Machine Learning

K – Nearest Neighbours (KNN)

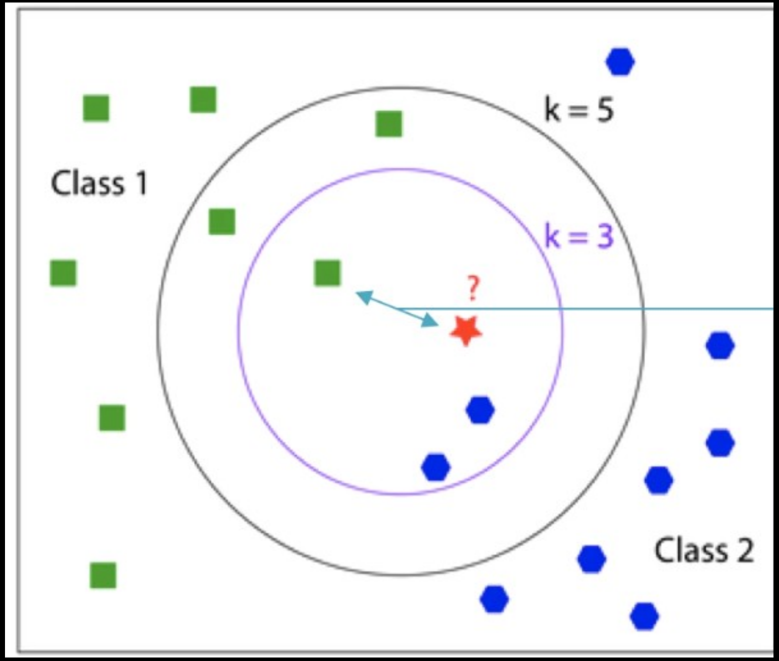
1. K – Nearest Neighbours

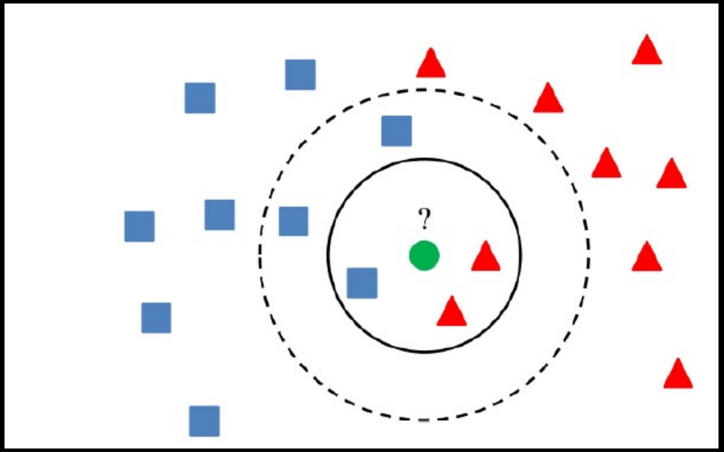
KNN - is the simplest algorithm for Classification in Supervised Learning.

To use KM it is necessary:

1 - Training data.
2 - Define the metric for the distance calculation.
3 - Define the value of K (number of closest neighbors that will be considered by the algorithm).

Consists of calculating the distance between the unknown example and the other examples in the training set





Distância Euclidiana

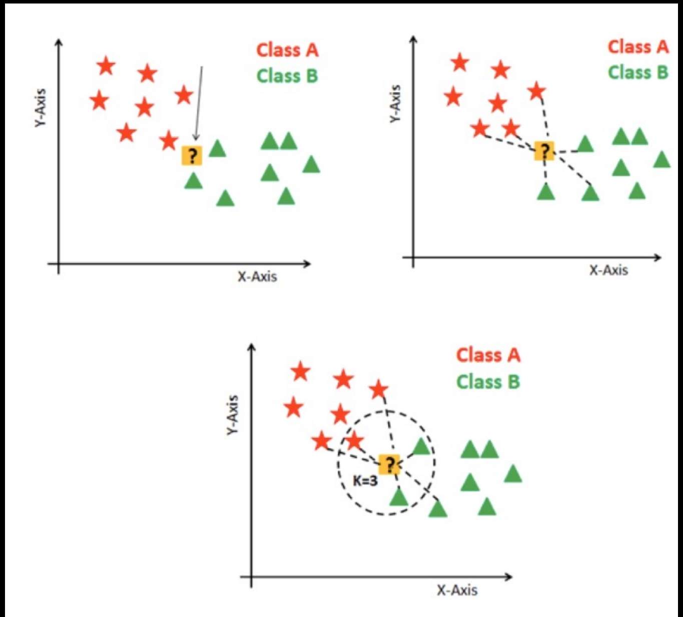
$$d(p,q)=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2}$$

