Machine Learning Project Report: Breast Cancer Prediction

Date: June 23, 2025
To: Professor Jagdeesh Kakarla
From: Mohammed Sarif Sheikh
Subject: Development and Evaluation of Machine Learning Models for Breast Cancer Prediction

1. Introduction

This report details the development and evaluation of machine learning models for the prediction of breast cancer (classifying tumors as either benign or malignant). The primary objective was to build and assess the performance of two distinct classification algorithms – Logistic Regression and a Multilayer Perceptron (MLP) Classifier – to determine their effectiveness in this critical medical diagnostic task.

2. Methodology

The project followed a standard machine learning pipeline, encompassing data acquisition, preprocessing, model training, and evaluation.

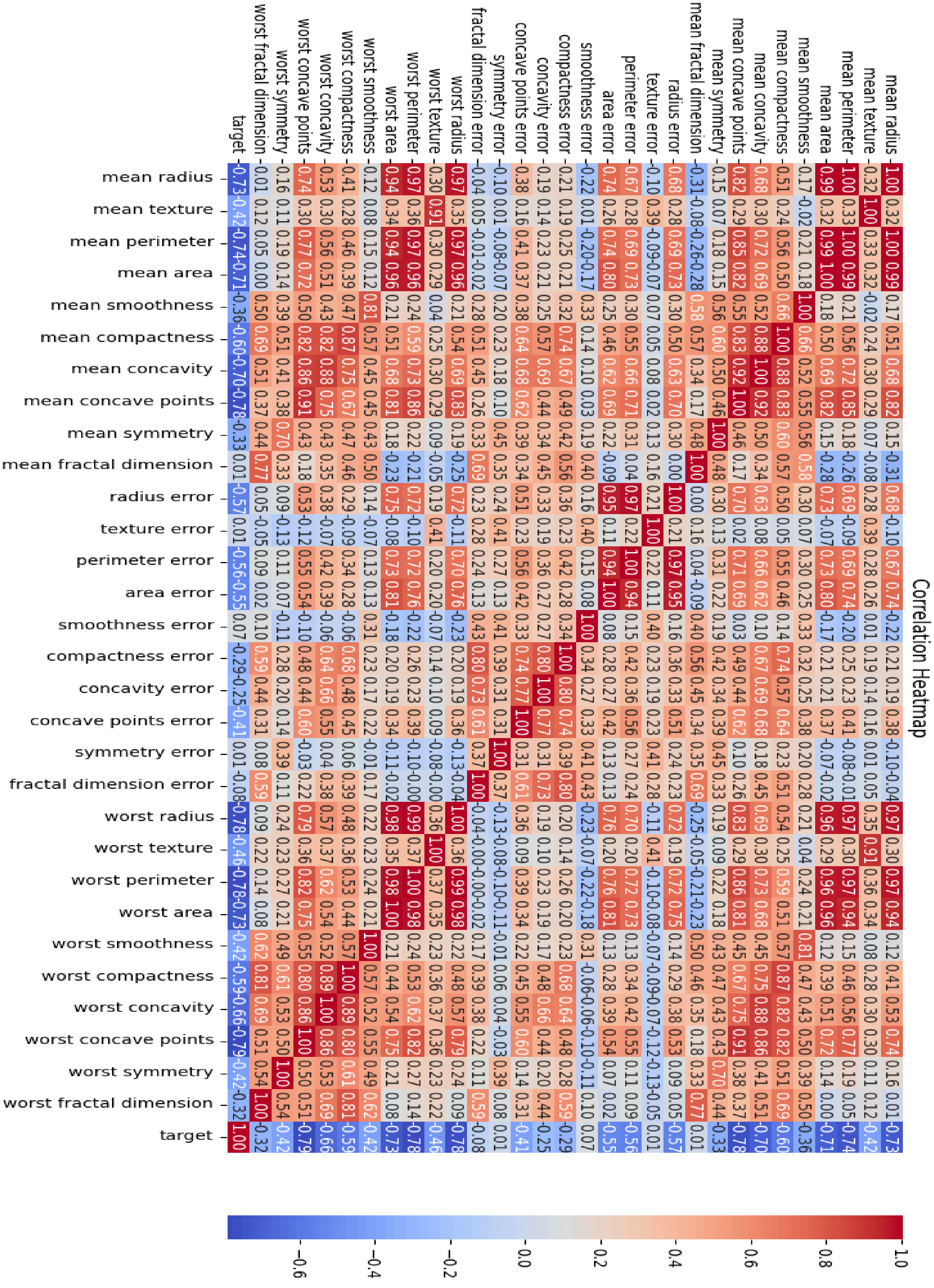
2.1 Data Acquisition

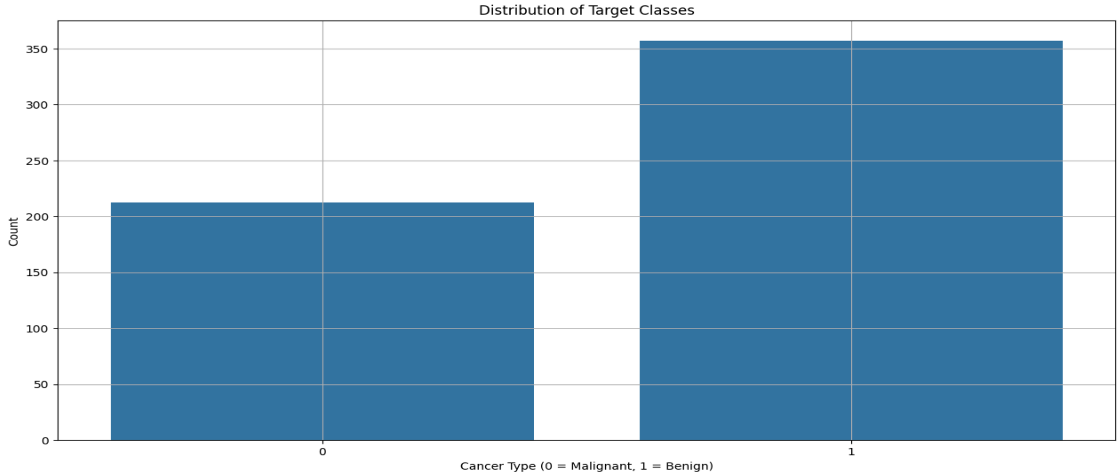
The dataset used for this project was the **Breast Cancer Wisconsin (Diagnostic) Dataset**, obtained from scikit-learn's load_breast_cancer utility. This dataset contains features computed from digitized images of a fine needle aspirate (FNA) of a breast mass, describing characteristics of the cell nuclei. The target variable indicates whether a tumor is benign (0) or malignant (1).

2.2 Exploratory Data Analysis (EDA)

Initial data exploration revealed the following:

- The dataset comprises 569 instances with 30 features and a target variable.
- No missing values were found across any of the features, simplifying the preprocessing stage.
- A correlation heatmap was generated to visualize relationships between all features, and specifically, feature correlations with the target variable were analyzed.


Correlation Heatmap
mean radius
mean texture
mean perimeter
mean area
mean smoothness
mean compactness
mean concavity
mean concave points
mean symmetry
mean fractal dimension
radius error
texture error
perimeter error
area error
smoothness error
compactness error
concavity error
concave points error
symmetry error
fractal dimension error
worst radius
worst texture
worst perimeter
worst area
worst smoothness
worst compactness
worst concavity
worst concave points
worst symmetry
worst fractal dimension
target
