Assignment Questions

Q1. Explain how the K-Nearest Neighbors (KNN) algorithm works. What are the key hyperparameters and distance metrics used.

Q2.

You are given the following data points:

Point	X	Y	Label
A	1	2	0
B	2	3	0
C	3	3	1
D	6	5	1

Predict the label of a new point **P(3,2)** using **K=3** and **Euclidean distance**.

Q3. Given data points: (1,1), (1.5,2), (3,4), (5,7), (3.5,5), (4.5,5), (3.5,4.5). Perform two iteration of K-Means with **K=2** using **initial centroids** as:

- Centroid 1: (1,1)
- Centroid 2: (5,7)

Case Study Questions

Supervised Learning

1. A bank wants to predict whether a loan applicant will default or not based on features such as income, loan amount, and credit score.
    - (a) Identify whether this is a classification or regression problem.
    - (b) Suggest a suitable supervised learning algorithm.
    - (c) Explain how you would split the data into training, validation, and test sets.

Unsupervised Learning – Clustering

2. A retail store has customer purchase data without labels. The store wants to segment customers into different groups for targeted marketing.
    - (a) Which machine learning approach is suitable here?
    - (b) Explain how K-Means clustering can be applied to this problem.
    - (c) Suggest how the results can help the store increase sales.

Overfitting and Underfitting

3. A student builds a decision tree to classify emails as "Spam" or "Not Spam." The model performs **98% accuracy** on the training data but only **60% accuracy** on test data.
    - (a) What problem is the student's model facing?
    - (b) Suggest two techniques to fix this problem.
    - (c) If the opposite occurred (low training and test accuracy), what issue would it indicate?

Hyperparameters & Validation

4. You are training a neural network for image classification (cats vs dogs).
    - (a) Explain the role of **learning rate**, **batch size**, and **epoch** in training this model.
    - (b) Why do we need a **validation set** during training?
    - (c) Suppose the validation accuracy decreases while training accuracy increases — what should you do?

K-Means Algorithm

5. Consider the dataset:
(185,72),(170,56),(168,60),(179,68),(182,72),(188,77)(185, 72), (170, 56), (168, 60), (179, 68), (182, 72), (188, 77)
    - (a) Apply **K-Means clustering** with K=2K = 2, choosing the first two points as initial centroids.
    - (b) Show two iterations with centroid updates and final cluster assignment.
    - (c) Explain how clustering can be useful in real-world applications like marketing.

Decision Trees

6. A company wants to predict whether a customer will buy a product or not. The dataset has features:
    - Age (Young, Middle-aged, Senior)
    - Income (High, Medium, Low)
    - Student (Yes, No)
    - Credit Rating (Fair, Excellent)
    - Target: Buys Product (Yes, No)
    - (a) Draw a **decision tree** using the above attributes.
    - (b) Explain how **Information Gain** or **Gini Index** is used to decide the root node.
    - (c) What are pruning and overfitting in decision trees?