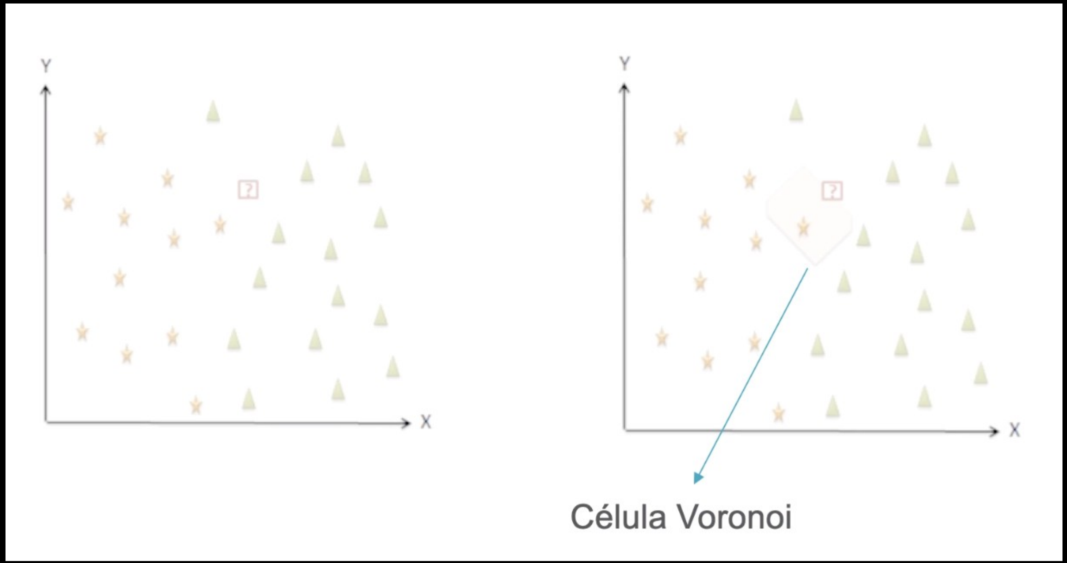
Note:

The data must be **Normalized** before applying the algorithm!

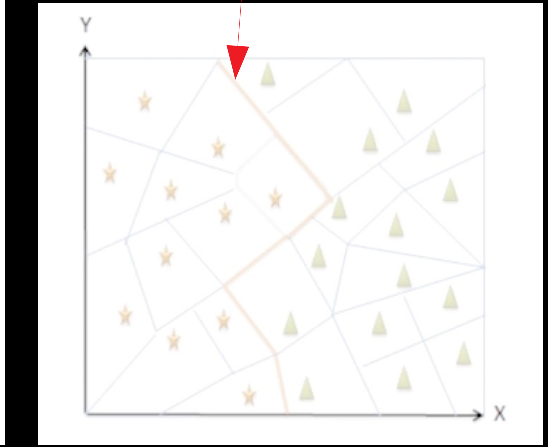
Other ways to measure distance:
- Manhattan distance
- Minkowsky distance
- Hamming distance

