Python for Computer Science and Data Science 2 (CSE 3652)

MINOR ASSIGNMENT-4: MACHINE LEARNING- CLASSIFICATION, REGRESSION AND CLUSTERING

1. Perform dimensionality reduction using scikit-learn's TSNE estimator on the Iris dataset, then graph the results.

2. Create a Seaborn pairplot graph for the California Housing dataset. Try the Matplotlib features to panning and zoom in on the diagram. These are accessible via the icons in the Matplotlib window.

3. Go to NOAA's Climate at a Glance page (Link) and download the available time series data for the average annual temperatures of New York City from 1895 to today (1895-2025). Implement simple linear regression using average annual temperature data. Also, show how does the temperature trend compare to the average January high temperatures?

4. Load the Iris dataset from the scikit-learn library and perform classification on it with the k-nearest neighbors algorithm. Use a KNeighborsClassifier with the default k value. What is the prediction accuracy?

5. You are given a dataset of 2D points with their corresponding class labels. The dataset is as follows:

Point ID	x	y	Class
A	2.0	3.0	0
B	1.0	1.0	0
C	4.0	4.0	1
D	5.0	2.0	1

   A new point P with coordinates $(3.0, 3.0)$ needs to be classified using the KNN algorithm. Use the Euclidean distance to calculate the distance between points.

6. A teacher wants to classify students as "Pass" or "Fail" based on their performance in three exams. The dataset includes three features:

Exam 1 Score	Exam 2 Score	Exam 3 Score	Class (Pass/Fail)
85	90	88	Pass
70	75	80	Pass
60	65	70	Fail
50	55	58	Fail
95	92	96	Pass
45	50	48	Fail

   A new student has the following scores:

   - Exam 1 Score: 72
   - Exam 2 Score: 78
   - Exam 3 Score: 75

   Classify this student using the K-Nearest Neighbors (KNN) algorithm with $k = 3$.

7. Using scikit-learn's KFold class and the cross_val_score function, determine the optimal value for k to classify the Iris dataset using a KNeighborsClassifier.

8. Write a Python script to perform K-Means clustering on the following dataset:

   $$\text{Dataset: } \{(1,1),(2,2),(3,3),(8,8),(9,9),(10,10)\}$$

   Use k=2 and visualize the clusters.

9. Write a Python script to perform K-Means clustering on the following dataset: Mall Customer Segmentation. Use $k = 5$ (also, determine optimal k via the Elbow Method) and visualize the clusters to identify customer segments.

   Expected Output:

   - Scatter plot showing clusters (e.g., "High Income-Low Spenders," "Moderate Income-Moderate Spenders").
   - Insights for targeted marketing strategies.

10. Perform the following tasks using the pandas `Series` object:

    (a) Create a Series from the list `[7, 11, 13, 17]`.

    (b) Create a Series with five elements where each element is `100.0`.

    (c) Create a Series with 20 elements that are all random numbers in the range 0 to 100. Use the `describe` method to produce the Series' basic descriptive statistics.

    (d) Create a Series called `temperatures` with the following floating-point values: `98.6`, `98.9`, `100.2`, and `97.9`. Use the `index` keyword argument to specify the custom indices 'Julie', 'Charlie', 'Sam', and 'Andrea'.

    (e) Form a dictionary from the names and values in Part (d), then use it to initialize a Series.