End Sem Open Lab exam

A student requires to attempt the Q. No = (Last 2 digits of Roll NO % 8 +1)

You can refer the manuals for Python syntax.

Plagiarism check of the code submitted will be done. Students who submit the copied code will be awarded Zero marks

1. Load the following dataset P**rostate_cancer.csv**

a. Prepare a new dataset **D** removing the feature 'id' and 'city' from the original dataset. The target label is 'diagnosis_result'.

b. Plot the pairs of numerical features in **D** using scatter plot.

c. Use min-max normalization to the dataset **D**.

d. Apply kNN algorithm with 80:20 training-testing split. The distance measure to be used is Manhattan distance.

e. Plot graph of k versus accuracy where k is the number of neighbours.

2. Load the following dataset **abalone.csv**

a. Prepare a new dataset **D** by keeping all the features except the last feature. The target feature is the first feature.

b. Plot the pairs of numerical features in **D** using scatter plot.

c. Apply kNN algorithm with 80:20 training-testing split. The distance measure to be used is Euclidean distance.

d. Plot graph of k versus precision where k is the number of neighbours.

3. Load the following dataset **car_purchasing_data.csv**

a. Prepare a new dataset **D** by keeping all the features except, 'customer_name', 'customer_email', 'country' and 'gender'. The target feature is 'car_purchase_amount'.

b. Perform standardization on **D**

c. Apply Linear regression with 70:30 training-testing split on **D** and generate the model

d. Plot the regression line over the training dataset as per the obtained model.

e. Plot graph of learning rate **α** versus RMSE for test dataset

4. Load the following dataset **bridges.csv**

a. Prepare a new dataset **D** with categorical values converted into numerical values. The first feature may be removed. The last feature is the target feature. The target feature has many values. The value 'WOOD' remains as it is but the other values may be changed to ‘NON-WOOD’. Thus the dataset becomes 2-class labelled dataset.

b. Missing value in each numerical feature can be replaced with corresponding mean value. Missing value in each categorical feature may be replaced with the corresponding mode value(The value that appears most times in that feature).

c. Plot the dataset and color the samples as per the label.

d. Apply Logistic regression with 80:20 training-testing split and generate the model.

e. Plot graph of training size versus accuracy for test dataset.

5. Load the following dataset **kidney-disease.csv**
    a. Prepare a new dataset **D** with categorical values converted into numerical values. The first feature may be removed.
    b. Plot the dataset and color the samples as per the label. The last feature is the target feature.
    c. Missing value in each numerical feature can be replaced with corresponding mean value. Missing value in each categorical feature may be replaced with the corresponding mode value (The value that appears most times in that feature).
    d. Apply Logistic regression with 70:30 training-testing split and generate the model.
    e. Plot graph of learning rate versus F1-score for test dataset.

6. Load the following dataset **indian-liver-patient.csv**
    a. Prepare a new dataset **D** by converting Boolean features into numerical features.
    b. Use min-max normalization on **D**
    c. Apply SVM classifier with RBF kernel and 70:30 training-testing split to generate the model.
    d. Plot graph with varied values for penalty term C and show how the precision changes.

7. Load the following dataset **Frogs-species.csv**
    a. Prepare a dataset **D** by extracting randomly 5 numerical features and The target feature.(The last feature is the target feature)
    b. The labels in the target feature can be changed to **Leptodactylidae** or **Non-Leptodactylidae**
    c. Apply SVM classifier with linear kernel and 70:30 training-testing split to generate the model.
    d. Plot graph with varied values for penalty term C and show how the recall changes.

8. Load the following image **fl-img.png**
    a. Apply PCA on the image.
    b. Find the eigen values and corresponding eigen vectors for the first 5 principal components
    c. Visualize the reconstructed image with reduced principal components
    d. Plot the graph PC versus reconstruction error for varied number of principal components.