**KNN (K- Nearest Neighbors) Algorithm**

1. What is K- Nearest neighbors?

* It is Supervised machine learning algorithm as target variable is known.
* Non parametric as it does not make an assumption about the underlying data distribution pattern.
* It is Lazy algorithm as KNN does not have a training step.
* All data points will be used only at the time of prediction. It Used for both Classification and Regression

1. What is K is K nearest neighbors?

* K is a number used to identify similar neighbors for the new data point.
* KNN takes K nearest neighbors to decide where the new data point with belong to. This decision is based on feature similarity.



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.

1. How do we chose the value of K?

* We can evaluate accuracy of KNN classifier using K fold cross validation.
* Using Elbow Method To Find Right Value OF K.

1. 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**
* **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
* **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 neighbor is maximum.
* **Step-6:** Our model is ready.

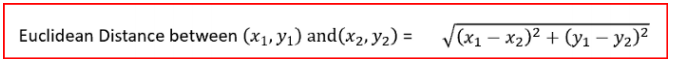
**All Distance:**

There are Three Types Of Distance

* Euclidian Distance (L2 Norm)
* Manhattan Distance (L1 Norm)
* Minkowski Distance (Lp Norm)
* Hamming Distance

Euclidian Distance (L2 Norm)

Euclidean distance is the **square root of the sum of squared distance** between two points. It is also known as L2 norm.



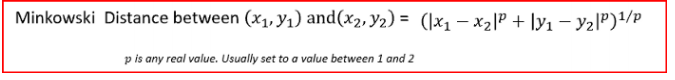
Manhattan Distance (L1 Norm)

Manhattan distance is the **sum of the absolute values of the differences between two points.** Also known as L1 norm.



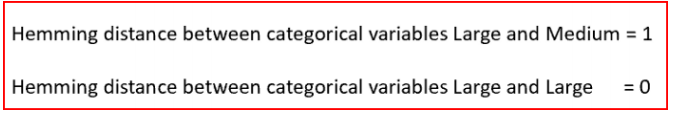
Minkowski Distance (Lp Norm)

Minkowski distance is the used to find distance similarity between two points. When p=1, it becomes Manhattan distance and when p=2, it becomes Euclidean distance.



Hamming Distance

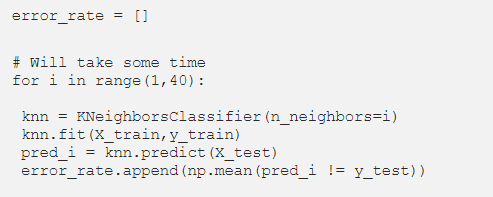
Hamming distance is used for categorical variables. In simple terms it tells us if the two categorical variables are same or not.



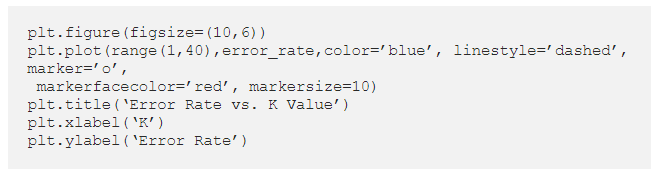
Elbow Method for K value :

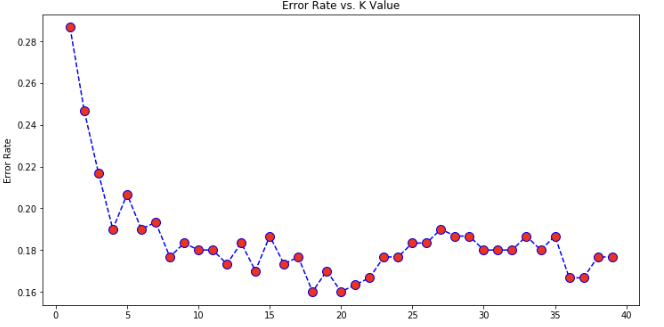
The optimal value of k reduces effect of the noise on the classification, but makes boundaries between classes less distinc.

Elbow method helps data scientists to select the optimal number of clusters for KNN clustering. It is one of the most popular methods to determine this optimal value of K.



To view Elbow Method





The optimal value of K is 20. The graph of the error rate increases after 20.

Retrain your model with the best K value and re-do the classification report and the confusion matrix.

