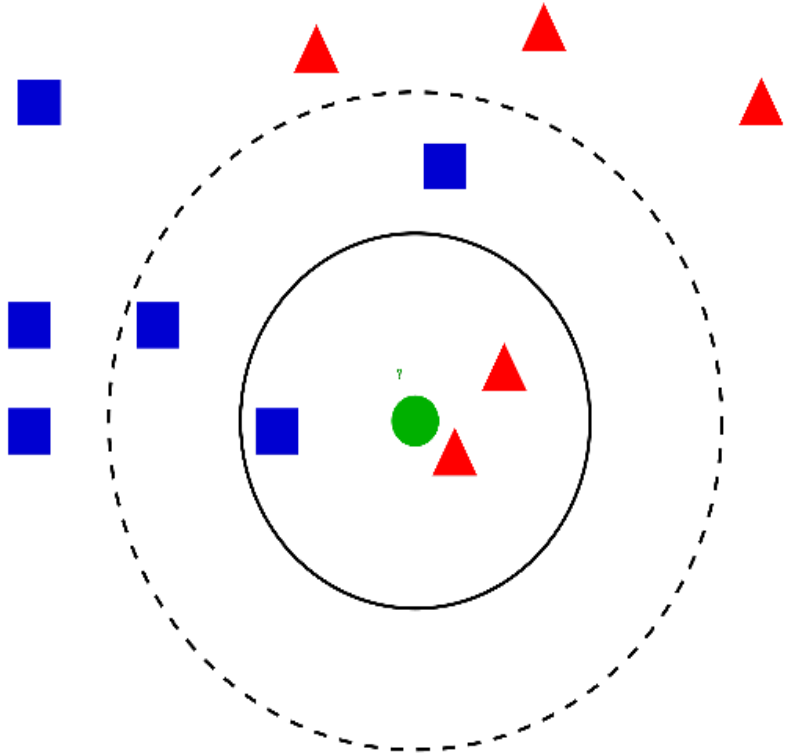
K nearest neighbor (KNN) algorithm

KNN

- k-nearest neighboralgorithm (k-NN) is a **non-parametric supervised learning method**

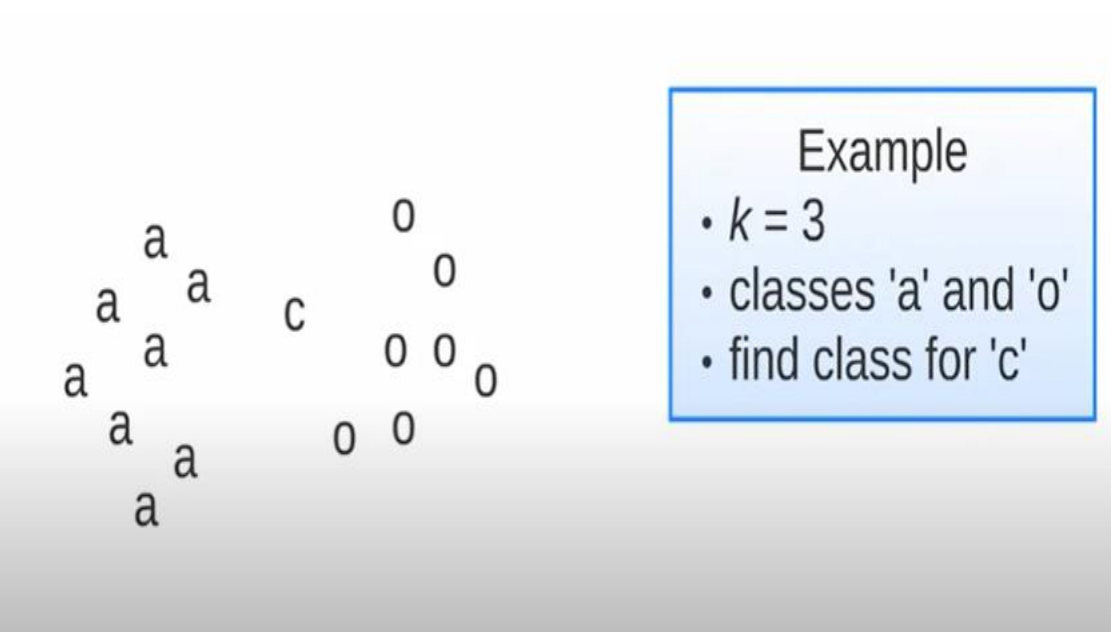
- KNN *predicts the label or value of a new data point by considering its K closest neighbours in the training dataset.*



