MACHINE LEARNING IN PRATICE

- 1. Data collection
- 2. Data exploration and preparation
- 3. Model training
- 4. Model evaluation
- 5. Model improvement

1. Training and Testing Datasets

The dataset used to create the model, with known attributes and target, is called the training dataset.

The validity of the created model will also need to be checked with another known dataset called the test dataset or validation dataset.

To facilitate this process, the overall known dataset can be split into a training dataset and a test dataset.

A standard rule of thumb is two-thirds of the data are to be used as training and one-third as a test dataset

Table 2.3 Training Dataset

Borrower	Credit Score (*X*)	Interest Rate (*Y*) (%)
01	500	7.31
02	600	6.70
03	700	5.95
05	800	5.40
06	800	5.70
08	550	7.00
09	650	6.50

Table 2.4 Test Dataset

Borrower	Credit Score (*X*)	Interest Rate (*Y*)
04	700	6.40
07	750	5.90
10	825	5.70

Lazy Learning – Classification Using Nearest Neighbors

K-Nearest Neighbor(KNN) Algorithm

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.

- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. 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.

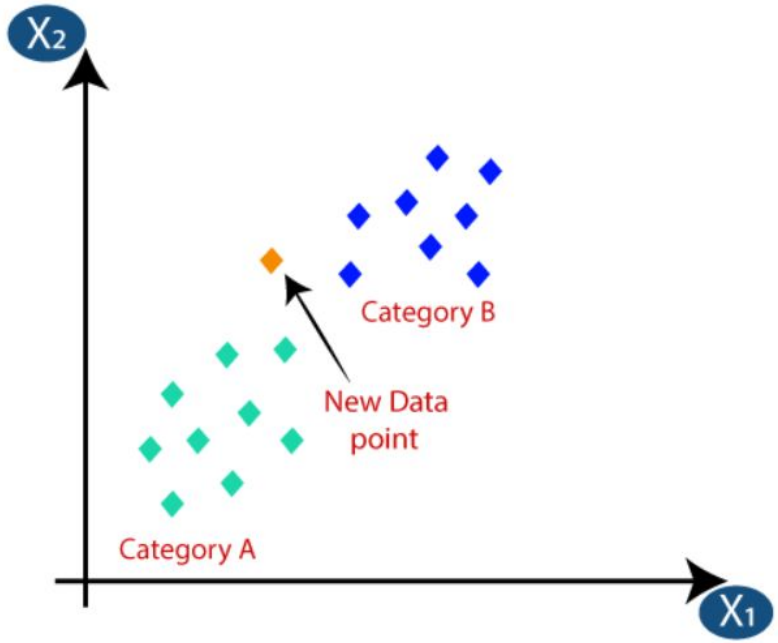
Why it called lazy learner ?

- It is also called a **lazy learner algorithm** because it

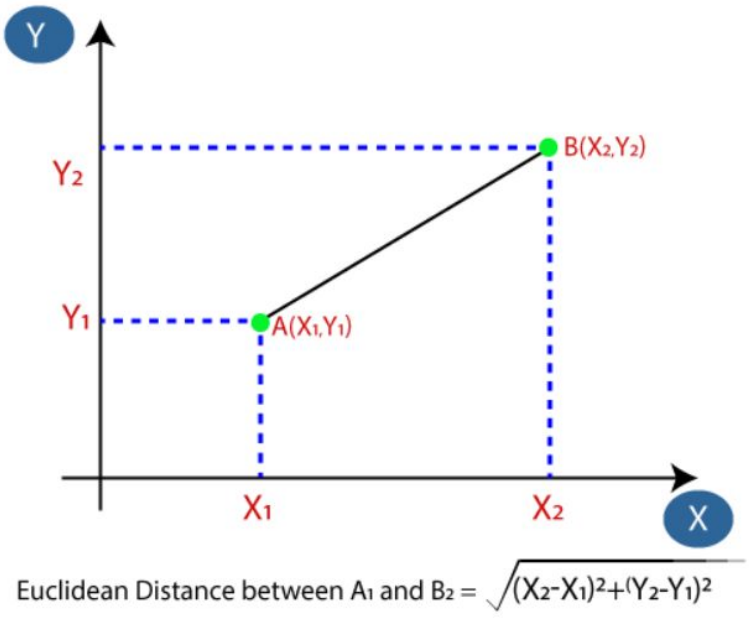
How does K-NN work?

