# Assignment- KNN & PCA

1. **What is K-Nearest Neighbors (KNN) and how does it work?**
   KNN is a non-parametric, supervised learning algorithm used for classification and regression. It works by finding the "K" closest data points (neighbors) to a given query point and making predictions based on their values. It relies on distance metrics such as Euclidean distance.

2. **What is the difference between KNN Classification and KNN Regression?**

   a. **KNN Classification:** The majority class among the K nearest neighbors determines the class label of the new data point.
   b. **KNN Regression:** The predicted value is the average (or weighted average) of the values of the K nearest neighbors.

3. **What is the role of the distance metric in KNN?**

   a. The distance metric determines how "closeness" is measured. Common metrics include:

      i. **Euclidean distance** (default, for continuous data)
      ii. **Manhattan distance** (better for grid-based data)
      iii. **Minkowski distance** (generalized version
      iv. **Cosine similarity** (for text and high-dimensional data)

4. **What is the Curse of Dimensionality in KNN?**

   a. As the number of dimensions increases, data points become more sparse, making distance calculations less meaningful. This degrades KNN's performance.

5. **How can we choose the best value of K in KNN?**

   a. Use cross-validation:
      i. A small K may lead to overfitting.
      ii. A large K may oversmooth the decision boundary (underfitting).
      iii. Typically, odd K values are chosen to avoid ties.

6. **What are KD Tree and Ball Tree in KNN?**
    a. These are data structures used to speed up nearest neighbor searches:
        i. **KD Tree (K-Dimensional Tree):** Works well for low-dimensional data.
        ii. **Ball Tree:** More efficient for high-dimensional data.
7. **When should you use KD Tree vs. Ball Tree?**
    a. **Use KD Tree** when dimensions are low (below ~30).
    b. **Use Ball Tree** when dimensions are higher.
8. **What are the disadvantages of KNN?**
    a. Computationally expensive (O(n) for each query)
    b. Memory-intensive (stores entire dataset)
    c. Sensitive to irrelevant and correlated features
    d. Affected by imbalanced datasets
9. **How does feature scaling affect KNN?**
    a. Since KNN relies on distance calculations, features with large scales dominate. Techniques like **Min-Max Scaling** or **Standardization (Z-score normalization)** improve performance.

10. **What is PCA (Principal Component Analysis)?**
    a. PCA is a dimensionality reduction technique that transforms data into a new coordinate system where the most variance is captured in fewer dimensions.
11. **How does PCA work?**
    a. Steps:
        i. Standardize the dataset.
        ii. Compute the covariance matrix.
        iii. Compute eigenvalues and eigenvectors.
        iv. Select the top eigenvectors (principal components).
        v. Transform the data into the new subspace.

12. **What is the geometric intuition behind PCA?**

    a. PCA finds new axes (principal components) along which the variance is maximized. These axes are perpendicular (orthogonal) to each other.

13. **What are Eigenvalues and Eigenvectors in PCA?**
    a. **Eigenvectors** define the new feature space (directions of maximum variance).
    b. **Eigenvalues** indicate the amount of variance explained by each eigenvector.

14. **What is the difference between Feature Selection and Feature Extraction?**
    a. **Feature Selection:** Selecting a subset of existing features.
    b. **Feature Extraction:** Creating new features from existing ones (e.g., PCA).

15. **How do you decide the number of components to keep in PCA?**
    a. Use the **explained variance ratio**:
        i. Keep components that explain ~95% of the variance.

16. **Can PCA be used for classification?**
    a. PCA itself is not a classifier, but it helps in preprocessing for classification by reducing dimensionality.

17. **What are the limitations of PCA?**
    a. Assumes linear relationships.
    b. Sensitive to scale (requires standardization).
    c. May lose interpretability.

18. **How do KNN and PCA complement each other?**
    a. PCA reduces dimensionality and mitigates the curse of dimensionality, improving KNN's efficiency.

19. **How does KNN handle missing values in a dataset?**

    a. Common approaches:
        i. **Imputation** (mean, median, KNN imputation)
        ii. **Ignoring missing values** (if a small percentage is missing)

20. **What are the key differences between PCA and Linear Discriminant Analysis (LDA)?**
    a. **PCA:** Unsupervised, maximizes variance.
    b. **LDA:** Supervised, maximizes class separability.