Interview Question

1. What are the key hyperparameters in KNN?

Ans:

Some key hyperparameters in KNN are:

1. **Number of Neighbors (K)**:
   * K represents the number of nearest neighbors to consider when making predictions. It is a crucial hyperparameter that significantly influences the model's performance. Choosing an appropriate value for K is essential to balance bias and variance in the model. A small K value can lead to high variance (overfitting), while a large K value can lead to high bias (underfitting).
2. **Distance Metric**:
   * KNN uses a distance metric to measure the distance between data points. The most common distance metrics are:
     + Euclidean distance: ∑𝑖=1𝑛(𝑥𝑖−𝑦𝑖)2∑*i*=1*n*​(*xi*​−*yi*​)2​
     + Manhattan distance: ∑𝑖=1𝑛∣𝑥𝑖−𝑦𝑖∣∑*i*=1*n*​∣*xi*​−*yi*​∣
     + Minkowski distance: A generalization of Euclidean and Manhattan distances, controlled by a parameter 𝑝*p*.
3. **Weighting Scheme**:
   * KNN can use different weighting schemes to give more or less importance to neighboring points when making predictions. The two main weighting schemes are:
     + Uniform weighting: All neighbors contribute equally to the prediction.
     + Distance weighting: Weight neighbors inversely proportional to their distance from the query point. Closer neighbors have a greater influence on the prediction.
4. **Algorithm**:
   * Although the basic algorithm of KNN is straightforward, there are variations that can be more efficient for large datasets. The two main algorithms are:
     + Brute-force: Computes distances between the query point and all training samples.
     + KD-tree or Ball-tree: Data structures used to organize the training samples, allowing for faster nearest neighbor search.
5. **Feature Scaling**:
   * Since KNN relies on distance metrics, the scale of features can significantly affect the algorithm's performance. It's often necessary to scale or normalize features to ensure that no single feature dominates the distance calculations.
6. **Leaf Size**:
   * This hyperparameter is relevant when using KD-tree or Ball-tree algorithms. It determines the number of samples at the leaf node of the tree. Smaller leaf sizes may result in a more exhaustive search but can also increase computational overhead.
7. What distance metrics can be used in KNN?

Ans:

1. Euclidean Distance
2. Manhattan Distance
3. Minkowski Distance
4. Chebyshev Distance
5. Hamming Distance
6. Cosine Similarity