- K can be odd numbers…K=3,5,7…..

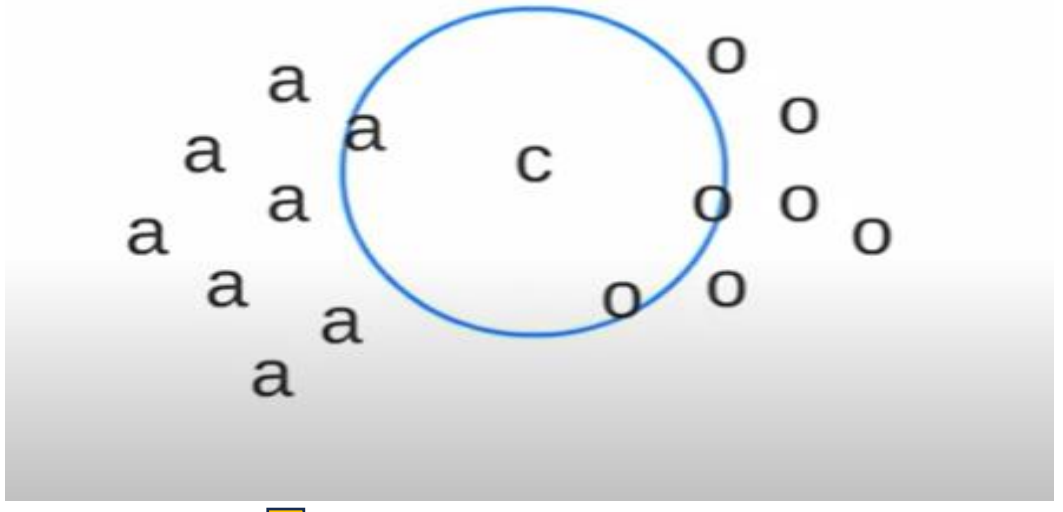
- The K-NN algorithm works by finding the K nearest neighbors to a given data point based on a distance metric, such as ***Euclidean distance***.

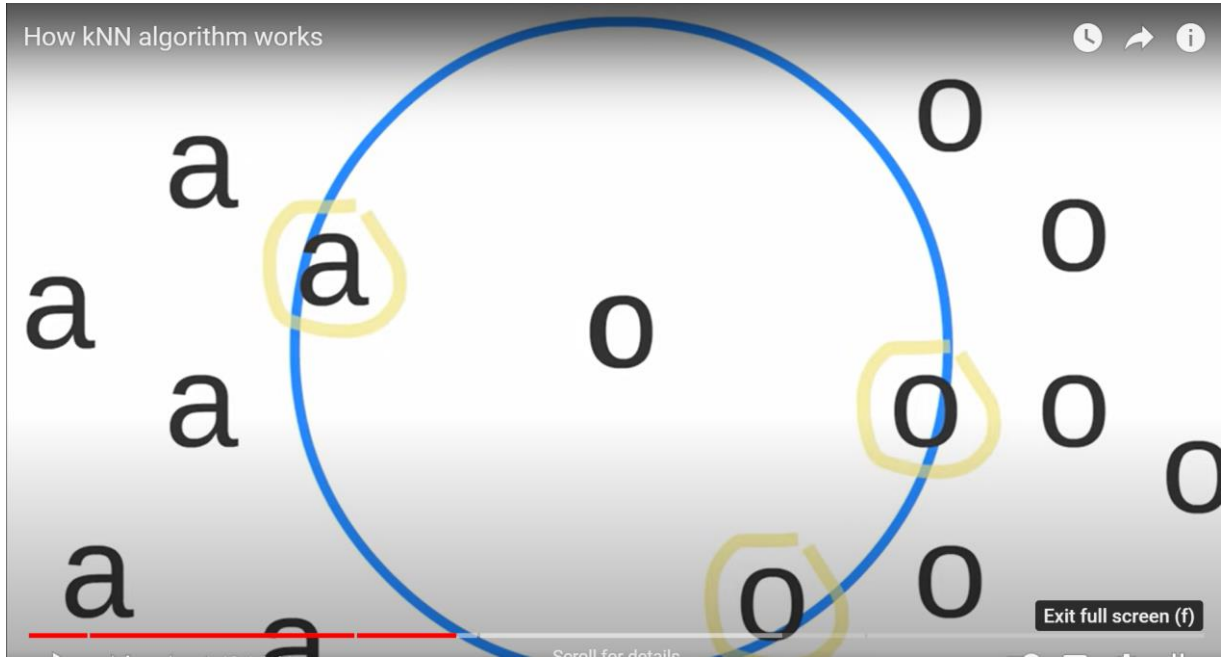
- The class or value of the data point is then determined by the **majority vote or average of the K neighbors**.



