K-Nearest Neighbours (K-NN)

K-NN algorithm is a type of supervised Machine Learning algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems.

The K-NN algorithm assumes that similar things exist in close proximity. In other words similar things are near to each other.

KNN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classify a new data point based on the similarity.

K-NN is a non-parametric algorithm, which means it does not make any assumptions on underlying data.

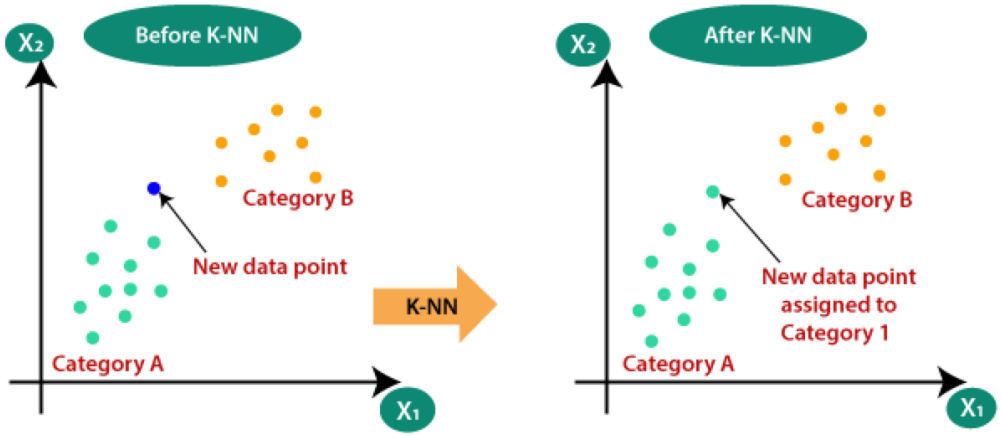
It is also called a Lazy-Learner algorithm because it does not learn from the training set immediately, instead it stores the dataset and at the time of classification, it performs an action on the dataset.

K-NN 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.

Example - Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or a dog. So, for this identification, our K-NN model will find the similar features of the new data set to the cats and dog images and based on the most similar features, it will put it in either cat or dog category.

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 x1, 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:



STEP 1: Choose the number K of neighbors

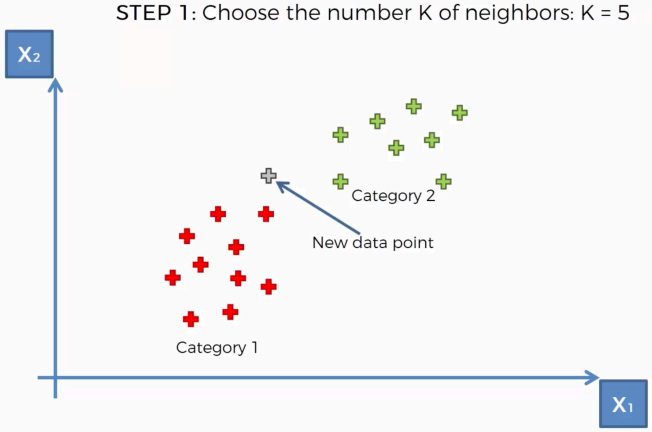
STEP 2: Take the K nearest neighbors of the new data point, according to the Euclidean distance

STEP 3: Among these K neighbors, count the number of data points in each category

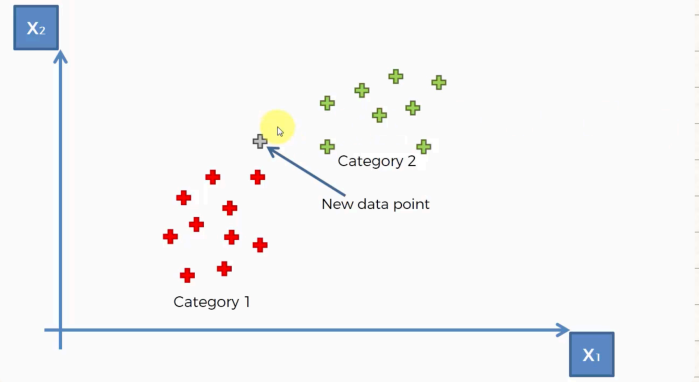
STEP 4: Assign the new data point to the category where you counted the most neighbors

