

K-Nearest Neighbor(KNN) Algorithm

- **K-Nearest Neighbor** is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- **K-NN algorithm** assumes the **similarity between the new case/new data point and available cases/data points** and **put the new case into the category that is most similar to the available category's data points.**

KNN Algorithm

- K-NN algorithm stores all the available data and classifies a new data point based on the similarity and the similarity between the two data points is calculated using the Euclidean distance .
- This means when new data appears then it can be easily classified into a well suite category by using K-NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

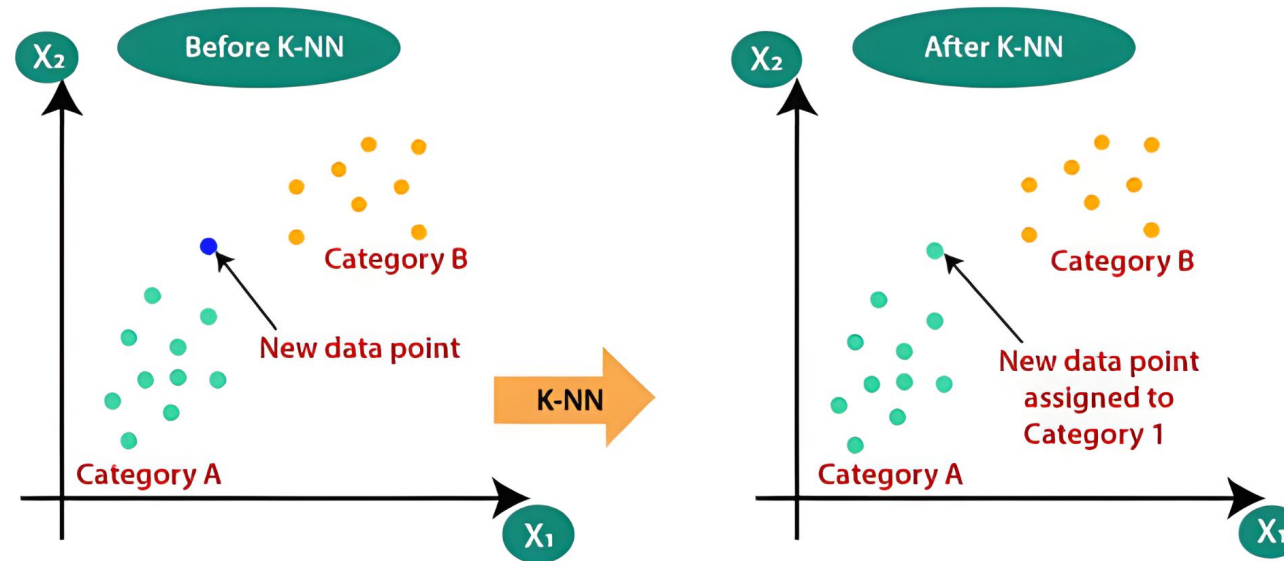
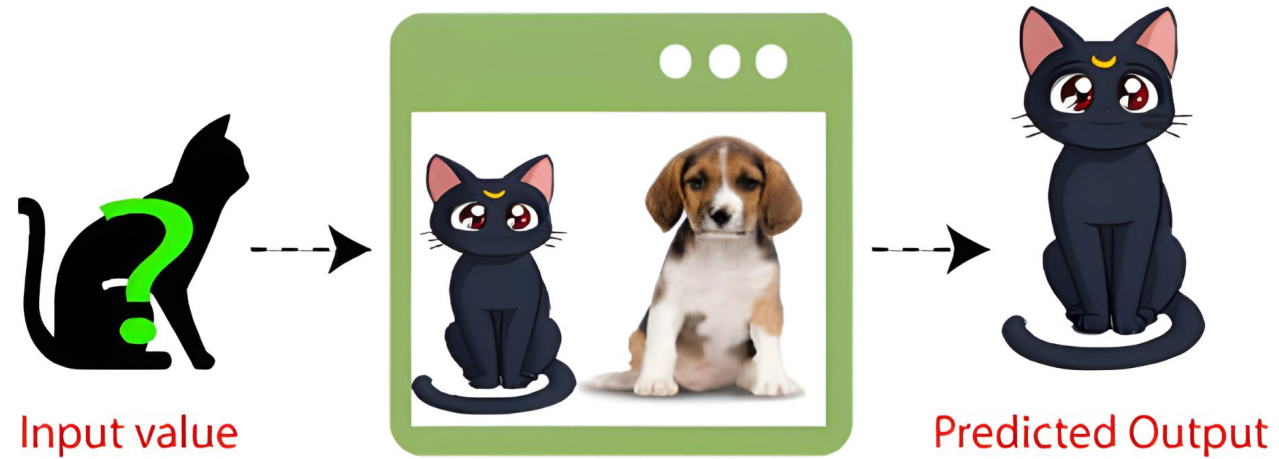
KNN Algorithm

