AIML Lab - Experiment 5

NAME: Manan Shukla

SAP: 500119574

ROLL NO: R2142230365

BATCH: 12

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 essential for ensuring reliability and accuracy. K-Fold Cross-Validation is a widely used technique for this purpose. It divides the dataset into K subsets or 'folds,' allowing models to train on different data portions while testing on unseen data. This process provides a robust performance estimate and helps prevent overfitting.

Datasets and Classification Algorithms

In this experiment, K-Fold Cross-Validation is applied to the Pima Indians Diabetes Dataset, Wine Quality Dataset, and Breast Cancer Wisconsin Dataset. Five classification algorithms are evaluated:

1. Logistic Regression
2. Decision Tree
3. Support Vector Machine (SVM)
4. K-Nearest Neighbors (KNN)

5. Linear Discriminant Analysis (LDA)

Steps of the Code

1. Import Packages: Libraries such as Pandas, Matplotlib, and Sklearn are imported for data manipulation, visualization, and applying machine learning algorithms.

2. Loading the Dataset: Each dataset is loaded into a DataFrame to facilitate analysis. The structure and type of data are confirmed for correctness.

3. Data Splitting: Features (X) and target variable (y) are separated. Using train_test_split, the data is divided into training and validation sets.

4. Model Definition: The machine learning models are defined: Logistic Regression, LDA, KNN, Decision Tree, Naive Bayes, and SVM.

5. Cross-Validation: Using Stratified K-Fold Cross-Validation (10 folds), each model's accuracy is tested on different parts of the dataset. Accuracy scores for each model are recorded.

6. Visualization: A boxplot of accuracy scores from cross-validation provides a comparative view of model accuracy and variance.

Conclusion

Support Vector Classifier (SVC) showed the best performance with the highest average accuracy, making it the most effective model for diabetes prediction in the Pima Indians Diabetes dataset.