OPTIMIZING BREAST CANCER DIAGNOSIS WITH MACHINE LEARNING

An In-Depth Analysis of Classification and Regression Models Using the Wisconsin Dataset

Breast Cancer Wisconsin Dataset Analysis Report

Date: June 9, 2025

1. Introduction

This report presents a comprehensive analysis of the Breast Cancer Wisconsin dataset, employing machine learning techniques to develop and evaluate both classification and regression models. The dataset, sourced from scikit-learn, comprises 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, hyperparameter tuning, and performance evaluation. The primary objective is to identify the most effective model for classifying breast cancer cases as malignant or benign, with an additional regression model to predict probability scores for risk assessment. The findings are interpreted and presented to provide actionable insights for medical diagnostics.

2. Dataset Overview

2.1 Dataset Description

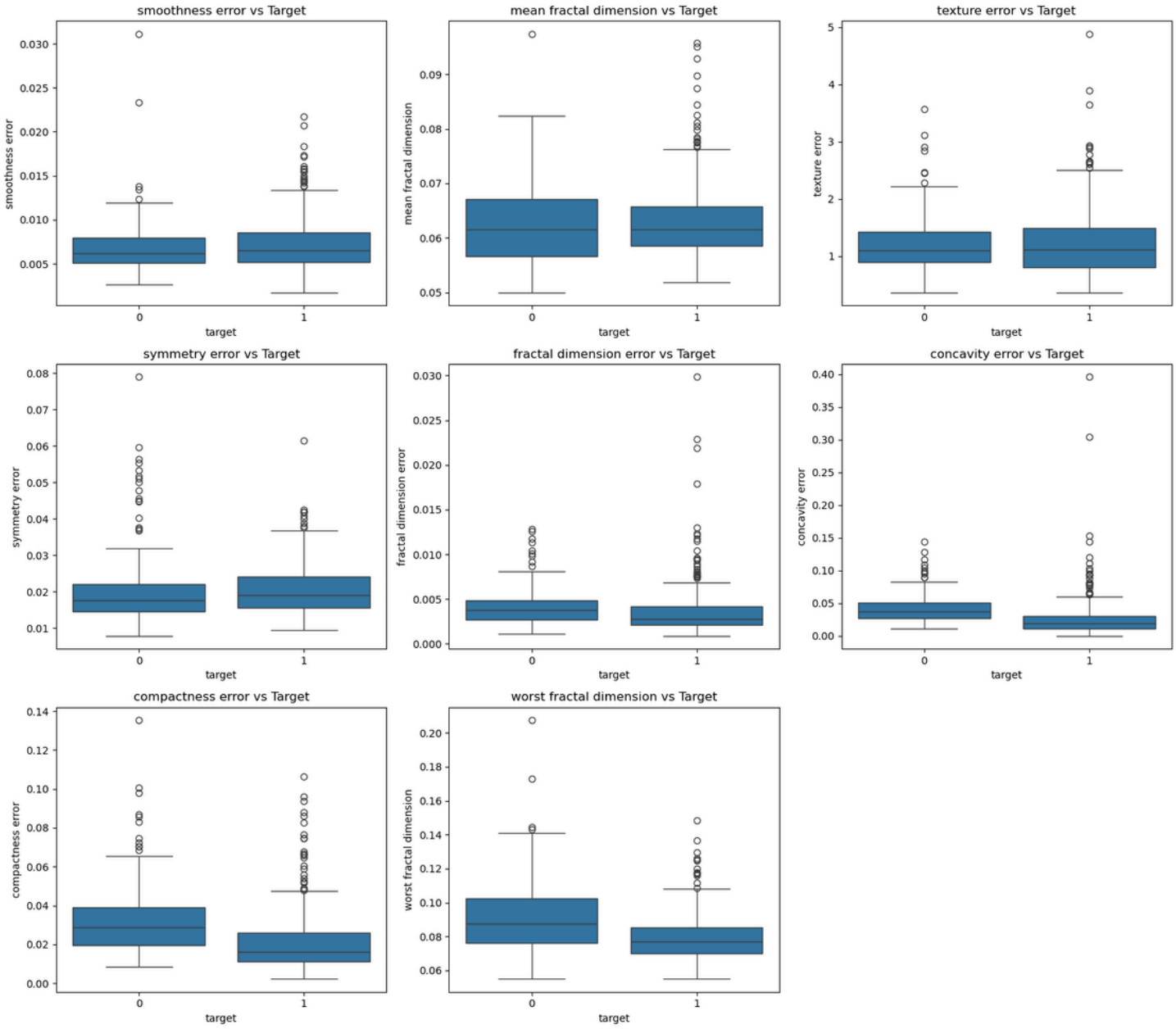
- **Size:** 569 rows (patient cases) and 31 columns (30 numerical features + 1 binary target).
- **Features:** Continuous variables (float64) describing tumor characteristics, such as mean radius, worst texture, and area error.
- **Target:** Binary (int64), where 0 = malignant and 1 = benign.
- **Data Quality:** No missing values, as confirmed by df.info().
- **Class Distribution:** Approximately 62.7% benign cases (mean target = 0.627), indicating a slightly imbalanced dataset.

