Mini Project: Logistic Regression on a Real Dataset

Objective:

In this mini-project, you will work with a real-world dataset to apply logistic regression for binary classification. You will go through the complete machine learning workflow, including data preprocessing, model training, evaluation, and interpretation of results.

Step 1: Load the Dataset

- Use the Breast Cancer Wisconsin dataset from Scikit-learn.
- Convert it into a Pandas DataFrame.
- Display the first five rows and check dataset details using `.info()` and `.describe()`.
- Plot class distribution to understand the balance of classes.

Step 2: Data Preprocessing

- Check for missing values and handle them if necessary.
- Standardize the feature variables using `StandardScaler`.
- Split the dataset into training and testing sets (80% training, 20% testing).
- Perform Exploratory Data Analysis (EDA):
  - Plot histograms of numerical features.
  - Use correlation heatmap to visualize feature relationships.

Step 3: Train the Logistic Regression Model

- Initialize and train a logistic regression model on the training dataset.
- Make predictions on the test dataset.
- Display model coefficients to understand feature importance.

Step 4: Evaluate the Model

- Calculate and display the accuracy score.
- Generate a classification report (precision, recall, F1-score, support).
- Plot a confusion matrix and interpret the misclassifications.
- Plot the ROC Curve and calculate AUC score.

Step 5: Interpretation of Results

- What do the evaluation metrics tell you about the model's performance?
- Are there specific features that strongly influence classification?

- How does logistic regression perform compared to a simple baseline model (e.g., always predicting the majority class)?
- Discuss potential improvements such as feature engineering, handling imbalanced data, or trying other classifiers.

Step 6 (Optional): Try Another Dataset

- Repeat the experiment with another binary classification dataset (e.g., Titanic dataset from OpenML or UCI Pima Diabetes dataset).
- Compare logistic regression's performance across different datasets.
- Experiment with hyperparameter tuning (e.g., adjusting `C` in logistic regression for regularization).