AIML Lab-5 Report: K-Fold Cross-Validation

Objective

The experiment explores how K-Fold Cross-Validation impacts the accuracy of machine learning classification algorithms applied to three datasets: Pima Indians Diabetes Dataset, Wine Quality Dataset, and Breast Cancer Wisconsin Dataset.

The classification algorithms studied include:

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

Introduction

K-Fold Cross-Validation is a robust evaluation technique to estimate the performance of machine learning models while minimizing overfitting. By dividing the data into K subsets (folds), the models are trained on different portions of the data and tested on unseen subsets. The experiment aims to:

- Assess the effectiveness of classification algorithms.
- Determine the most accurate algorithm for the datasets.

Steps Followed

1. **Importing Libraries**: Key Python libraries used include Pandas (data manipulation), Matplotlib (visualization), and Sklearn (model building and evaluation).

2. **Loading the Dataset**: Datasets like the Pima Indians Diabetes dataset were loaded and

structured for preprocessing.

3. **Data Splitting**: Features (X) and target variable (y) were separated. Data was split into training and validation sets using train_test_split.

4. **Model Selection**: Six algorithms were chosen: Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbors, Decision Tree, Naive Bayes, and SVM.

5. **Applying K-Fold Cross-Validation**: A stratified 10-Fold Cross-Validation was employed for model evaluation. Accuracy scores were calculated for all models.

6. **Visualization**: Boxplots were created to compare the models' accuracy distributions.

Key Observations

- **Performance Comparison**: Support Vector Classifier (SVC) demonstrated the highest average accuracy. Boxplots revealed the variance in accuracy across folds, highlighting SVC's stability.

- **Practical Application**: SVC was tested on a new patient dataset, effectively predicting diabetes outcomes based on features like glucose levels, BMI, and age.

Conclusion

The experiment showed that the SVM model performed best for predicting outcomes in the Pima Indians Diabetes dataset. This conclusion underscores its potential for real-world applications in healthcare diagnostics.