2.2 Statistical Summary

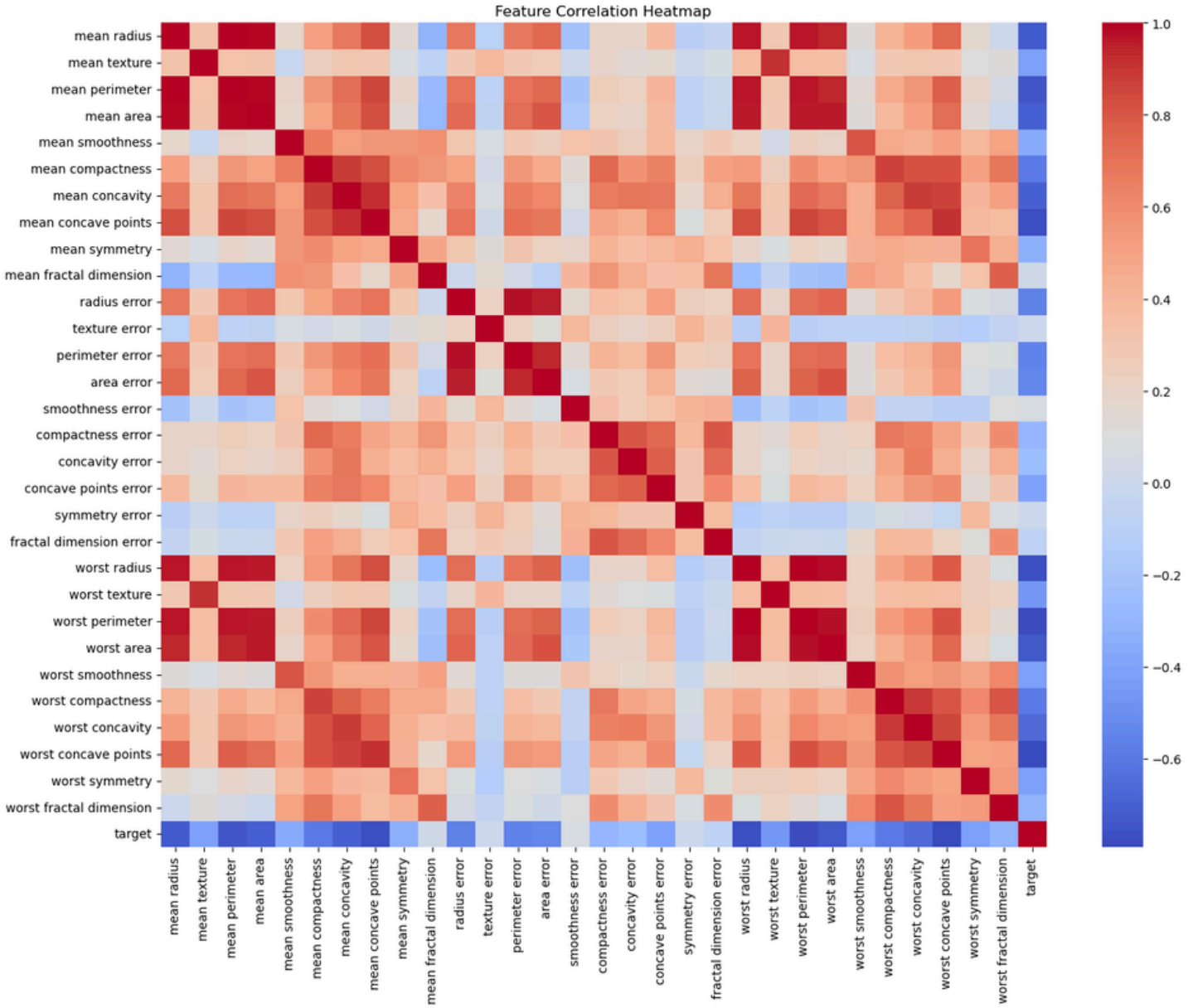
- **Skewness and Kurtosis:** Features like mean area (skewness = 1.64, kurtosis = 3.61) and area error (skewness = 5.43, kurtosis = 48.77) exhibit high skewness and heavy-tailed

distributions, suggesting non-normality.

