Assignment 3: Implement Linear Regression to Predict Wine Quality

Objective:
The goal of this assignment is to implement and evaluate linear regression models on two real-world datasets. You will compare model performance on **Red Wine** and **White Wine** datasets from the UCI Machine Learning Repository, and determine on which dataset linear regression performs better.

Datasets:

Use the following datasets from the UCI Machine Learning Repository:

- Wine Quality - Red Wine
- Wine Quality - White Wine

Each dataset contains physicochemical properties of wines, with the target variable being the **quality** score (integer between 0 and 10).

Assignment Tasks:

You are required to implement your solution in a **Python Jupyter Notebook**. Your notebook should contain the following sections:

1. Data Loading and Preprocessing

- Load both datasets (Red and White wine).
- Check for missing or invalid data and handle appropriately.
- Normalize or scale features if needed.
- Optionally perform feature selection or dimensionality reduction and justify your choices.

2. Data Visualization

- Provide visual exploration of the datasets.
- Use plots (e.g., histograms, boxplots, pairplots, correlation heatmaps) to understand feature distributions and relationships.
- Compare the red and white wine datasets visually.

3. Linear Regression Modeling

- Implement a linear regression model on each dataset.
- Use **train-valid-test split** for evaluation
- Plot regression results (e.g., predicted vs actual quality).

4. Hyperparameter Tuning

- Linear regression doesn't have many hyperparameters, but experiment with:
    - **Learning rate**
    - **Number of Epochs**

5. Evaluation Metrics

- Implement following metrics for both datasets:
    - Mean Absolute Error (MAE)
    - L1 Loss
    - L2 Loss
    - Mean Square Error(MSE)

6. Performance Comparison and Conclusion

- Based on the evaluation metrics, determine on which dataset linear regression (and its regularized variants) performs better.

- Provide a thoughtful discussion on **why** the performance differs between the two datasets (consider data distribution, feature importance, etc.).

Deliverables:

- Submit a clean and well-commented Jupyter Notebook (.ipynb).

- All code, visualizations, and explanations should be included in the notebook.

- Submission deadline: 12.10.25(12:00 am)