Breast Cancer Prediction - Logistic Regression Explanation

This document explains key components and methods used in the Breast Cancer Prediction project using Logistic Regression.

1. Logistic Regression:

A supervised machine learning algorithm used for binary classification. It uses the sigmoid function to convert linear outputs into probabilities between 0 and 1.

2. LabelEncoder:

Used to convert categorical labels ('M', 'B') into numeric form (1 for Malignant, 0 for Benign) using: LabelEncoder().fit_transform(df['diagnosis'])

3. fit_transform() vs transform():

- fit_transform(): Computes the required statistics and applies the transformation.
- transform(): Applies the already computed transformation (e.g., using training data stats on test data).

4. StandardScaler:

Scales features to have mean = 0 and standard deviation = 1.

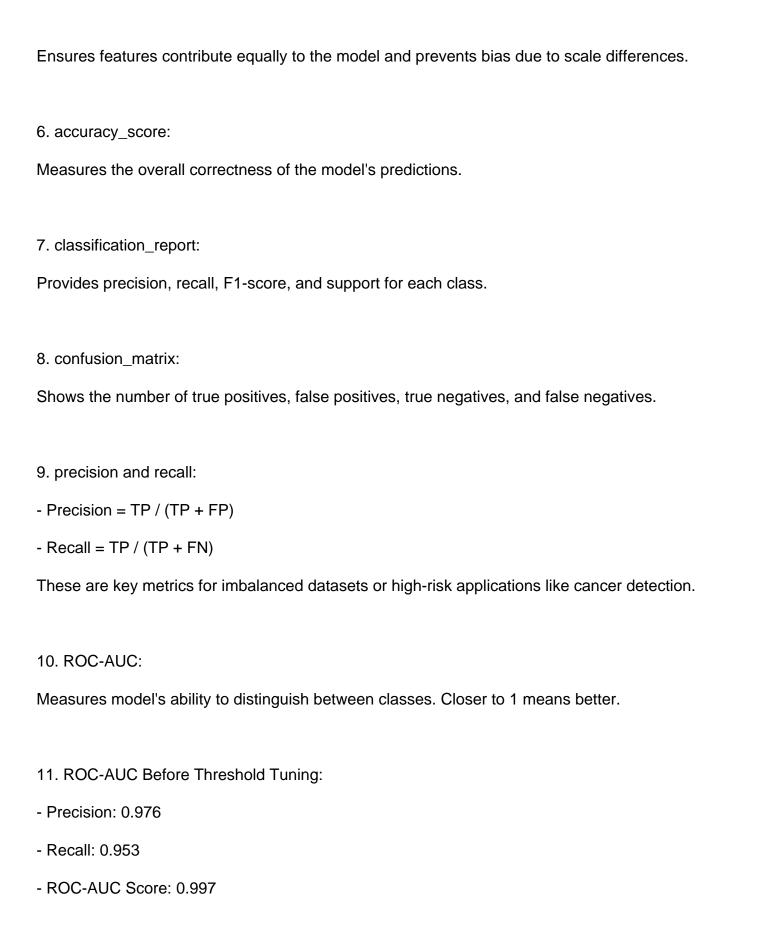
This helps improve model convergence and performance.

Example:

- Original: [10, 20, 30]

- Standardized: [-1.22, 0, 1.22]

5. Purpose of Standardizing Features:



12. Threshold Tuning and Sigmoid Function:

Sigmoid(z) = $1 / (1 + e^{-z})$, maps scores to probabilities.

By adjusting the threshold (default = 0.5), we can control the trade-off between precision and recall.

A loop was used to evaluate precision, recall, and F1-score at 100 thresholds.

13. ROC-AUC After Threshold Tuning (Optimal Threshold = 0.38):

- Precision: 0.9767

- Recall: 0.9767

- F1 Score: 0.9767

- ROC-AUC: 0.997

Result: Balanced precision and recall, with slightly better performance for the specific business goal.

Conclusion:

This project demonstrates how careful preprocessing, logistic regression modeling, and threshold tuning can significantly improve classification performance in medical datasets.