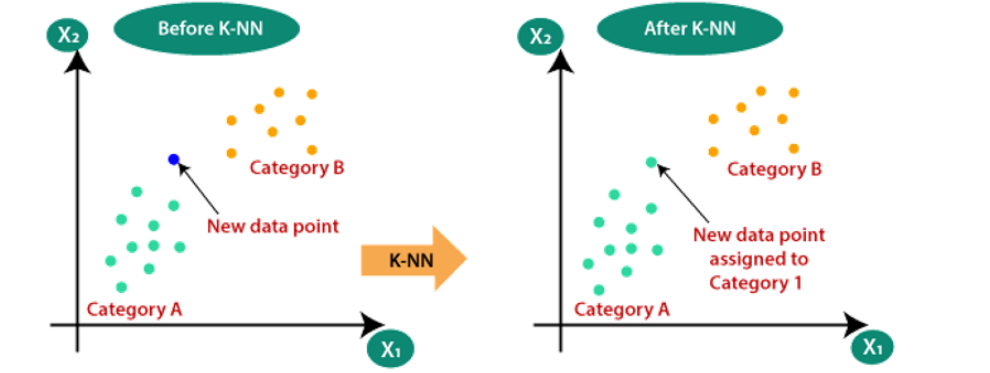
# K-Nearest Neighbors (KNN)

**What is the K-Nearest Neighbors Algorithm?**  
Also known as K-Nearest Neighbors (KNN), this algorithm is a popular method used in machine learning to solve classification and regression problems. It determines the class of a data point by considering the classes of its closest neighbors.



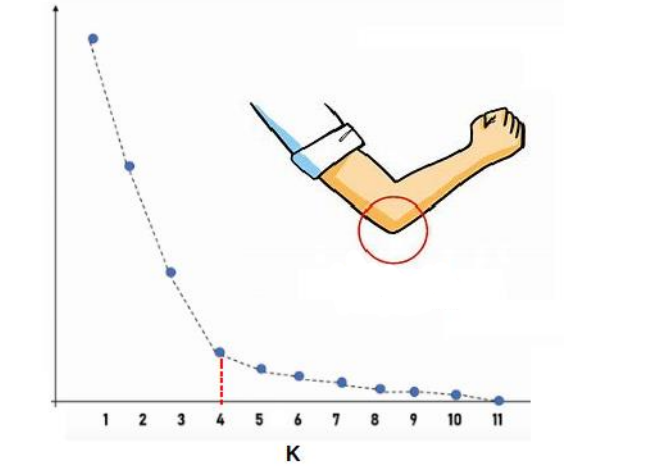
**Working Principle of the KNN Algorithm**

**1- Selecting the Number of Neighbors (k value)**  
With a small k value, the model becomes more sensitive to the data, increasing the risk of overfitting, which may result in inaccurate predictions. Conversely, with a large k value, the model generalizes better but might overlook smaller classes.

Typically, the k value should be chosen as an odd number to ensure that ties between the nearest classes are avoided.  
Methods like the Elbow Method can be used to experiment with different k values, evaluate the model's performance, and select the most suitable k value.

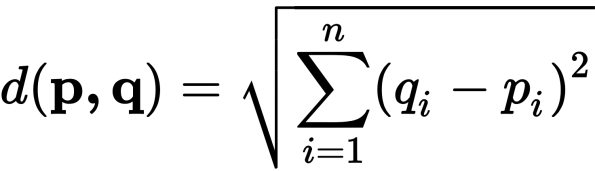
**Elbow Method**  
The Elbow Method is a technique used to select the K value in clustering or classification algorithms like KNN. This method evaluates the model's performance by experimenting with different K values and aims to determine the optimal K value.

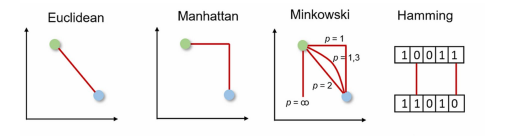
The Elbow Method is based on the principle that as the K value increases, the rate of improvement in the model's performance slows down, resulting in a curve that bends like an elbow. This "elbow point" represents the K value that provides the best performance for the model.



