Breast Cancer Detection using Logistic Regression

By **Astitva Pandey(2201CS20)**

1. Introduction to Logistic Regression

Logistic Regression is a statistical method used for binary classification problems. It models the probability that a given input belongs to a particular class. The output of Logistic Regression is a probability value between 0 and 1, which is then mapped to one of the two classes using a decision threshold (commonly 0.5).

The model uses the logistic (sigmoid) function to map the linear combination of features to a probability score. Unlike linear regression, which predicts continuous values, logistic regression is specifically designed for classification tasks.
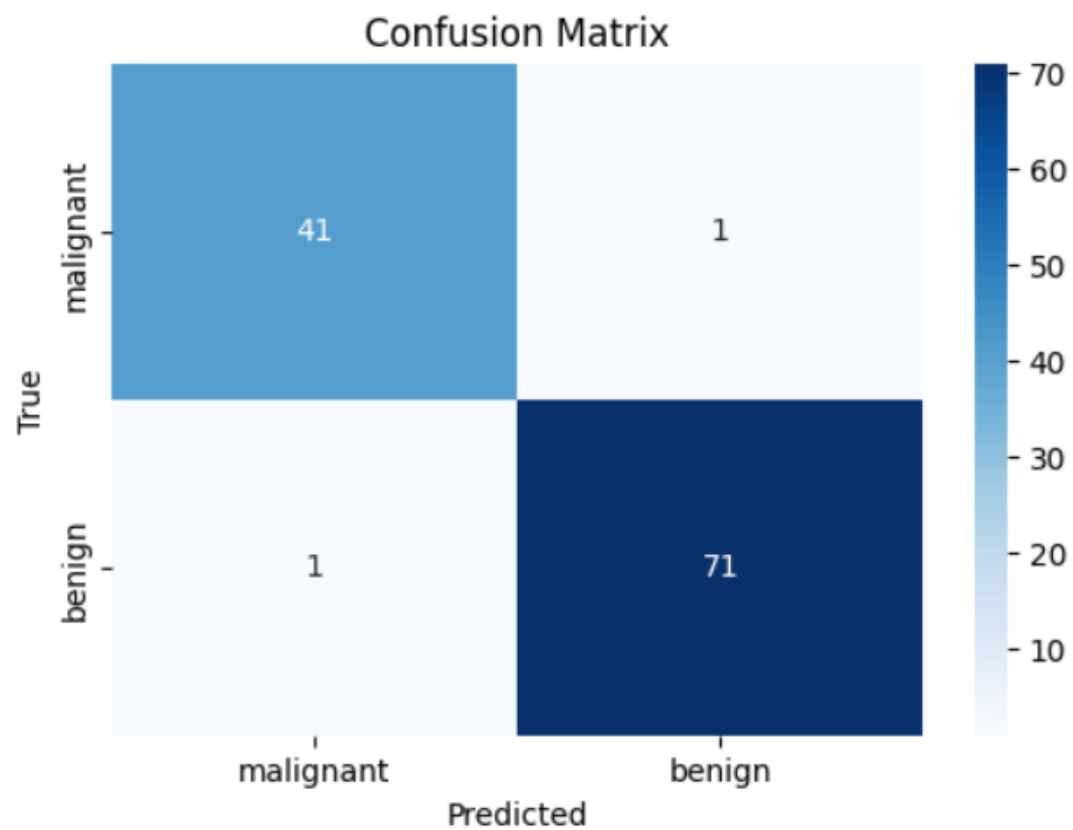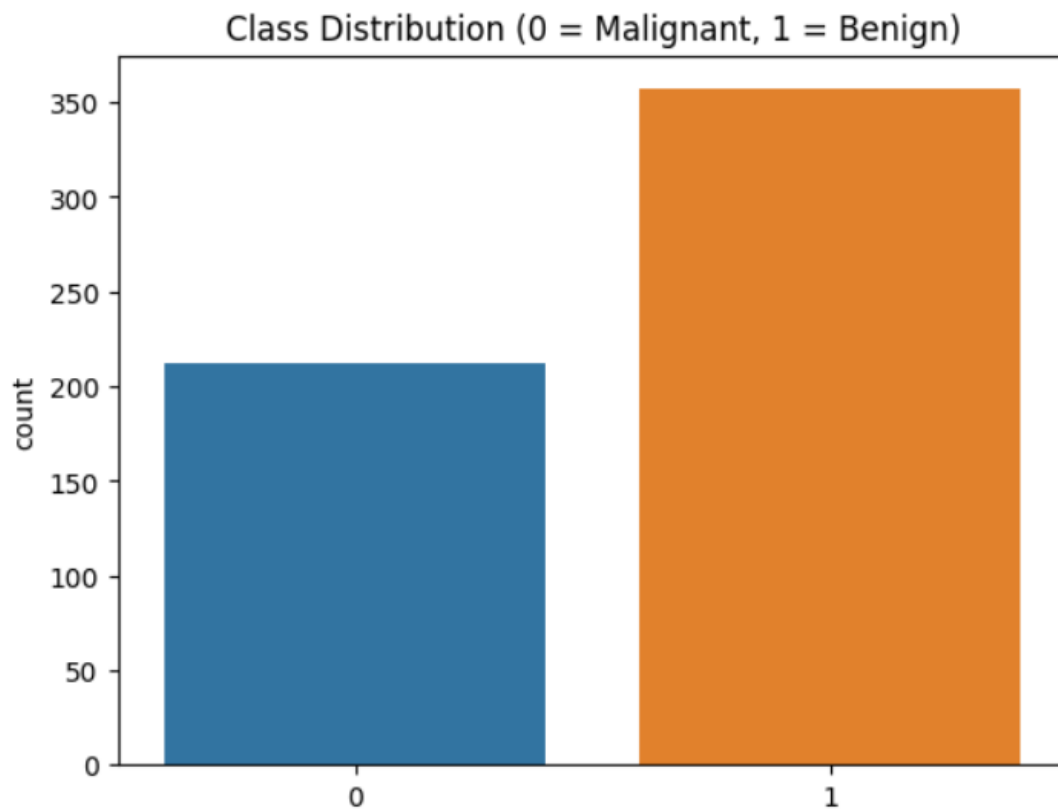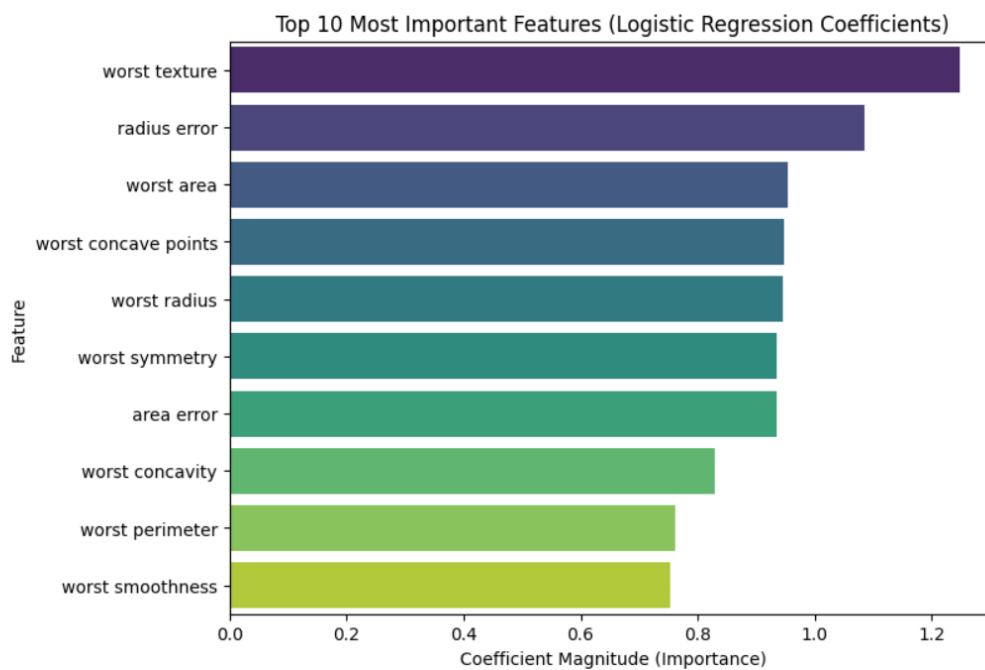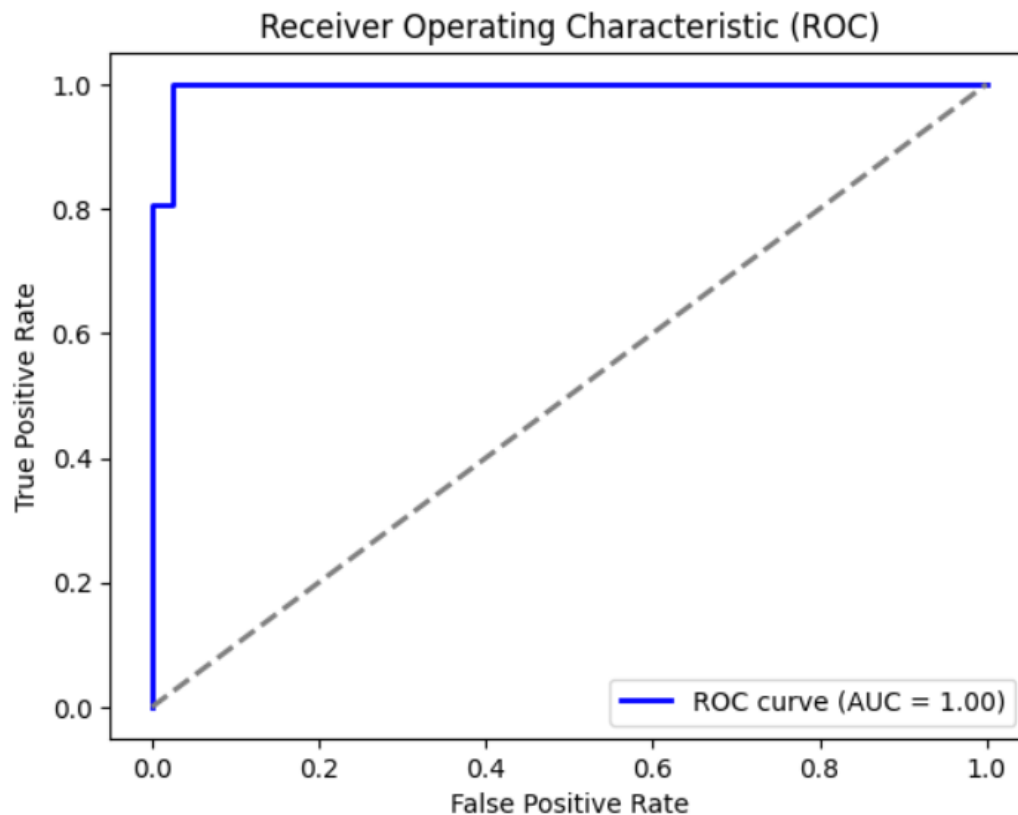
2. Application in Breast Cancer Detection