- **K-NN** is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
- It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- **KNN algorithm** at the training phase just stores the dataset and when it gets **new data**, then it classifies that data into a category that is much similar to the new data.

KNN Algorithm

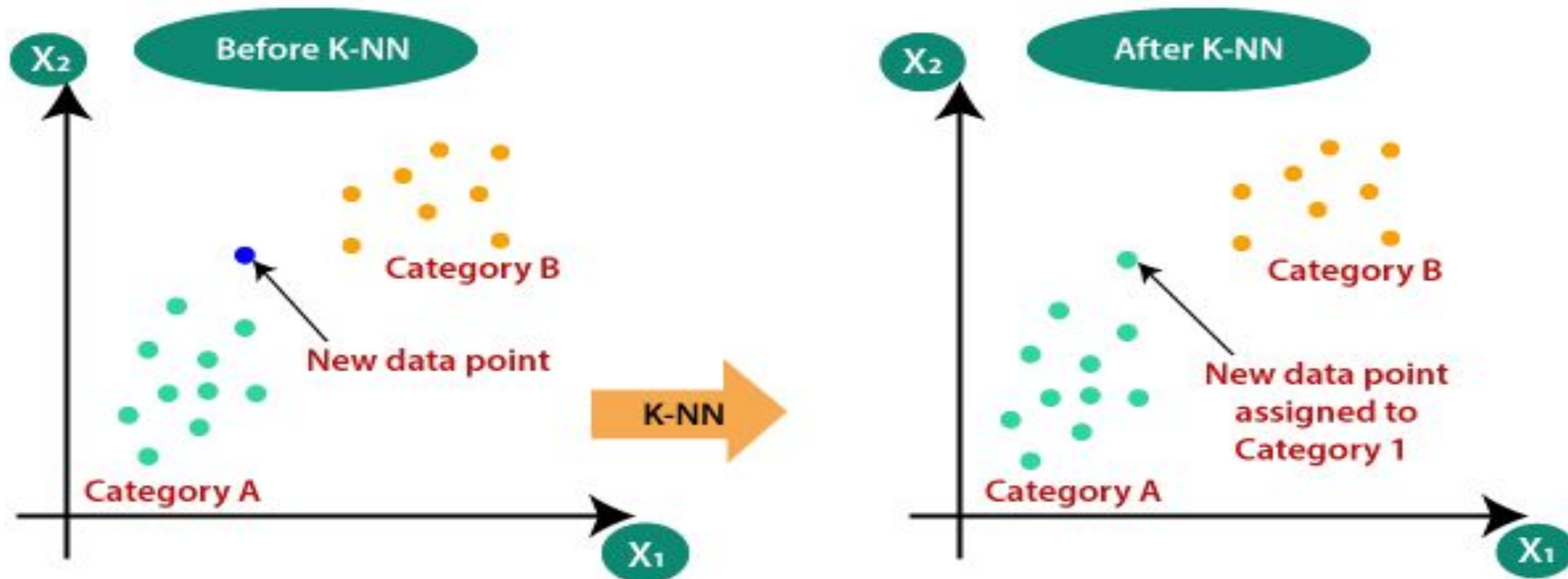
- **Example:** Suppose, we have an image of **an animal** that looks similar to cat and dog, but we want to know either it is **a cat or dog**.
- So for this identification, we can use the KNN algorithm, as it works on a **similarity measure**.
- Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features it will put it in either **cat or dog** category.