**2- Distance Calculation**  
KNN calculates the distance between a test sample and each instance in the training dataset. The neighbors with the smallest k distances are selected. These neighbors are the training samples that have the most similar features to the test sample.The most common distance metrics include Euclidean, Manhattan, Minkowski, and Hamming distances. Among these, the Euclidean distance is the most widely used; however, the choice of distance metric can vary depending on the structure of the data.

**Euclidean Distance Formula:**



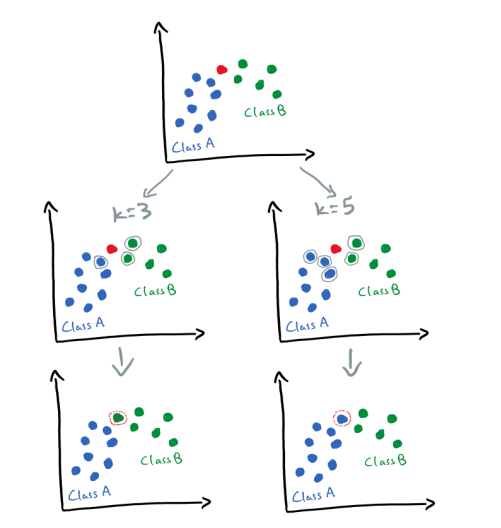


The appropriate distance metric is calculated, and then the k nearest neighbors are selected based on the calculated distances.

1. **Class Prediction**

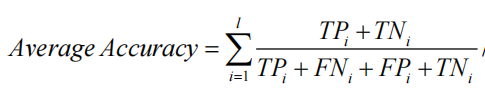
The classes of the selected neighbors are examined. Typically, the most frequent class among the neighbors is assigned as the predicted class for the test sample.

The model training is quick, but during the prediction phase, since all the data needs to be used, it can be time-consuming.



Performance Calculation  
The observations in the test dataset are compared with the observations in the training dataset, and a category is assigned to them. Since we know the class of the observations in the test dataset, the model's performance can be evaluated.

The most commonly used method for performance evaluation is the average accuracy.



**TP (True Positive):** Correctly classified as positive.  
**TN (True Negative):** Correctly classified as negative.  
**FP (False Positive):** Incorrectly predicted as positive.  
**FN (False Negative):** Incorrectly predicted as negative.

## Advantages of the KNN Algorithm

* **Simplicity and Accuracy:** It is easy to understand and apply. Users can quickly grasp the basic logic of the algorithm.
* **Ease of Parameter Tuning:** The model does not learn any parameters during the training phase; all data is stored. This enables the model to be set up quickly.
* **Instance-based Learning:** When new data is added, the model can be easily updated. When the training data changes, predictions can be made with the new data without retraining the model.

## Disadvantages of the KNN Algorithm

* **Curse of Dimensionality:** Storing all the training data can be memory-intensive. This is a significant disadvantage for large datasets.
* **Sensitivity to Data:** Noisy data and imbalanced class distributions can negatively impact the model's performance. In particular, predicting minority classes can be difficult.
* **High Computational Cost:** For large datasets, calculating the distance with all training data for each prediction can be time-consuming. This issue becomes more pronounced with high-dimensional data.

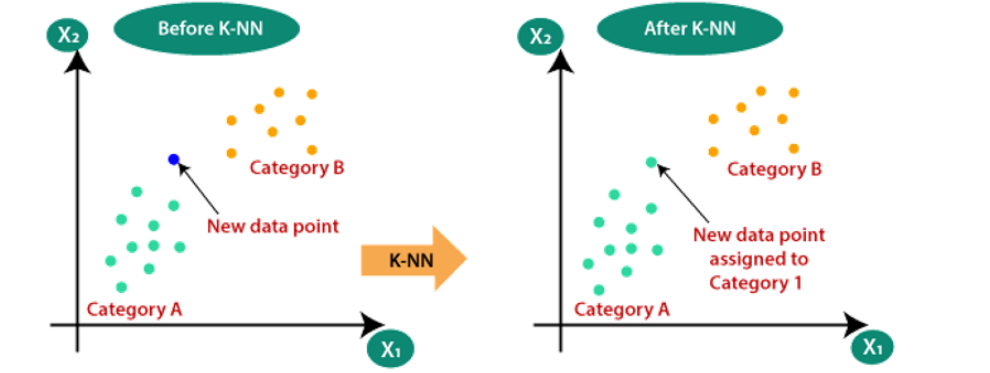
## Kaynakça

* Zhongheng Zhang, (2016), Department of Critical Care Medicine, Jinhua Municipal Central Hospital, Jinhua Hospital of Zhejiang University, Jinhua 321000, China.
* BTK Academy, Machine Learning Course, *[www.btkakademi.gov.tr/portal/course/makine-ogrenmesi-30123](http://www.btkakademi.gov.tr/portal/course/makine-ogrenmesi-30123)*