In this project, Logistic Regression is applied to the Breast Cancer Wisconsin dataset. This dataset contains features computed from digitized images of breast mass. The goal is to classify tumors as either malignant (cancerous) or benign (non-cancerous).
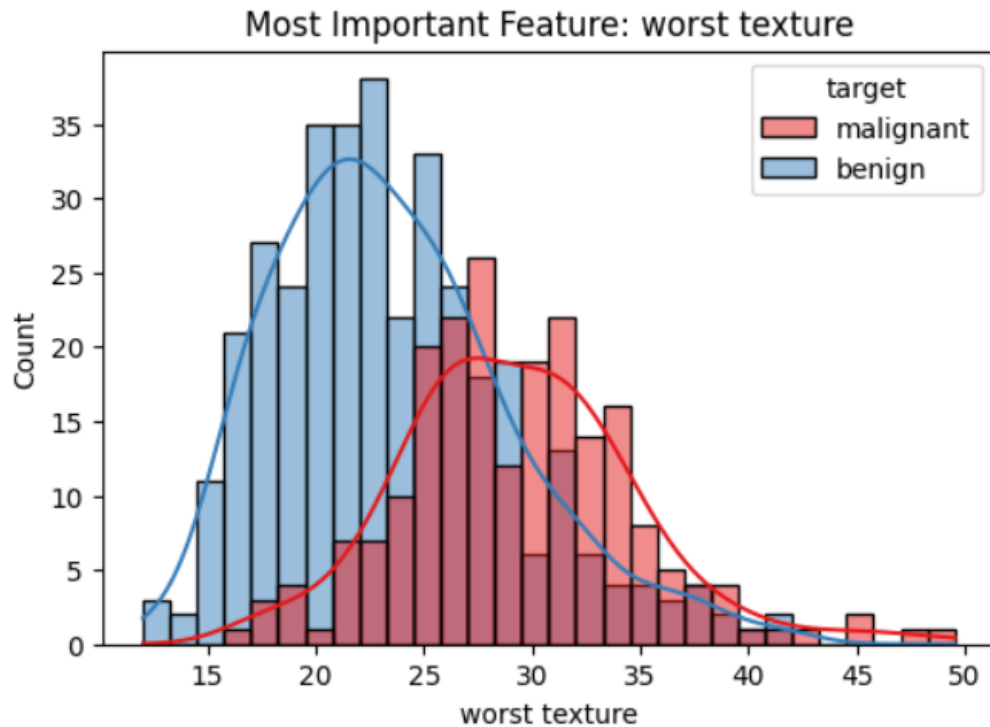
Steps followed in applying Logistic Regression:
1. Data was loaded from the scikit-learn dataset.
2. The dataset was split into training and testing sets.
3. Features were standardized to ensure that all input values have comparable scales.
4. Logistic Regression model was trained on the training data.
5. Model predictions were evaluated using accuracy, confusion matrix, classification report, and ROC curve.

3. Visualizations

Class Distribution (0 = Malignant, 1 = Benign)

Confusion Matrix

Receiver Operating Characteristic (ROC)

ROC curve (AUC = 1.00)

Top 10 Most Important Features (Logistic Regression Coefficients)

Most Important Feature: worst texture

4. Conclusion

Logistic Regression is a simple yet powerful algorithm for binary classification problems. In this case, it successfully distinguishes between malignant and benign tumors with high accuracy. The model also provides interpretable coefficients that help in understanding the importance of each feature.