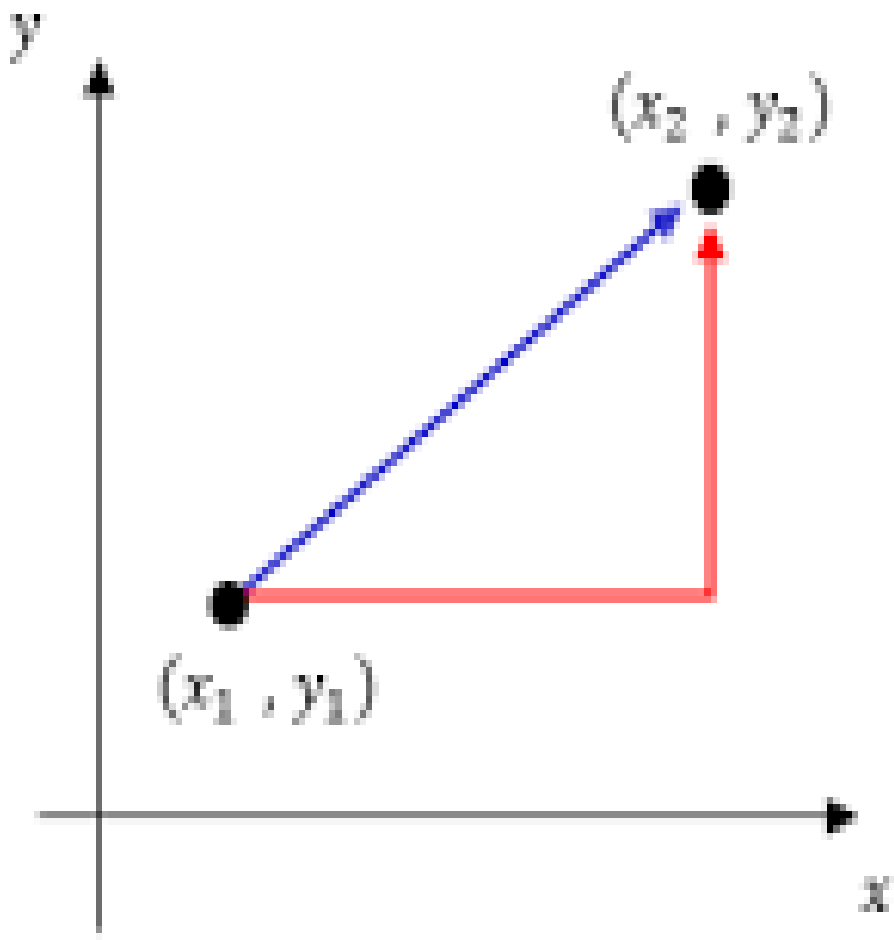
Example

- $k = 3$
- classes 'a' and 'o'
- find class for 'c'