# EN YAKIN KOMŞU (KNN)

## K-En Yakın Komşu Algoritması Nedir?

Bir diğer adıyla K-Nearest Neighbors olan bu algoritma, makine öğrenimi alanında sınıflandırma ve regresyon problemleri çözmek için kullanılan popüler bir algoritmadır. Bu algoritma, bir veri noktasının sınıfını belirlemek için, o noktaya en yakın komşularının sınıflarını dikkate alır.



## KNN Algoritması Çalışma Prensibi

#### Komşu sayısı seçilmesi (k değeri)

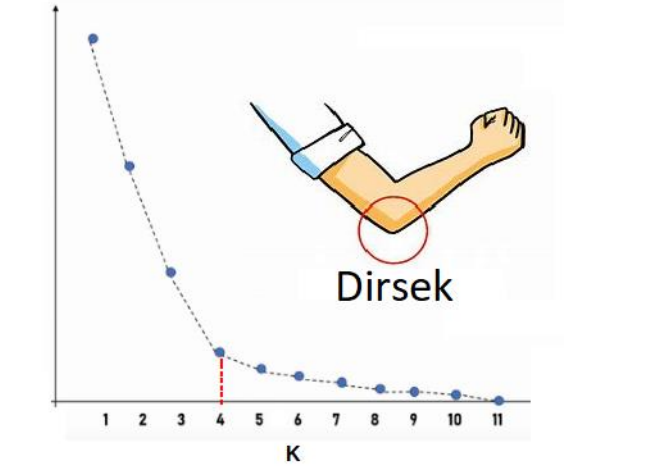
Küçük bir k değeri ile model veriye daha duyarlı hale gelir ve aşırı uyum (overfitting) riski artar yani doğru tahminlerde bulunamaz. Büyük bir k değerinde ise model genelleştirme yapar ancak küçük sınıfları göz ardı edebilir.

Genellikle doğru bir tahmin yapılabilmesi için k değeri tek sayı olarak seçilmelidir. Bu sayede eşit en yakın sınıflara sahip komşıuların eşit olma durumunun önüne geçilir.

Elbow gibi yöntemler kullanılarak farklı k değerlerini deneyerek modelin performansını değerlendirir ve en uygun k değerini seçer.

**Elbow Metodu:**

Bu metot KNN gibi kümeleme veya sınıflandırma algoritmalarında K değerini seçmek için kullanılan bir tekniktir. Bu yöntem, farklı K değerlerini deneyerek modelin performansını değerlendirir ve optimum K değerini belirlemeye çalışır. Elbow metodu, K değerinin artmasıyla birlikte modelin performansındaki azalışın hızının azalması ve eğrinin dirsek gibi bükülmesi prensibine dayanır.Bu dirsek noktası, K değerinin en iyi performansı sağladığı noktayı temsil eder

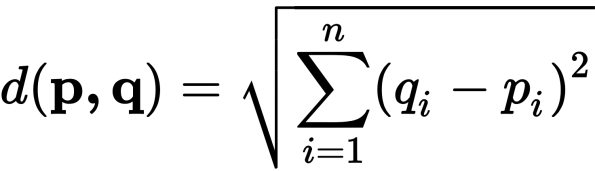


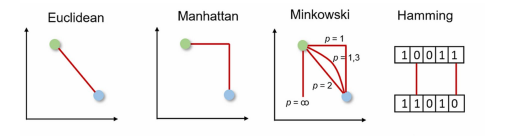
#### Mesafe hesaplama

KNN, bir test örneği ile eğitim veri kümesindeki her bir örnek arasındaki mesafeyi hesaplar. Hesaplanan mesafelerden en küçük k değere sahip olan komşular seçilir. Bu komşular, test örneği ile en benzer özelliklere sahip olan eğitim örnekleridir.

