Week 2 - Introduction to Scipy and Pandas

DS3010 - Introduction to Machine Learning

Instructions

1. Provide commented, indented code. Variables should have meaningful names.
2. Submit one .ipynb file containing all answers. The name should be [student name][roll_number] assignment[number].ipynb
3. Read the questions carefully before answering. If a question asks to follow a particular approach or to use a specific data structure, then it must be followed.
4. Write questions in separate text blocks in Jupyter Notebook before the code blocks containing answers.
5. All plots should have appropriate axis labels, titles, and legends.

Task for the Lab

1. Scipy exercise:

1.Solve a linear algebra system which is given as [Marks : 1]

x+5y+10z+7w=10

2x+12y+7z+1w=18

10x+8y+3z+2w=20

5x+4y+7z+5w=30

a. Using numpy create input array, solution array, then Solve and print the results
b. Verify your Results.
c. Find the determinant, trace of the square matrix.
d. Find the eigenvalues and eigenvectors.
e. Verify if the matrix is symmetric.

2. Generate a set of eight random 2D points. [Marks : 1]

a. Find the smallest polygon that covers all of the given points

Hint: You can use the Convex Hull algorithm

b. Visualize the points and the convex hull using matplotlib.

c. Plot the edges.

d. Highlight the area covered by the convex hull by filling it with a semi-transparent color.

2. Pandas exercise:

3. Consider the following Python dictionary data and list index labels. Create a DataFrame birds: [Marks : 1]

```
data = {
      'birds': ['Cranes', 'Cranes', 'plovers', 'spoonbills', 'spoonbills', 'Cranes', 'plovers', 'Cranes', 'spoonbills', 'spoonbills', 'Cranes', 'Cranes', 'plovers', 'plovers', 'spoonbills', 'spoonbills', 'Cranes', 'plovers', 'spoonbills', 'Cranes', 'Cranes', 'plovers', 'spoonbills', 'Cranes'],
    'age': [3.5, 4, 1.5, np.nan, 6, 3, 5.5, np.nan, 8, 4, 3.5, 2.5, 4.5, 3, 2.0, 5.0, 6.0, 7.0, 2.0, 4.0, 3.0, 5.5, 4.5, 7],
   'visits': [2, 4, 3, 4, 3, 4, 2, 2, 3, 2, 2, 1, 3, 2, 4, 3, 2, 1, 5, 3, 4, 1, 2, 5],
    'priority': ['yes', 'yes', 'no', np.nan, 'no', 'no', 'no', 'yes', 'no', 'no', 'yes', 'no', 'yes', 'no', 'no', 'yes', 'no', 'yes', 'yes', 'no', 'no', 'yes', 'no', 'yes']
}

labels = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j', 'k', 'l', 'm', 'n', 'o', 'p', 'q', 'r', 's', 't', 'u', 'v', 'w', 'x']
```

a. Fill nan with respective series mode value.

b. Select the rows where the birds is a Cranes and the age is between 2 and 4 (inclusive)

c. Sort dataframe (birds) first by the values in the 'age' in descending order, then by the value in the 'visits' column in ascending order.

d. Drop duplicate rows and make these changes permanent. Show dataframe after changes.
e. Create a grouped bar chart showing the count of each priority type (yes or no) for each bird type. What does this reveal about the relationship between bird type and priority?

4. You are provided with a housing price dataset. Consider the column "median house value" as the dependent feature and the remaining columns as the independent features. Load the dataset as a pandas dataframe and do the following tasks: [Marks : 2]

• EDA:

a. Generate the descriptive statistics of the features in the dataset and check for any missing values. Fill the missing values (if any) using the appropriate technique and justify.

b. Visualize the distribution of the categorical data using a pie-chart.

• Data Preprocessing:

c. Deal with the categorical features. Convert categorical features into numerical values.

d. Separate features and target column. Then carry out feature scaling.

e. Define multicollinearity. Check for multicollinearity in the dataset.

Extra Question:

5. Use the football dataset https://www.kaggle.com/datasets/sayanroy729/fifa-worldcup-2022-results.
    a. Find out the total percentages that each team made on target. Display the result as a python dictionary where the keys are the team list and the values are the percentage values. Round off the percentage values up to 2 decimal places.
    b. Find out how many times the teams played in this Fifa World Cup-2022. On top of this, find out the ranks of the teams.
    c. Drop all the duplicate rows permanently.
    d. Drop the columns: "Sl No", "Match No.", "Red Cards" and "Pts" permanently.
    e. Find out the rank based on the "Team" column and save the result by adding a new column named "Rank".
    f. Change the datatype of this column to integer (by using np.int16).