Distance Metrics Used in KNN Algorithm

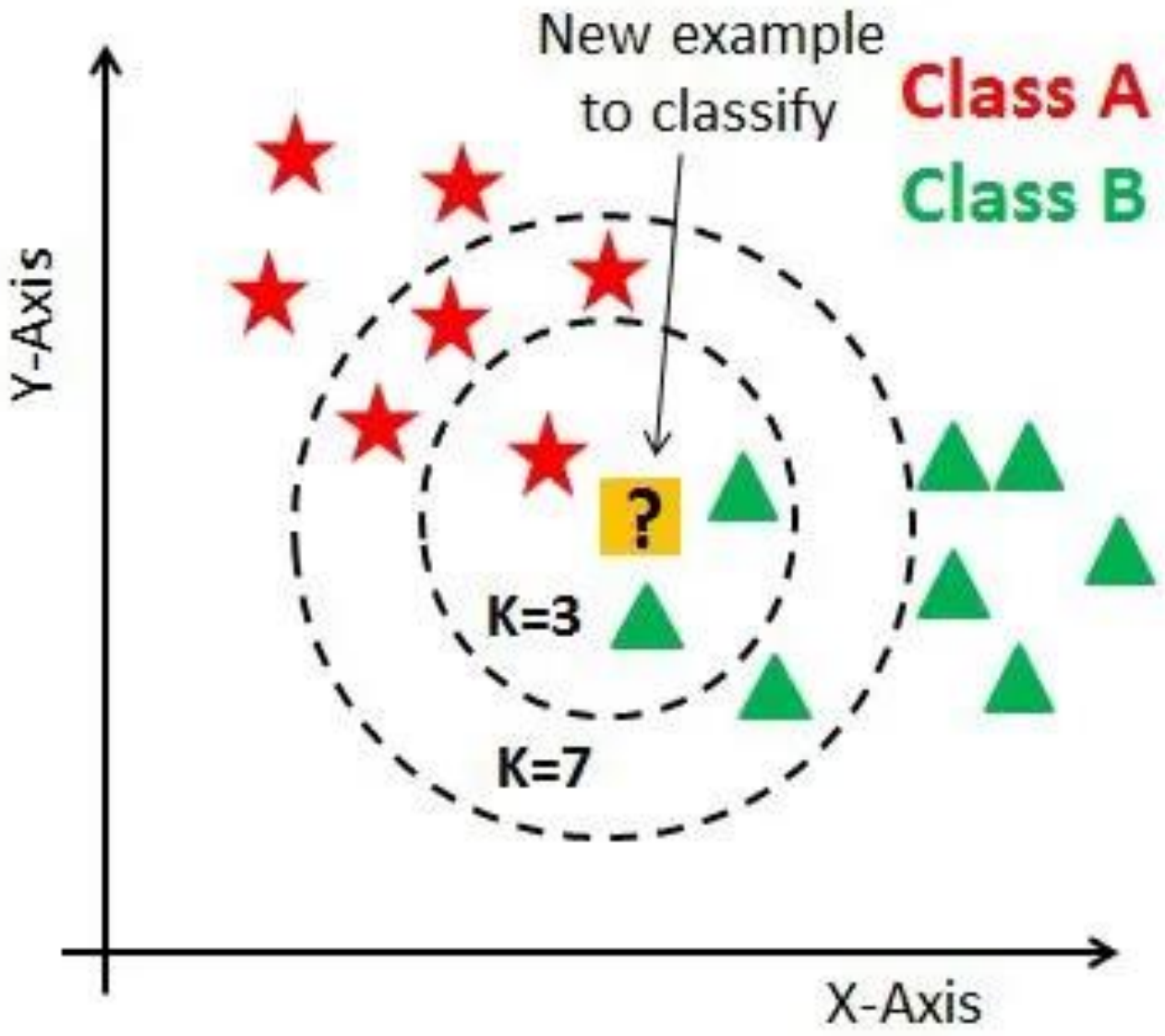


Manhattan Distance L^1

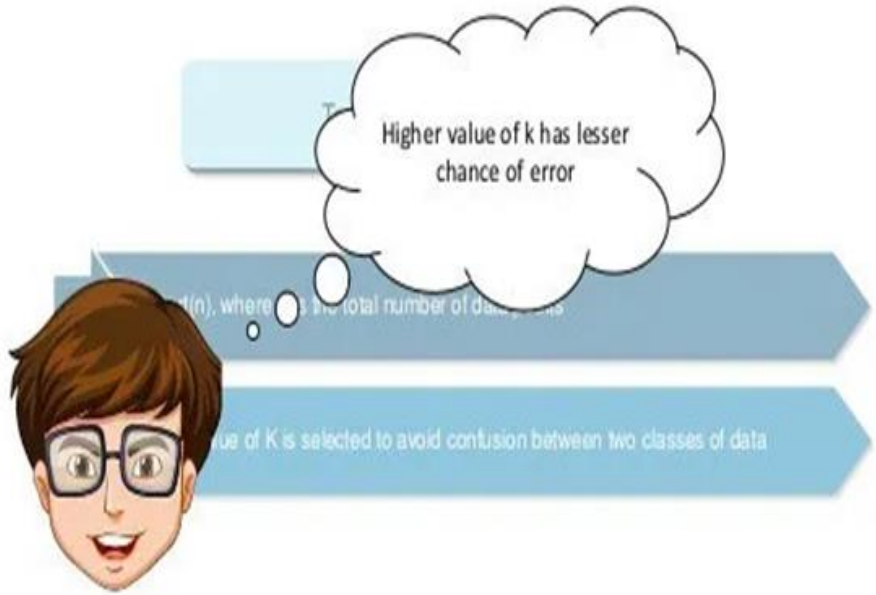
Euclidean Distance L^2

$$L^1 = |x_2 - x_1| + |y_2 - y_1|$$

$$L^2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$



How do we choose the factor 'k'?



KNN algorithm

1.Load the data

2. Initialize the value of **k**

3. Calculate the **distance** between test data and each row of training dataset by using Euclidean Distance.

4. **Sort the calculated distances** in ascending order based on distance values

5. Get top **k** rows from the sorted array

6. Get the most frequent class of these rows (majority Vote)

7. Return the predicted class

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 e.g.

Name	Age	Gender	sport
Ajay	32	M	Football
Mark	40	M	Neither
Sara	16	F	Cricket
Zaira	34	F	Cricket
Sachin	55	M	Neither