En yaygın mesafe ölçütleri olarak; Euclidean (öklid), Manhattan, minkowski ve hamming mesafeleridir. Bunlardan en yaygın kullanılanı öklid olurken verilerin yapısına göre mesafe hesaplama yöntemlerinde farklılık gösterilebilir.

Öklid yöntemi formülü:



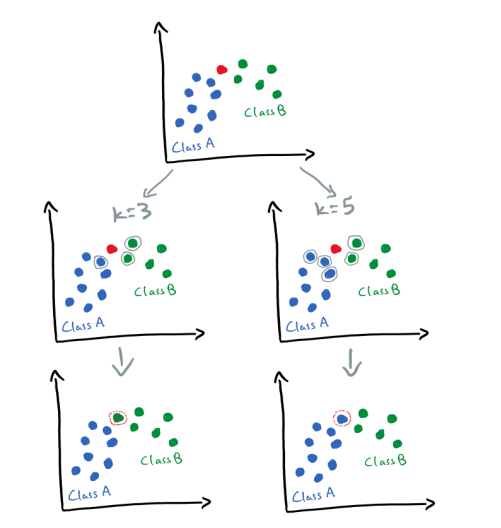


Bu mesafelerden uygun olanı hesaplanıp seçilen k komşu sayısı kadar komşu seçilir.

#### Sınıf tahmini

Seçilen komşuların sınıfları incelenir. Genellikle, en sık görülen sınıf, test örneği için tahmin edilen sınıf olarak atanır.

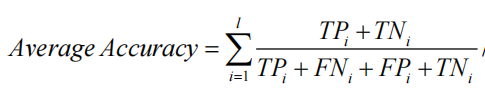
Modelin eğitilmesi hızlıdır, ancak tahmin aşamasında tüm verilerin kullanılması gerektiği için zaman alıcı olabilir.



## Performans hesaplama

Test veri kümesindeki gözlemleri eğitim veri kümesindeki gözlemlerle karşılaştırarak bunlara bir kategori atar. Test veri kümesindeki gözlemlerin sınıfını bildiğimiz için modelin performansı değerlendirilebilir.

Performans değerlendirmede kullanılan en yaygın yöntem ortalama doğruluk olarak kullanılabilir.



 **TP (True Positive)**: Doğru şekilde pozitif olarak sınıflandırılanlar.

 **TN (True Negative)**: Doğru şekilde negatif olarak sınıflandırılanlar.

 **FP (False Positive)**: Yanlış pozitif tahmin edilenler.

 **FN (False Negative)**: Yanlış negatif tahmin edilenler.

## KNN Algoritmasının Avantajları

* Basit ve doğruluk: Anlaşılması ve uygulanması kolaydır. Kullanıcılar, algoritmanın temel mantığını hızlıca kavrayabilir.
* Parametre ayarının kolaylığı: Model, eğitim aşamasında herhangi bir parametre öğrenmez; tüm veriler saklanır. Bu, modelin hızlı bir şekilde kurulmasını sağlar.
* Örnek bazlı öğrenme: Yeni veriler eklendiğinde modelin güncellenmesi kolaydır. Eğitim verisi değiştiğinde, modelin yeniden eğitilmesine gerek kalmadan yeni verilerle tahmin yapılabilir.

## KNN Algoritmasının Dezavantajları

* Boyutluluk problemi: Tüm eğitim verilerinin saklanması, bellek açısından maliyetli olabilir. Bu, büyük veri setleri için önemli bir dezavantajdır.
* Veriye duyarlılık:Gürültülü veriler ve dengesiz sınıf dağılımları, modelin performansını olumsuz etkileyebilir. Özellikle, azınlık sınıflarının doğru tahmin edilmesi zorlaşabilir.
* Yüksek hesaplama maliyeti: Büyük veri setlerinde, her tahmin için tüm eğitim verileri ile mesafe hesaplamak zaman alıcı olabilir. Bu, özellikle yüksek boyutlu verilerde daha belirgin hale gelir.

## Kaynakça

* Zhongheng Zhang, (2016), Department of Critical Care Medicine, Jinhua Municipal Central Hospital, Jinhua Hospital of Zhejiang University, Jinhua 321000, China.
* BTK Akademi Makine öğrenmesi kursu, *[www.btkakademi.gov.tr/portal/course/makine-ogrenmesi-30123](http://www.btkakademi.gov.tr/portal/course/makine-ogrenmesi-30123)*