Breast Cancer Wisconsin Dataset Analysis Report

June 9, 2025

1 Introduction

This report presents the analysis of the Breast Cancer Wisconsin dataset using various machine learning techniques. The dataset, sourced from scikit-learn, contains 569 patient cases with 30 numerical features and a binary target variable (0 = malignant, 1 = benign). The analysis includes data exploration, preprocessing, feature selection, model training, and performance comparison across multiple classifiers. The goal is to identify the most effective model for classifying breast cancer cases as malignant or benign.

2 Dataset Overview

2.1 Dataset Description

The dataset comprises:

- **569 rows**: Each representing a unique patient case.
- **31 columns**: 30 numerical features (e.g., mean radius, worst texture) and 1 binary target column (0 = malignant, 1 = benign).
- **Feature Types**: All features are continuous (float64), and the target is binary (int64).
- **No Missing Values**: The dataset is complete, as confirmed by the `df.info()` output.

2.2 Statistical Summary

A statistical summary revealed:

- Features like `mean area` (skewness = 1.64, kurtosis = 3.61) and `area error` (skewness = 5.43, kurtosis = 48.77) exhibit high skewness and kurtosis, indicating non-normal distributions.
- Several features, such as `radius error` and `perimeter error`, show significant outliers, as visualized through boxplots.

3 Data Preprocessing

3.1 Train-Test Split

The dataset was split into training (80%) and testing (20%) sets using scikit-learns `train_test_split` with a random state of 0 for reproducibility.

3.2 Feature Scaling

Features were standardized using `StandardScaler` to ensure zero mean and unit variance, which is critical for models like Logistic Regression, SVM, and KNN.

3.3 Feature Selection

LassoCV (L1 regularization with 5-fold cross-validation) was applied to select relevant features. This reduced the feature set from 30 to 27 by eliminating features with zero coefficients, improving model efficiency and reducing overfitting risk.

4 Model Training and Evaluation

4.1 Logistic Regression (Baseline)

A Logistic Regression model was trained with `max_iter=500` and achieved:

- **Accuracy**: 0.956
- **Precision**: 0.98 (benign), 0.92 (malignant)
- **Recall**: 0.94 (benign), 0.98 (malignant)
- **F1 Score**: 0.96 (weighted average)

Despite a convergence warning, the model performed well, indicating robust classification capability.

4.2 Model Comparison with Pipeline and Hyperparameter Tuning

Five classifiers were evaluated using a `Pipeline` with `StandardScaler` and `GridSearchCV` (5-fold cross-validation) to optimize hyperparameters. The models and their best parameters are:

- **SVM**: `{C: 1, kernel: rbf}`
- **Logistic Regression**: `{C: 1}`
- **KNN**: `{n_neighbors: 5}`
- **Random Forest**: `{n_estimators: 20, max_depth: 10}`
- **Decision Tree**: `{criterion: entropy, max_depth: 5}`

The performance metrics on the test set are summarized below:

Table 1: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1 Score
SVM	0.982	0.971	1.000	0.985
Logistic Regression	0.965	0.970	0.970	0.970
KNN	0.956	0.931	1.000	0.964
Random Forest	0.947	0.955	0.955	0.955
Decision Tree	0.912	0.983	0.866	0.921

5 Insights

- **SVM (RBF Kernel)** outperformed all models with an F1 Score of 0.985, achieving perfect recall (1.0) for the benign class, making it ideal for medical applications where missing positive cases is critical.
- **Logistic Regression** provided stable performance (F1 = 0.970) with lower computational complexity, suitable for scenarios requiring interpretability.
- **KNN** achieved perfect recall but lower precision (0.931), indicating potential false positives, which could be acceptable in recall-sensitive contexts.
- **Random Forest** offered balanced performance but was slightly outperformed by SVM and KNN.
- **Decision Tree** had the lowest performance (F1 = 0.921), likely due to overfitting or underfitting at the chosen depth.

6 Conclusion

The analysis demonstrates that SVM with an RBF kernel is the most effective model for classifying breast cancer cases in this dataset, balancing high accuracy, precision, and recall. Logistic Regression and KNN are viable alternatives, with KNN excelling in recall-sensitive scenarios. Feature selection via LassoCV effectively reduced dimensionality, enhancing model efficiency. Future work could explore handling outliers, addressing skewness through transformations, or testing ensemble methods to further improve performance.