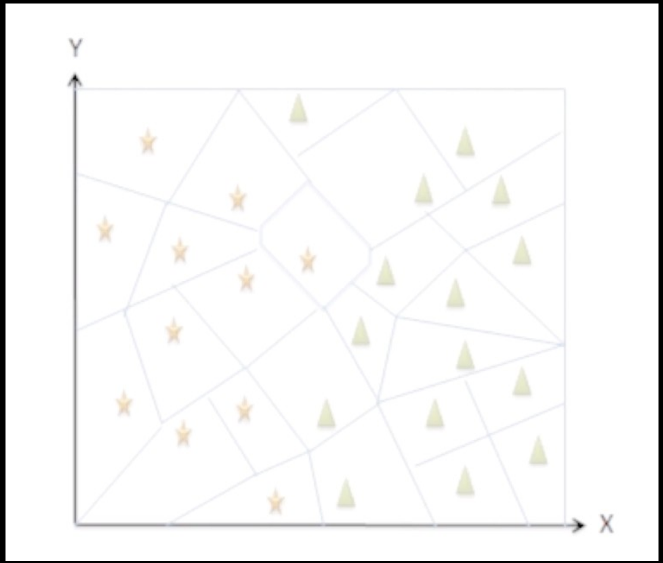
2. KNN and Voronoi Cell Structure





Border

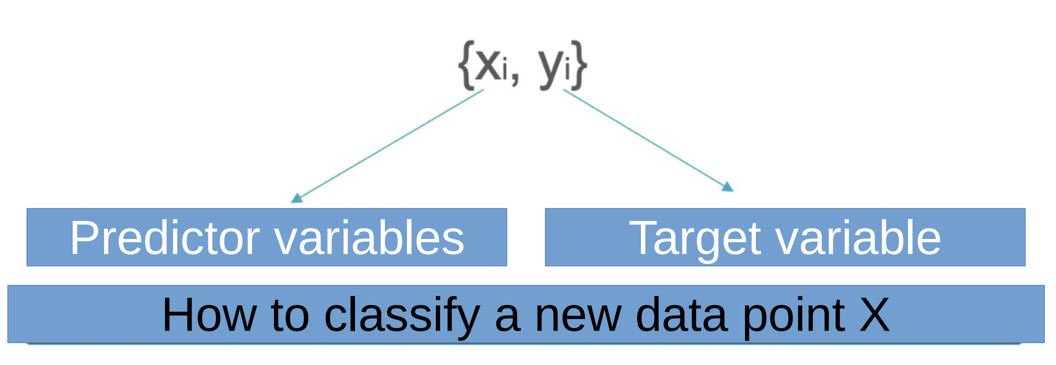




The size of each cell is determined by the number of examples available in the training data set.

The outliers must be removed from the data set.

3. How KNN works

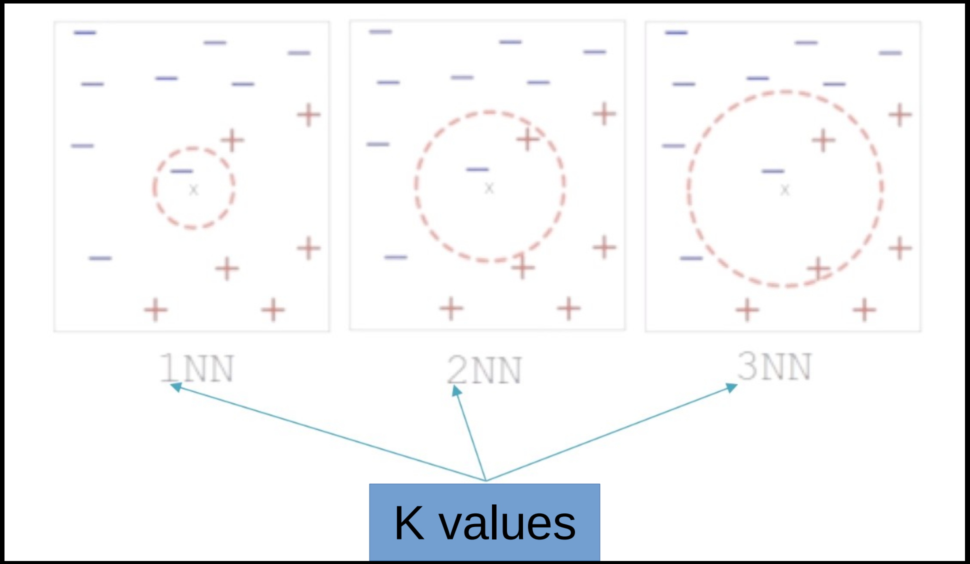


$\{x_i, y_i\}$

Predictor variables

Target variable

How to classify a new data point X



K values

1 - The distance is computed between X and Xi for each value of Xi.

2 - The closest neighbor to Xin and its respective class is chosen.

3 - The most frequent value of y in the list yi1, yi2, ..., yin is returned.

The main purpose of the distance measure is to identify data that are similar and that are not similar.

4. Math Distance Measures

As well as the K value, the distance measurement directly influences the performance of models created with KNN.

Minkowski distance

$$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$$

Hamming distance

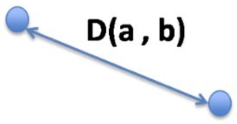
$$D_H = \sum_{i=1}^{k}|x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

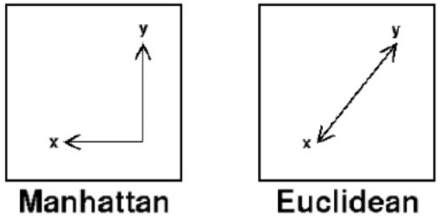
$$x \neq y \Rightarrow D = 1$$

Euclidean distance

$$D(a,b) = \sqrt{\sum_{i=1}^{n}(b_i - a_i)^2}$$



Manhattan and Euclidean distances are the most common



The accuracy of the classification using the KNN algorithm strongly depends on the data model. Most of the time, the attributes need to be normalized to avoid distance measures being dominated by a single attribute.