KNN Classifier



Why do we need a K-NN Algorithm?

Suppose there are two categories, i.e., **Category A** and **Category B**, and we have **a new data point x_1** , so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset. Consider the below diagram:



How does K-NN work?

The K-NN working can be explained on the basis of the below algorithm:

Step-1: Select the number K of the neighbors

Step-2: Calculate the Euclidean distance of K number of neighbors from new data point.

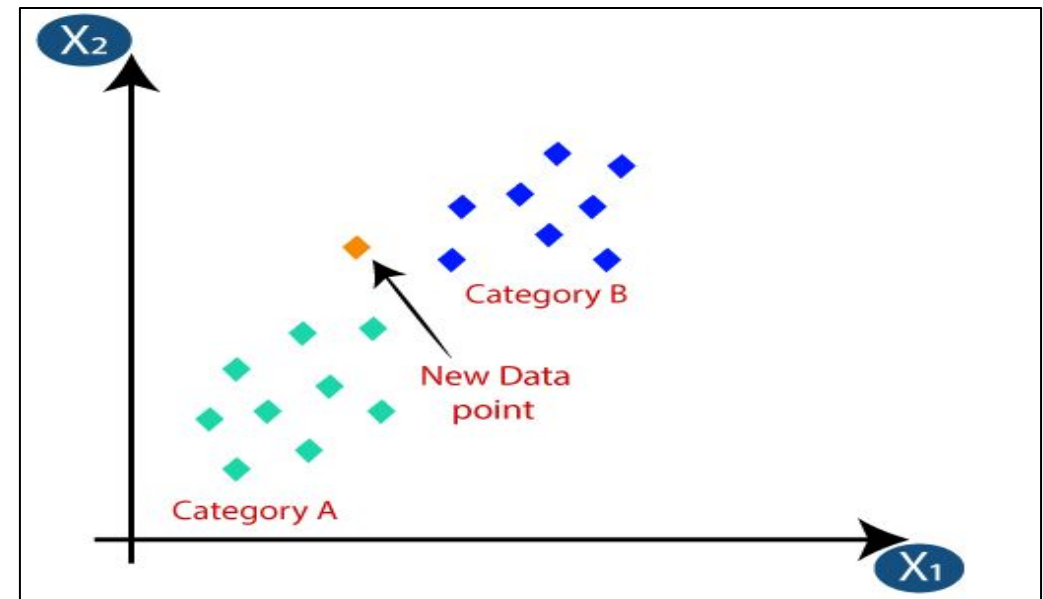
Step-3: Take the K nearest neighbors as per the calculated Euclidean distances.

Step-4: Among these k neighbors, count the number of the data points in each category.

Step-5: Assign the new data points to that category for which the number of the neighbors is maximum.