Rahul	40	M	Cricket
Pooja	20	F	Neither
Smith	15	M	Cricket
Laxmi	55	F	Football
Michael	15	M	Football
Anjali	**5**	**F**	**?**

Suppose K=3

Name	Age	Gender	Distance	Class of sport
Ajay	32	0	27.02	Football
Mark	40	0	35.01	Neither
sara	16	1	11.00	Cricket
Zaira	34	1	9.00	Cricket
Sachin	55	0	50.01	Neither
Rahul	40	0	35.01	Cricket
Pooja	20	1	15.00	Neither
Smith	15	0	10.00	Cricket
Laxmi	55	1	50.00	Football
Michael	15	0	10.05	Football

male = 0

Female = 1

Ajay male = 0 Age = 32

$$= \sqrt{(5-32)^2 + (1-0)^2}$$

$$= \sqrt{729 + 1}$$

$$= 27.02$$

Zaira 9 → Cricket

Smith 10 → Cricket

Michael 10.05 → Football

Anjali	5	F	Cricket

Select the small 3 distances because K=3

Solve using KNN

Ques:) Perform KNN-classification Algorithm on following dataset and Predict the class for x ($P_1 = 3$ and $P_2 = 7$). K=3

P_1	P_2	Class
7	7	False
7	4	False
3	4	True
1	4	True