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0



2.3 Feature Selection

To reduce dimensionality and focus on the most impactful features, a correlation-based feature selection approach was employed. The top 9 features most highly correlated (by absolute value) with the 'target' variable were selected for model training. These selected features were saved as selected_features.pkl to ensure consistency in deployment.

2.4 Data Preprocessing

- **Scaling:** All selected features were scaled using StandardScaler. This step is crucial for algorithms sensitive to feature magnitudes, such as neural networks and logistic regression, ensuring that features with larger numerical ranges do not unduly influence the model. The fitted scaler was saved as scaler.pkl.
- **Data Splitting:** The preprocessed data was split into training and testing sets using an 80/20 ratio (test_size=0.2). To maintain the original class distribution in both sets, stratify=y was applied, which is essential for classification tasks to prevent biased evaluation. A random_state=45 was set for reproducibility.

3. Models Used

Two distinct classification models were trained and evaluated:

3.1 Logistic Regression

A Logistic Regression model was chosen as a robust and interpretable baseline classifier. It is a linear model used for binary classification that estimates probabilities of class membership.

3.2 Multilayer Perceptron (MLP) Classifier

A Multilayer Perceptron (MLP) Classifier, a type of artificial neural network, was implemented for its ability to learn complex non-linear relationships. The specific configuration was:

- **Hidden Layers:** Two hidden layers with sizes (64, 32). This means the first hidden layer contains 64 neurons, and the second contains 32 neurons.
- **Activation Function:** The relu (Rectified Linear Unit) activation function was used for all neurons in the hidden layers. The ReLU function, $f(x)=\max(0,x)$, helps introduce non-linearity and mitigates vanishing gradients.
- **Maximum Iterations:** max_iter was set to 1000, allowing sufficient training epochs for convergence.
- **Random State:** random_state=45 was set for reproducibility.

Both trained models were saved as logistic_regression_model.pkl and mlp_classifier_model.pkl respectively.

