Report on K-Fold Cross-Validation in Machine Learning (Lab-5)

K-Fold Cross-Validation in Machine Learning

Name: Chaitanaya Sharma

SAP ID: 500122763

Roll No: R2142230335

Batch: 11

1. Experiment Overview

The lab focuses on the application of K-Fold Cross-Validation to evaluate the accuracy of various machine learning algorithms. The objective is to test the performance of five classification algorithms on three datasets:

- Pima Indians Diabetes Dataset
- Wine Quality Dataset
- Breast Cancer Wisconsin Dataset

2. Introduction

In machine learning, model evaluation is critical to assess reliability. K-Fold Cross-Validation is a method that helps mitigate overfitting by dividing the dataset into K equal parts (folds). The model is trained on K-1 folds and tested on the remaining fold, providing a comprehensive measure of performance.

The algorithms used are:

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

Report on K-Fold Cross-Validation in Machine Learning (Lab-5)

3. Datasets Used

- Pima Indians Diabetes Dataset: Medical data of women for predicting diabetes.
- Wine Quality Dataset: Data related to wine characteristics for quality classification.
- Breast Cancer Wisconsin Dataset: Features related to cancer cells for malignancy prediction.

4. Steps in Experiment

1. Importing Packages: Necessary libraries (Pandas, Matplotlib, Sklearn) were used for analysis.
2. Loading the Dataset: Each dataset was loaded and verified using Pandas.
3. Data Splitting: Features and target variables were separated and split for training/testing.
4. Model Definition: Logistic Regression, LDA, KNN, Decision Tree, Naive Bayes, and SVM were initialized.
5. Cross-Validation: Stratified K-Fold Cross-Validation (10 folds) was used to evaluate models.
6. Visualization: A boxplot was used to compare model accuracy scores.

5. Results

The Support Vector Classifier (SVC) achieved the highest accuracy across the folds, suggesting its effectiveness in predicting outcomes.

6. Practical Application

A new patient's data was tested with the SVC model, successfully predicting the likelihood of diabetes. This highlights the potential of machine learning in healthcare.

7. Conclusion

K-Fold Cross-Validation is a robust method for evaluating models. SVC performed best among the algorithms tested, particularly for medical predictions.