Open Source Software Lab

Lab Test 2

Wednesday – 3 to 5 PM

Time Duration: 50 Minutes **Maximum Marks: 20Marks**

Note:

• No extra time will be provided for form submissions. Any responses submitted after the deadline willnot be accepted.

• Please create a Word document with your answers, along with screenshots of the output. Upload aword file on Google Classroom which contains the following:

- Link to your GitHub account
- Codes for questions 1 along with the URL of the repository

• Save your file using the following format: (Batch_Enrollment_StudentName_LabTest_2.docx)

Odd Numbered Systems

Q1. [15 Marks] You are tasked with classifying the Iris dataset using a machine learning model. Perform the following tasks:

1. Load the Iris dataset from sklearn.datasets and display the first 10 rows of the dataset.
2. Provide a summary of the dataset, including the number of instances, features, and target classes. Also, describe each feature in terms of its data type and range (minimum, maximum).
3. Visualize the relationships between the features using a heatmap of the correlation matrix.
4. Split the dataset into training (80%) and testing (20%) sets.
5. Standardize the features using `StandardScaler`.
6. Visualize the distribution of the features (before and after scaling) using histograms or boxplots.
7. Train a **Logistic Regression** model to classify the Iris species.
8. Use cross-validation (e.g., 5-fold) to tune the hyperparameter `C` (regularization strength) and choose the best model based on accuracy.
9. Evaluate the model performance on the test set using appropriate classification metrics (Accuracy, Precision, Recall, F1-score).

Q2. [5 Marks] Write a Python function that takes a NumPy array of strings and returns the longest string from the array, but only if the string has more than 5 characters. If no string has more than 5 characters, return the shortest string.

Even Numbered Systems

Q1. [15 Marks] You are given the Breast Cancer Wisconsin dataset (available in sklearn.datasets). The task is to classify whether a tumor is malignant or benign based on several features. Perform the following tasks:

1. Load the Breast Cancer dataset from sklearn.datasets and display the first 10 rows of the dataset.
2. Display the first 10 rows of the dataset and provide basic statistics for each feature (mean, standard deviation, min, and max).
3. Check for missing values and explain why handling missing values is important for this dataset.
4. Split the data into training (70%) and testing (30%) sets.
5. Apply **StandardScaler** to scale the features.
6. Use a correlation matrix to identify any features that are highly correlated.
7. Build a **K-Nearest Neighbors (KNN)** model to classify the tumors as malignant or benign.
8. Use cross-validation to select the optimal value for `k` (number of neighbors). Report the accuracy of the model for the best `k` value and explain why you chose that value.
9. Evaluate the KNN model using the test set. Report the **Confusion Matrix** and **Classification Report**. Based on the results, interpret the model's performance (e.g., precision, recall, F1-score).

Q2. [5 Marks] Given a list of strings, write a Python function using NumPy to return a new list where each string is reversed and the length of the reversed string is greater than or equal to 5 characters. If a string is shorter than 5 characters, exclude it from the result.