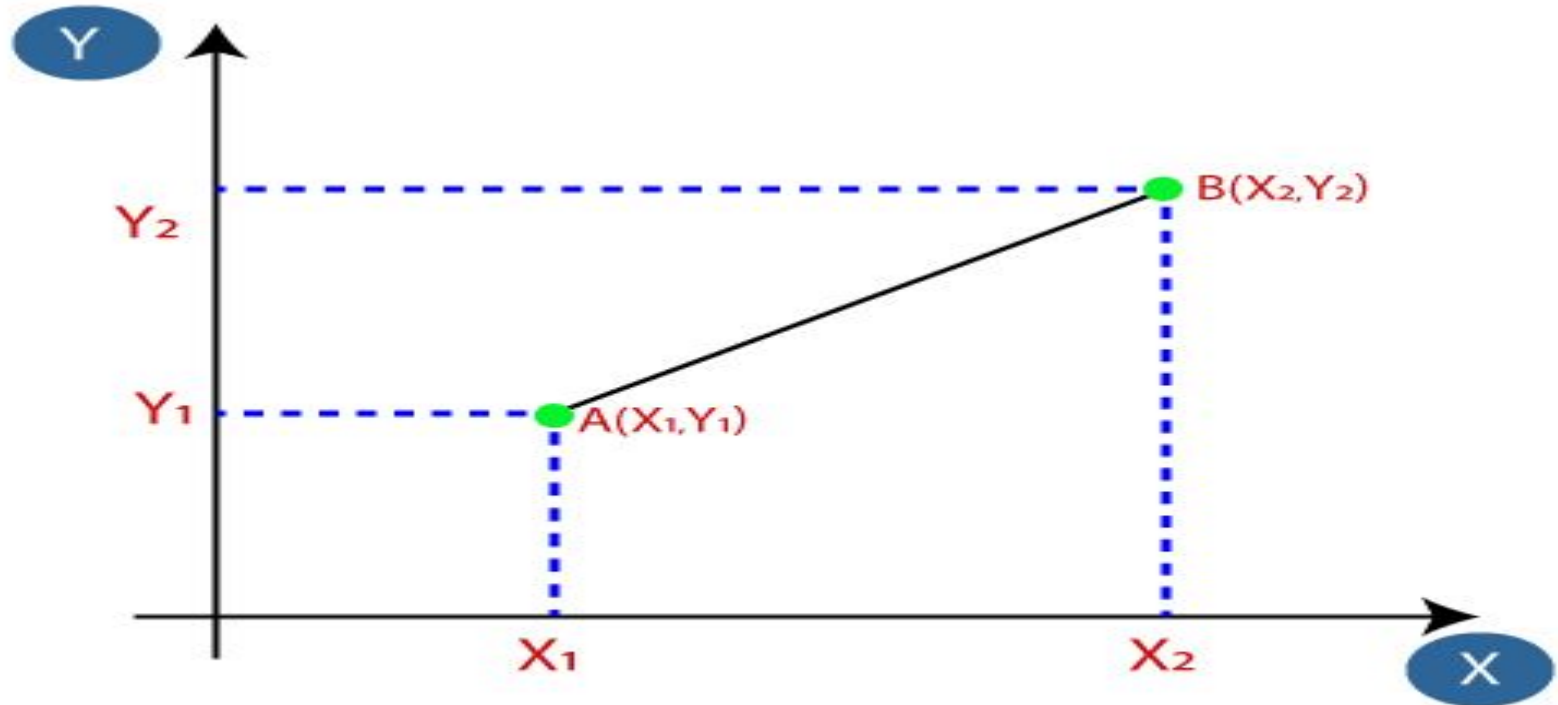
Step-6: Our model is ready.

Suppose we have a new data point and we need to put it in the required category.
Consider the image:



