-Cancer-Classification