4. Results and Evaluation

The models' performance was assessed on the unseen test set using common classification metrics: Accuracy, Precision, Recall, F1-Score, and a Confusion Matrix.

4.1 Logistic Regression Evaluation

Metric	**Value**
Accuracy	0.9474
Precision	0.9459
Recall	0.9722
F1 Score	0.9589

Confusion Matrix:

```
[[38  4]
 [ 2 70]]
```

- **True Negatives (TN):** 38 (Correctly predicted Benign)
- **False Positives (FP):** 4 (Incorrectly predicted Malignant - Type I error)
- **False Negatives (FN):** 2 (Incorrectly predicted Benign - Type II error)
- **True Positives (TP):** 70 (Correctly predicted Malignant)

4.2 MLP Classifier Evaluation

Metric	Value
Accuracy	0.9561
Precision	0.9467
Recall	0.9861
F1 Score	0.9660

Confusion Matrix:

```
[[38  4]
 [ 1 71]]
```

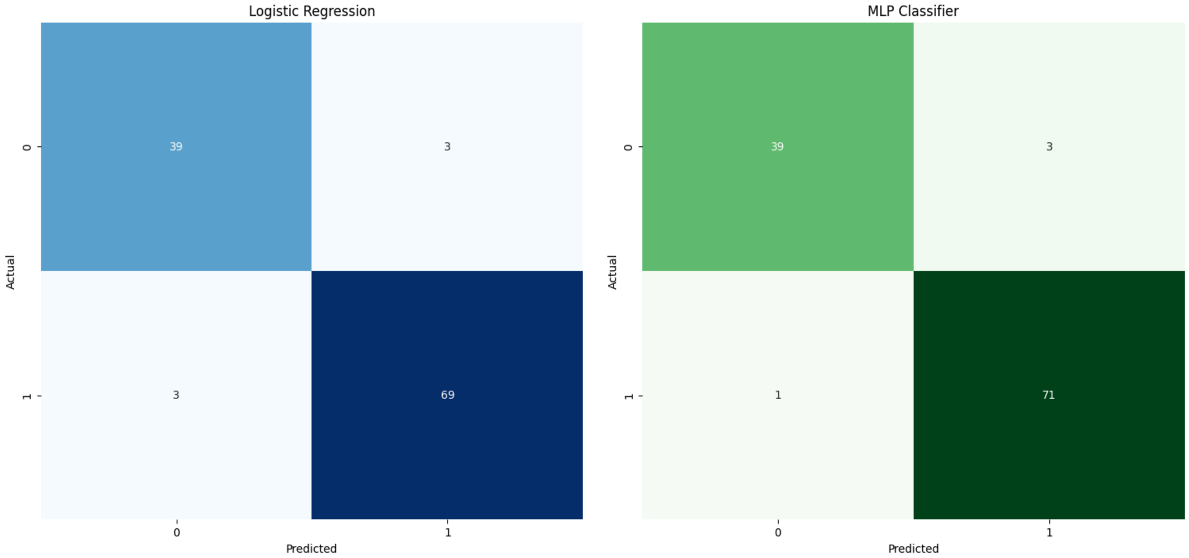
- **True Negatives (TN):** 38 (Correctly predicted Benign)
- **False Positives (FP):** 4 (Incorrectly predicted Malignant - Type I error)
- **False Negatives (FN):** 1 (Incorrectly predicted Benign - Type II error)
- **True Positives (TP):** 71 (Correctly predicted Malignant)

4.3 Interpretation and Comparison

Both models demonstrate strong performance in classifying breast cancer. The **MLP Classifier, however, shows a slight but significant advantage**, particularly in:

- **Accuracy:** Marginally higher (0.9561 vs. 0.9474).
- **Recall:** Substantially higher (0.9861 vs. 0.9722). This is critical in medical diagnosis as it signifies the model's ability to correctly identify nearly all actual malignant cases, thus minimizing "missed diagnoses" (False Negatives). The MLP only missed 1 malignant case compared to Logistic Regression's 2.
- **F1-Score:** Reflects a better balance between precision and recall for the MLP.

Both models exhibit the same number of False Positives (4), meaning they produce a similar rate of "false alarms" (benign cases incorrectly flagged as malignant), which would lead to further, albeit unnecessary, investigations.



5. Conclusion

The developed machine learning pipeline for breast cancer prediction is highly effective. Both Logistic Regression and the MLP Classifier achieved excellent results on the Breast Cancer Wisconsin (Diagnostic) dataset. The **MLP Classifier, with its neural network architecture, demonstrated superior performance, especially in Recall**, making it a more desirable model for this specific application due to its reduced rate of false negatives.

The models, along with the data scaler and selected features, have been successfully saved, enabling their seamless integration into a production environment or a user-friendly application.