Your Model is Ready

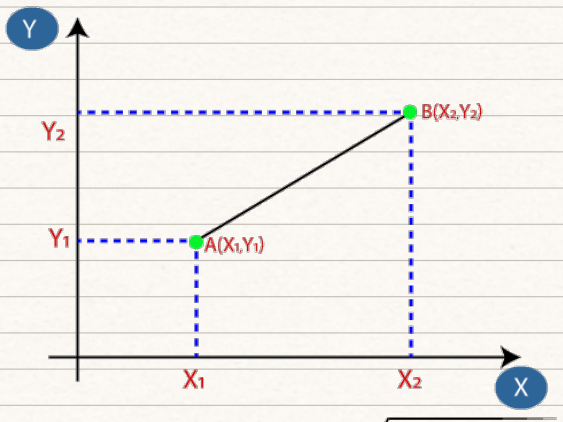
STEP 1: Choose the number K of neighbors: K = 5



STEP 2: Take the K = 5 nearest neighbors of the new data point, according to the Euclidean distance



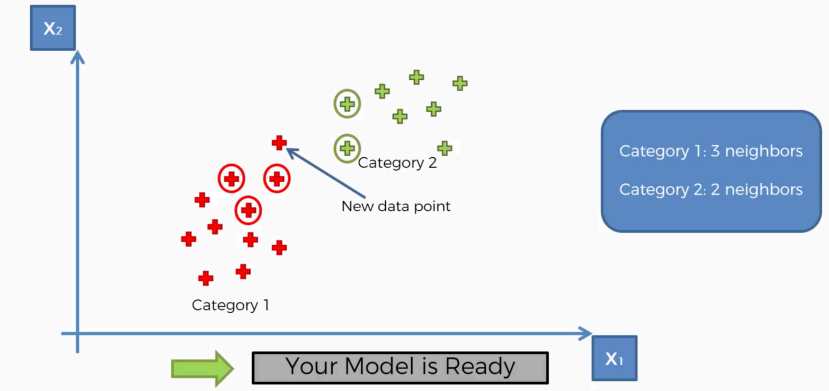
The Euclidean distance is the distance between two points which can be calculated as follows:-



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

STEP 3: Among these K neighbors, count the number of data points in each category

STEP 4: Assign the new data point to the category where you counted the most neighbors



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. 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.

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.

Q.

Sr.	Maths	CS	Result
1	4	3	Fail
2	6	7	Pass
3	7	8	Pass
4	5	5	Fail
5	8	8	Pass

Query → $x = (Maths = 6, CS = 8)$, $k = 3$

Results are out and X student got the above marks. Using the table, find out whether the student is Pass or Fail ?

Euclidean Distance :-

$$d = \sqrt{(x_{o1} - x_{A1})^2 + (x_{o2} - x_{A2})^2}$$

where O stands for observed variable and
A stands for Actual Variable.

Using this formula.

① $\sqrt{(6-4)^2 + (8-3)^2} = \sqrt{29} = 5.38$

② $\sqrt{(6-6)^2 + (8-7)^2} = 1$

③ $\sqrt{(6-7)^2 + (8-8)^2} = 1$

④ $\sqrt{(6-5)^2 + (8-5)^2} = \sqrt{10} = 3.16$

⑤ $\sqrt{(6-8)^2 + (8-8)^2} = 2$

As $k=3$, for this given problem, we will consider Sr. no. 2, 3 and 5 as our nearest neighbours and will make the prediction based on that

Sr.	Maths	CS	Result	Distance
1	4	3	Fail	5.38
2	6	7	Pass	1
3	7	8	Pass	1
4	5	5	Fail	3.16
5	8	8	Pass	2

According to the above observation, we can see that 3 students got passed and no student got failed.

So, the x student is declared Passed.