Lab-5: K-Fold Cross-Validation Analysis

Prepared by: Arjun Neema

SAP ID: 500124012

Roll No: R2142230341

Batch: 11

Topic: K-Fold Cross-Validation

Experiment Objective:

This lab experiment investigates the impact of K-Fold Cross-Validation on the accuracy of several machine learning classification algorithms:

- Logistic Regression
- Decision Tree
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
- Linear Discriminant Analysis (LDA)

The evaluation is conducted using three datasets:

1. Pima Indians Diabetes Dataset
2. Wine Quality Dataset
3. Breast Cancer Wisconsin Dataset

Overview of the Experiment:

Introduction:

K-Fold Cross-Validation is a critical technique for ensuring reliable evaluation of machine learning models by dividing datasets into K subsets. This experiment emphasizes its utility in reducing overfitting and providing robust performance estimates. The aim is to determine the most effective algorithm for the aforementioned datasets.

Steps of the Experiment:

1. Importing Libraries:

Libraries such as Pandas, Matplotlib, and Scikit-learn were utilized for data handling, visualization, and model implementation.

2. Dataset Loading:

Data from the three datasets was loaded and verified for consistency and correctness.

3. Data Splitting:

Data was split into features (X) and target (y). Using train_test_split, subsets for training and validation were created.

4. Model Definition:

Classification algorithms were defined for testing. Each model's performance was assessed based on accuracy.

5. K-Fold Cross-Validation:

Stratified K-Fold Cross-Validation with 10 folds was applied to each model, ensuring robust evaluation across all datasets.

6. Visualization:

A boxplot visualization summarized the accuracy distribution of each model, facilitating comparative analysis.

Results and Key Findings:

1. Algorithm Comparison:

Performance was compared across all models using the accuracy scores obtained during cross-validation.

2. Best Performing Model:

The Support Vector Classifier (SVC) achieved the highest average accuracy, emerging as the top performer for the Pima Indians Diabetes dataset.

3. Practical Validation:

SVC's predictive capability was further tested on a new patient dataset, reinforcing its effectiveness in real-world applications.

Conclusion:

The SVC algorithm demonstrated superior performance and reliability, making it the best choice for predicting outcomes in the tested datasets. Its practical application in diagnosing conditions like diabetes provides valuable insights for healthcare professionals.