- 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 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.

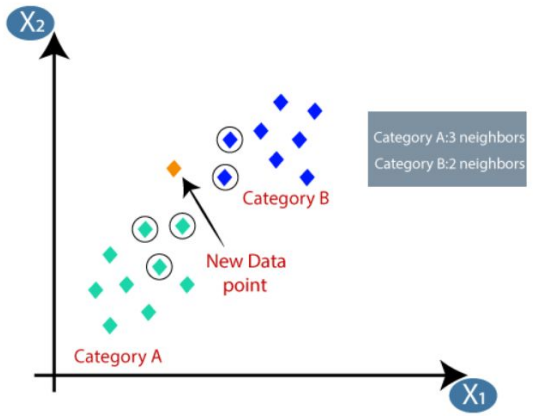
Suppose it is need to put a new point in the required category. Consider the below image:



- Firstly, choose the number of neighbors, so we will choose the k=5.
- Next, calculate the **Euclidean distance** between the data points. The Euclidean distance is the distance between two points. It can be calculated as:



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:



3 nearest neighbors are from category A, hence this new data point must belong to category A.

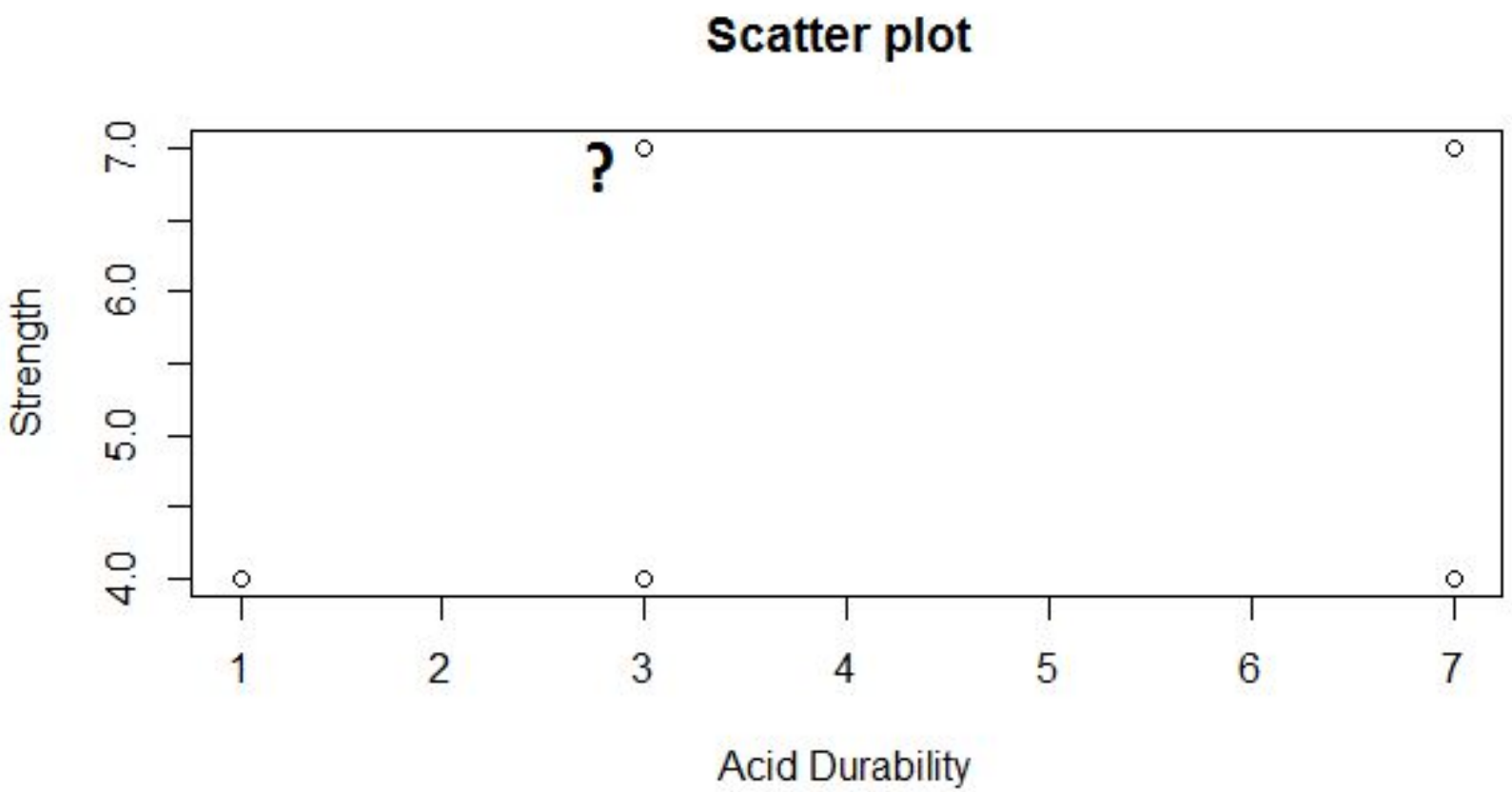
Example

