Elements of AIML Lab - LAB-5

Name: Yashi Vijayvargiya

SAP: 500124219 | ROLL NO: R2142230340 | BATCH: 11

Topic: K-Fold Cross-Validation

Experiment Question

How does K-Fold Cross-Validation influence the accuracy of various machine learning classification algorithms (Logistic Regression, Decision Tree, Support Vector Machine, K-Nearest Neighbors, and Linear Discriminant Analysis) when applied to the Pima Indians Diabetes Dataset, Wine Quality Dataset, and Breast Cancer Wisconsin Dataset?

Introduction

In machine learning, evaluating model performance is crucial for reliability and accuracy. K-Fold Cross-Validation is widely used to mitigate overfitting by dividing data into K folds, allowing models to be trained on different portions of data while tested on unseen data. This report applies K-Fold Cross-Validation to three datasets: the Pima Indians Diabetes Dataset, Wine Quality Dataset, and Breast Cancer Wisconsin Dataset. We evaluate five algorithms - Logistic Regression, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Linear Discriminant Analysis (LDA) - for predicting outcomes.

Steps of the Code

1. Import Packages: Import libraries - pandas, matplotlib, and sklearn for data manipulation, visualization, and model application.

2. Load Dataset: Load the Pima Indians Diabetes dataset using pandas. Confirm dataset structure.

3. Data Splitting: Divide the data into features (X) and target (y) and split into training and validation sets.

4. Model Definition: Define models - Logistic Regression, LDA, KNN, Decision Tree, and SVM for

comparison.

5. Cross-Validation: Evaluate each model using Stratified K-Fold Cross-Validation with 10 folds for robust testing.

6. Visualization: Use boxplot to compare model accuracy distributions across folds.

7. Conclusion: SVC showed the highest accuracy, making it the best model for diabetes prediction in this dataset.

8. New Patient Prediction: Test SVC model on new patient data to assess diabetes risk prediction.

Conclusion

The Support Vector Classifier (SVC) emerged as the best-performing model with the highest average accuracy across folds, indicating its effectiveness for diabetes prediction in the Pima Indians Diabetes Database. This model can now predict diabetes risk for new patients, providing valuable insights for healthcare assessment.