**KNN - Interview Questions and Answers**

1. **What are the key hyperparameters in KNN?**

The key hyperparameters in the K-Nearest Neighbors (KNN) algorithm are:

**Number of Neighbors (k):**

This is the most important hyperparameter. It determines the number of nearest neighbors to consider when making a prediction. A smaller value of k can lead to a model that captures noise in the training data (overfitting), while a larger value of k might lead to a model that oversimplifies the patterns in the data (underfitting). Choosing the right value for k is crucial for the performance of the KNN algorithm.

**Distance Metric:**

This hyperparameter defines how the distance between data points is calculated. Common distance metrics include:

**Euclidean Distance:** The most common distance metric, suitable for continuous features.

**Manhattan Distance:** Useful when the data is high-dimensional or when the features are not continuous.

**Minkowski Distance:** A generalized form that includes both Euclidean and Manhattan distances as special cases.

**Hamming Distance:** Used for categorical variables.

**Weighting Function:**

This hyperparameter determines how the influence of each neighbor is weighted when making a prediction.

**Options include:**

**Uniform Weighting:** All neighbors contribute equally to the prediction.

**Distance Weighting:** Closer neighbors have a higher weight, meaning their contribution is more significant than that of farther neighbors. This can often improve performance in cases where closer neighbors are more likely to have similar outcomes.

**Algorithm for Computing Nearest Neighbors:**

This hyperparameter controls the method used to compute the nearest neighbors efficiently.

**Common algorithms include:**

**Brute Force:** Computes the distance between the query point and all other points in the training set. This is accurate but computationally expensive for large datasets.

**KD-Tree:** Efficient for low-dimensional data, it partitions the space to reduce the number of points considered when searching for neighbors.

**Ball Tree:** Suitable for high-dimensional data and works by constructing a tree structure that helps in faster neighbor searches.

**Leaf Size (for KD-Tree and Ball Tree algorithms):**

This hyperparameter affects the speed of the construction and query processes in tree-based algorithms. A smaller leaf size can lead to more accurate results but increases computational cost. Choosing the right combination of these hyperparameters is essential for optimizing the performance of the KNN algorithm for a specific dataset. This typically involves experimentation and cross-validation.

1. **What distance metrics can be used in KNN?**

**Distance Metric:**

This hyperparameter defines how the distance between data points is calculated.

**Common distance metrics include:**

**Euclidean Distance:** The most common distance metric, suitable for continuous features. It calculates the straight-line distance between two points in Euclidean space.

**Manhattan Distance (L1 Norm or Taxicab Distance):** Useful when the data is high-dimensional or when the features are not continuous. Computes the distance between two points by summing the absolute differences of their coordinates.

**Minkowski Distance:** A generalized form that includes both Euclidean and Manhattan distances as special cases. It has a parameter 𝑝

**p that determines the type of distance:**

If 𝑝 = 1

p=1, Minkowski distance is equivalent to Manhattan distance.

If 𝑝 = 2

p=2, Minkowski distance is equivalent to Euclidean distance.

**Hamming Distance:** Used for categorical variables. Counts the number of positions at which the corresponding elements are different.

Suitable for cases where features are not continuous and can be compared directly.

**Chebyshev Distance:** Measures the maximum absolute difference between the coordinates of two points. Suitable for scenarios where the maximum difference in any dimension is of interest.

**Cosine Similarity (converted to a distance measure):** Measures the cosine of the angle between two vectors in a multi-dimensional space. Often used for text classification and high-dimensional spaces where the magnitude of the vectors is not as important as their direction.

Likewise, there are many distance metrics can be used in KNN. In K-Nearest Neighbors (KNN), several distance metrics can be used to measure the similarity or dissimilarity between data points. The choice of distance metric can significantly impact the performance of the KNN algorithm, especially depending on the nature of the data.