Points	X1 (Acid Durability)	X2(strength)	Y=Classification
P1	7	7	BAD
P2	7	4	BAD
P3	3	4	GOOD
P4	1	4	GOOD

KNN Example

Points	X1(Acid Durability)	X2(Strength)	Y(Classification)
P1	7	7	BAD
P2	7	4	BAD
P3	3	4	GOOD
P4	1	4	GOOD
P5	3	7	?

Scatter Plot



Euclidean Distance From Each Point

KNN				
Euclidean Distance of P5(3,7) from	P1	P2	P3	P4
	(7,7)	(7,4)	(3,4)	(1,4)

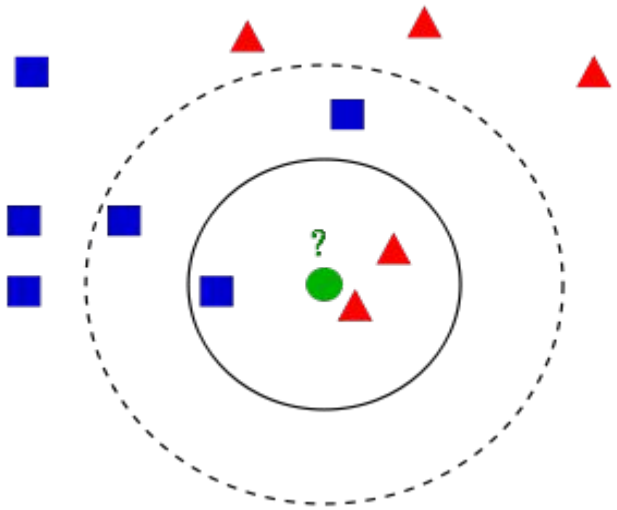
3 Nearest NeighBour

Euclidean Distance of P5(3,7) from	P1	P2	P3	P4
	(7,7)	(7,4)	(3,4)	(1,4)
Class	BAD	BAD	GOOD	GOOD

KNN Classification

Points	X1(Durability)	X2(Strength)	Y(Classification)
P1	7	7	BAD
P2	7	4	BAD
P3	3	4	GOOD
P4	1	4	GOOD
P5	3	7	GOOD

Variation In KNN



Different Values of K