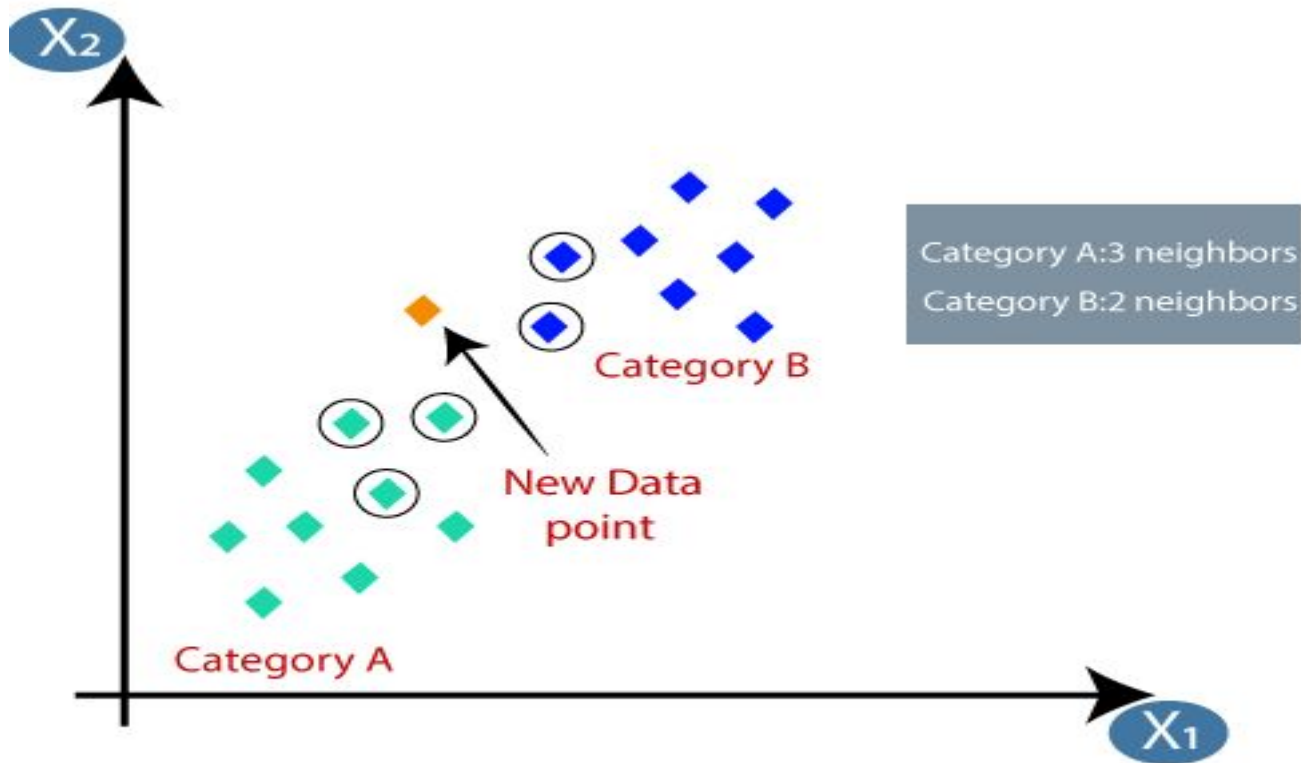
KNN Algorithm

- Firstly, we will choose the number of neighbors, so we will choose the $k=5$.
- Next, we will calculate the Euclidean distance between the data points from new data point.
- The Euclidean distance is the distance between two data points, which we have already studied in geometry. It can be calculated as:



$$\text{Euclidean Distance between } A_1 \text{ and } B_2 = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

- By calculating the Euclidean distance we got the nearest neighbors, as **three nearest neighbors in category A** and two nearest neighbors in category B. Consider the below image:



- As we can see the 3 nearest neighbors are from category A, hence this new data point must belong to category A.

KNN Algorithm

- How to select the value of K in the K-NN Algorithm?
- Below are some points to remember while selecting the value of K in the K-NN algorithm:
 - ✓ There is no particular way to determine the best value for "K", so we need to try some values to find the best out of them (with min. error).
 - ✓ The most preferred value for K is 5.
 - ✓ A very low value for K such as K=1 or K=2, can be noisy and lead to the effects of outliers in the model.
 - ✓ Large values for K are good, but it may find some difficulties.

KNN Algorithm

Advantages of KNN Algorithm:

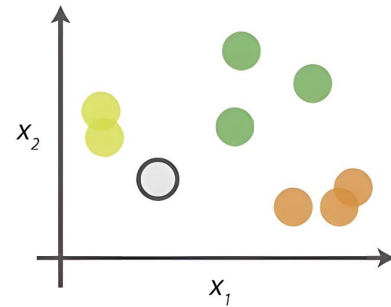
- It is simple to implement.
- It is robust to the noisy training data
- It can be more effective if the training data is large.

Disadvantages of KNN Algorithm:

- Always needs to determine the value of K which may be complex some time.
- The computation cost is high because of calculating the distance between the data points for all the training samples.

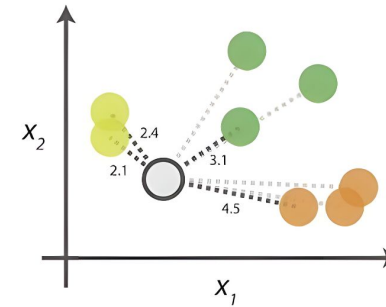
KNN Algorithm

0. Look at the data











Say you want to classify the grey point into a class. Here, there are three potential classes - lime green, green and orange.

1. Calculate distances









Start by calculating the distances between the grey point and all other points.