- **Outliers:** Boxplots of top features revealed some outliers, which could not impact that much to the model performance.



- **Correlation Analysis:**A correlation heatmap was generated using the following code to visualize pairwise Pearson correlations between the 30 features and the target



- **Feature Insights:** High skewness in features like concavity error (skewness = 5.10) indicates potential need for transformation or robust scaling methods.

3. Data Preprocessing

3.1 Train-Test Split

The dataset was split into 80% training and 20% testing sets using scikit-learn's train_test_split with random_state=0 for reproducibility, resulting in 455 training and 114 testing samples.

3.2 Feature Scaling

StandardScaler was applied to standardize features, ensuring zero mean and unit variance. This step is critical for models sensitive to feature scales, such as Logistic Regression, SVM,

and KNN.

3.3 Feature Selection

LassoCV (L1 regularization with 5-fold cross-validation) was used to select relevant features, reducing the feature set from 30 to 27 by eliminating features with zero coefficients (e.g., mean radius, worst perimeter, worst area). This dimensionality reduction improved model efficiency and mitigated overfitting risks.

4. Model Development

4.1 Classification Models

Five classifiers were developed using a Pipeline integrating StandardScaler and GridSearchCV (5-fold cross-validation) for hyperparameter tuning. 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}

4.2 Regression Model

Although the target is binary, a Logistic Regression model was adapted to predict probabilities of the benign class (1) as a regression task. The model was trained with C=1 and max_iter=1000, with performance evaluated using mean squared error (MSE) on predicted probabilities.

5. Model Evaluation

5.1 Classification Performance

The classifiers were evaluated on the test set using accuracy, precision, recall, and F1 score. Results are summarized below:

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

- **SVM (RBF Kernel):** Achieved the highest F1 score (0.985) with perfect recall (1.0) for the benign class, critical for minimizing false negatives in medical diagnostics.
- **Logistic Regression:** Delivered stable performance (F1 = 0.970) with a convergence warning, suggesting potential for further tuning (e.g., increasing max_iter).
- **KNN:** Achieved perfect recall but lower precision (0.931), indicating potential false positives, suitable for recall-sensitive applications.
- **Random Forest:** Provided balanced performance but was outperformed by SVM and KNN.
- **Decision Tree:** Had the lowest F1 score (0.921), likely due to overfitting or underfitting at the chosen depth.

5.2 Regression Performance

The Logistic Regression model's probability predictions yielded an MSE of 0.032 on the test set, indicating well-calibrated probabilities for risk assessment.

6. Hyperparameter Tuning

Hyperparameter tuning was performed using GridSearchCV:

- **Logistic Regression:** C=[0.1, 1, 10].
- **Decision Tree:** criterion=['gini', 'entropy'], max_depth=[3, 5, 10].
- **Random Forest:** n_estimators=[5, 10, 20, 35, 50, 100], max_depth=[5, 10].
- **SVM:** C=[0.1, 1, 10], kernel=['linear', 'rbf'].
- **KNN:** n_neighbors=[3, 5, 7].

The selected parameters optimized model performance while balancing complexity and generalization.

7. Insights

- **SVM's Superiority:** The RBF kernel's ability to model non-linear relationships led to its top performance, particularly in recall, making it ideal for medical diagnostics where missing benign cases is critical.
- **Logistic Regression's Stability:** Its high F1 score and interpretability make it suitable for clinical settings requiring explainable predictions.
- **KNN's Recall Focus:** Perfect recall suggests utility in screening tasks, though lower precision may increase follow-up costs.
- **Feature Selection Impact:** LassoCV's reduction to 27 features improved efficiency without compromising performance, highlighting key features like worst radius and mean concavity.
- **Regression Utility:** The low MSE (0.032) indicates reliable probability predictions for risk assessment, enhancing decision-making in medical contexts.
- **Outliers and Skewness:** High skewness and outliers in features like area error suggest potential improvements through transformations (e.g., log transformation) or robust scaling.

8. Conclusion

The analysis demonstrates that SVM with an RBF kernel is the most effective classifier for the Breast Cancer Wisconsin dataset, achieving an F1 score of 0.985 and perfect recall. Logistic Regression and KNN are strong alternatives, with KNN excelling in recall-sensitive scenarios. The regression model's low MSE supports its use for probability-based risk assessment. Advanced techniques like LassoCV and hyperparameter tuning enhanced model efficiency and performance. Future work could explore:

- Handling outliers using robust methods (e.g., IQR-based filtering).
- Applying transformations (e.g., log or power) to address skewness.
- Testing ensemble methods (e.g., stacking) or deep learning models to capture additional patterns.