Lab 9 Assignment Task: Implementing Random Forest on the Wine Quality Dataset

Dataset

Name: Wine Quality Dataset (Red Wine)
Source: UCI Machine Learning Repository
Download Link: https://archive.ics.uci.edu/ml/datasets/wine+quality
Direct CSV (Red Wine): https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv

Lab Tasks

- Load the Dataset: Load the Wine Quality dataset using pandas. Display the first five rows and check for missing values.
- Data Preparation: Split the data into features (X) and target (y). Divide into training and testing sets (70% train, 30% test).
- Train a Decision Tree Classifier: Use default parameters of DecisionTreeClassifier. Record its accuracy on the test set.
- Train a Random Forest Classifier: Use RandomForestClassifier with n_estimators=100, max_depth=5, random_state=42. Evaluate its accuracy and compare with the Decision Tree.
- Hyperparameter Tuning: Increase n_estimators to 200 and note the effect on accuracy. Remove max_depth limit and see if the model overfits.
- Feature Importance: Extract and display rf.feature_importances_. Identify which chemical properties most affect wine quality.
- Visualization (Optional): Plot a bar chart of feature importances. Compare Decision Tree vs Random Forest results visually.

Exercises

- Why might the Random Forest perform better than the single Decision Tree on this dataset?
- How does the max_depth parameter influence overfitting in this Random Forest model?
- Does increasing the number of trees (n_estimators) always improve accuracy? Why or why not?
- Which features were identified as most important by the Random Forest? Do they make sense chemically (e.g., alcohol, acidity, sugar)?
- How could you further improve model performance using data preprocessing or feature scaling?