Data Mining Lab (Assignment-07)

B.Tech 6th semester

(Language/Platform: Python)

Imagine you are a Data Scientist working for a botanical research organization. The organization wants to develop an automated system to classify different species of flowers based on certain measurable features. They provide you with the Iris dataset, which contains sepal length, sepal width, petal length, and petal width as features, along with their corresponding species labels (Setosa, Versicolor, and Virginica). Your task is to build and analyze multiple classification models to determine the best-performing approach for accurate species identification.
Using the given scenario, answer the following questions:

1. Building a Decision Tree Model

Load the Iris dataset and implement a Decision Tree classifier using Python. Evaluate the model's performance by computing and displaying its accuracy. How well does this model classify the species?

2. Implementing k-NN Classification

Train a k-Nearest Neighbors (k-NN) classifier on the Iris dataset. Compute and display its accuracy. How does k-NN perform compared to the Decision Tree classifier?

3. Comparing Decision Tree and Random Forest

Implement a Random Forest classifier on the Iris dataset. Compare its performance with the Decision Tree classifier in terms of accuracy and effectiveness. Does the ensemble approach of Random Forest improve classification performance?

4. Support Vector Machine for Classification

Implement an SVM classifier on the Iris dataset and evaluate its accuracy. Compare its performance with the previously implemented models (k-NN, Decision Tree, and Random Forest). Which model gives the best accuracy, and why?

5. Impact of Feature Scaling

Apply feature scaling (such as StandardScaler or MinMaxScaler) to the Iris dataset. Analyze its impact on the accuracy of the k-NN and SVM models. Does scaling improve classification performance? Why is feature scaling important for distance-based algorithms?

6. Cross-Validation for Performance Evaluation

Use cross-validation to assess the performance of Decision Tree, Random Forest, k-NN, and SVM classifiers on the Iris dataset. Compare their accuracies across different folds. Which model maintains consistency in performance?

7. Visualizing Decision Boundaries

Reduce the Iris dataset's dimensionality using Principal Component Analysis (PCA) and plot the decision boundaries of the Decision Tree, k-NN, SVM, and Random Forest classifiers in a 2D space. How well can each model distinguish between the species?

8. Optimizing the Random Forest Model

Perform hyperparameter tuning on the Random Forest classifier using GridSearchCV or RandomizedSearchCV to improve its accuracy. Which hyperparameters influence model performance the most, and what are the best settings?

9. Implementing a Naïve Bayes Classifier

Train a Naïve Bayes classifier on the Iris dataset and compare its accuracy with the other classifiers. Given that Naïve Bayes assumes feature independence, how does it perform on this dataset compared to tree-based and distance-based models?

10. Beyond Accuracy: Evaluating Model Performance

Compute and compare the precision, recall, and F1-score for all implemented classifiers. Which model provides the best balance between precision and recall? Why is it important to analyze these metrics beyond just accuracy?