2. Find neighbours

Point Distance			
		2.1	→ 1st NN
		2.4	→ 2nd NN
		3.1	→ 3rd NN
		4.5	→ 4th NN

Next, find the nearest neighbours by ranking points by increasing distance. The nearest neighbours (NNs) of the grey point are the ones closest in dataspace.

3. Vote on labels

Class	# of votes	
	2	➔ Class  wins the vote! Point  is therefore predicted to be of class  .
	1	
	1	

Vote on the predicted class labels based on the classes of the k nearest neighbours. Here, the labels were predicted based on the $k=3$ nearest neighbours.

K Nearest Neighbors is a classification algorithm that operates on a very simple principle.

Training Algorithm:

1. Store all the Data

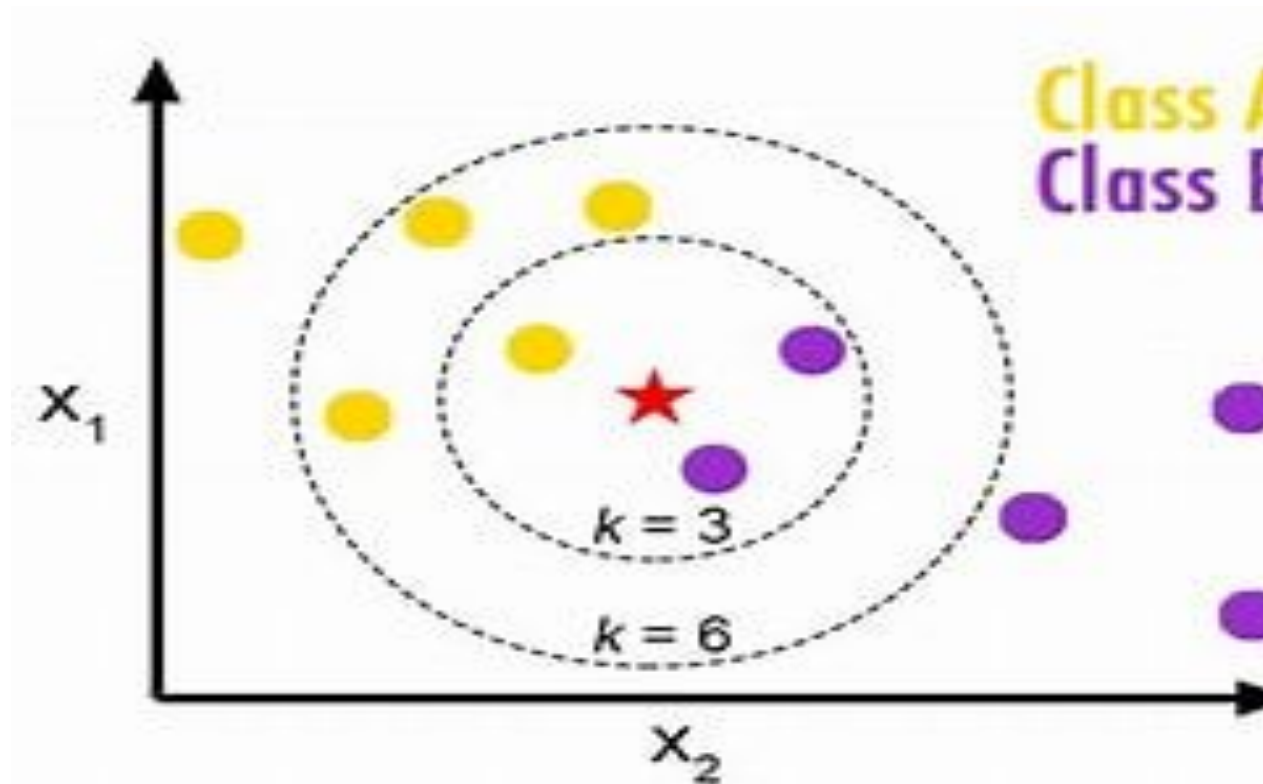
Prediction Algorithm:

1. Calculate the distance from new data point x to all data points in your data set.

2. Sort the data points in your data in the increasing order of distance from x .

3. Predict the majority class label of the “ k ” closest points

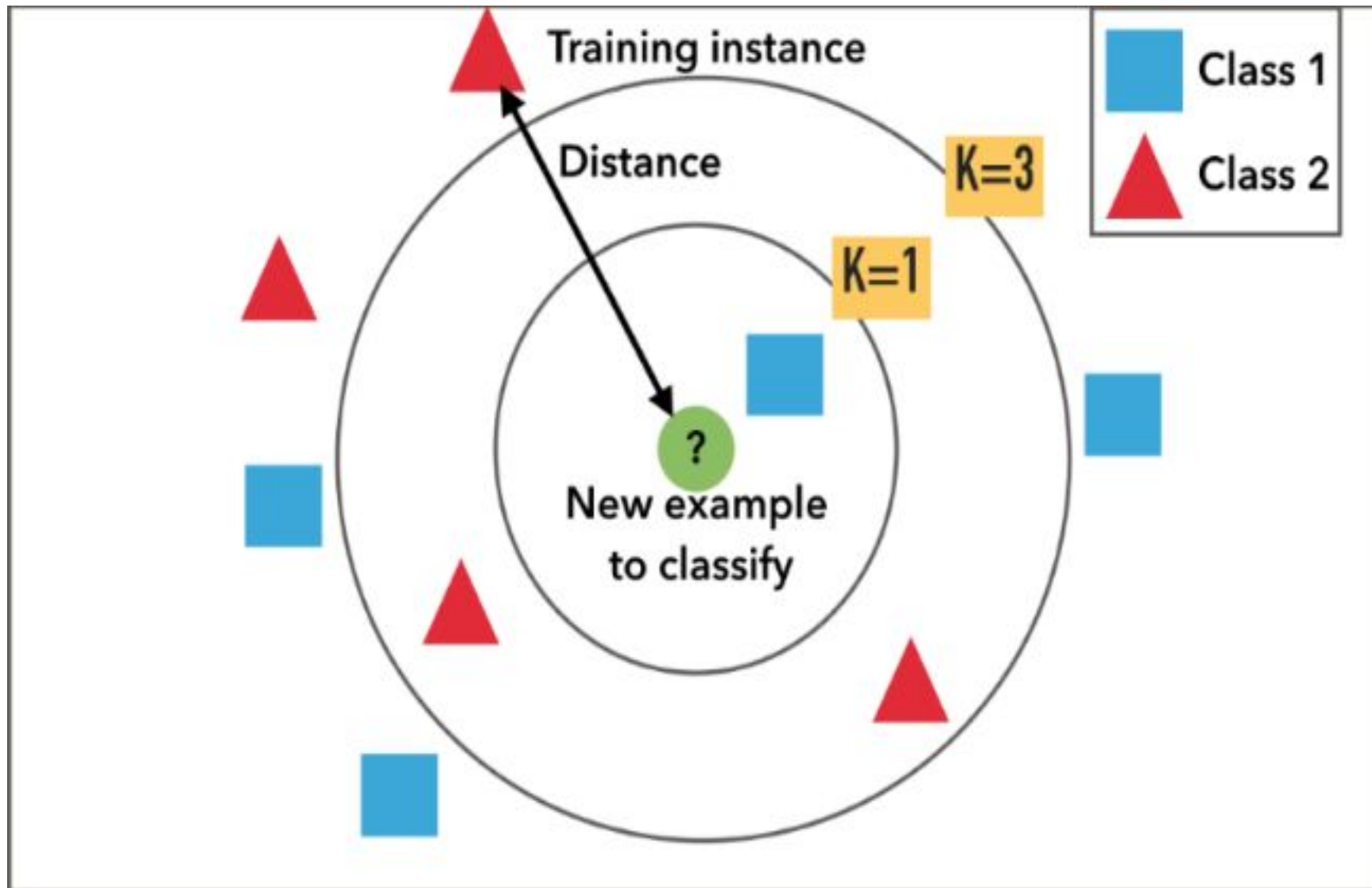
Choosing a K will affect what class a new point is assigned to:



In given example if $k=3$ then new point will be in class B but if $k=6$ then it will be in class A.

Because majority of points in $k=6$ circle are from class A.

KNN Algorithm



KNN Algorithm

K Nearest Neighbors - Example

