Training Notes: K-Nearest Neighbors (K-NN) Algorithm

1. Introduction to K-Nearest Neighbors (K-NN)

- **Definition**: K-NN is a **supervised machine learning algorithm** used for **classification** and **regression**.
- **Key Idea**: It classifies a new data point based on the majority class (for classification) or average value (for regression) of its **'K' nearest neighbors** in the feature space.
- **Instance-Based Learning**: It is a **lazy learner** (does not learn a model during training; computation happens at prediction time).

---

2. How K-NN Works

Step-by-Step Process

1. **Choose the number of neighbors (K)**: Typically an odd number to avoid ties.
2. **Calculate Distance**:
   - Common distance metrics:
     - **Euclidean Distance** (most common)
     - **Manhattan Distance**
     - **Minkowski Distance** (generalized form)
3. **Find K-Nearest Neighbors**: Identify the K closest data points.
4. **Majority Voting (Classification)**: Assign the class that appears most among neighbors.
5. **Average (Regression)**: Predict the average value of the neighbors.

Example (Classification)

- **New data point**: Find its 3 nearest neighbors.
- **Neighbors' classes**: [Class A, Class A, Class B]

- **Prediction**: Class A (majority vote).

---

3. Choosing the Right K Value

- **Small K (e.g., K=1)**:
  - High variance (overfitting, sensitive to noise).
- **Large K (e.g., K=20)**:
  - High bias (underfitting, smoother boundaries).
- **Optimal K**: Found using **cross-validation**.