Supervised Learning

(K-Nearest Neighbor)

Dr. Virendra Singh Kushwah

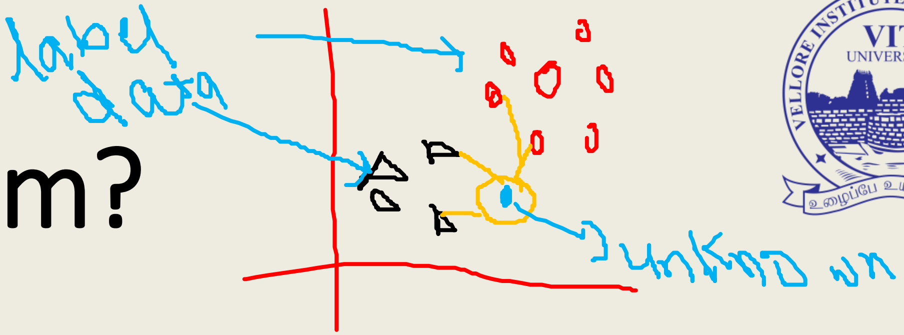
Assistant Professor Grade-II

School of Computing Science and Engineering

Virendra.Kushwah@vitbhopal.ac.in

7415869616

What is KNN algorithm?



- KNN which stands for K Nearest Neighbor is a Supervised Machine Learning algorithm that classifies a new data point into the target class, depending on the features of its neighboring data points.
- K nearest neighbors or KNN Algorithm is a simple algorithm which uses the entire dataset in its training phase. Whenever a prediction is required for an unseen data instance, it searches through the entire training dataset for k-most similar instances and the data with the most similar instance is finally returned as the prediction.
- k-NN is often used in search applications where you are looking for similar items, like find items similar to this one.

K-N.N How many data points are more closer?

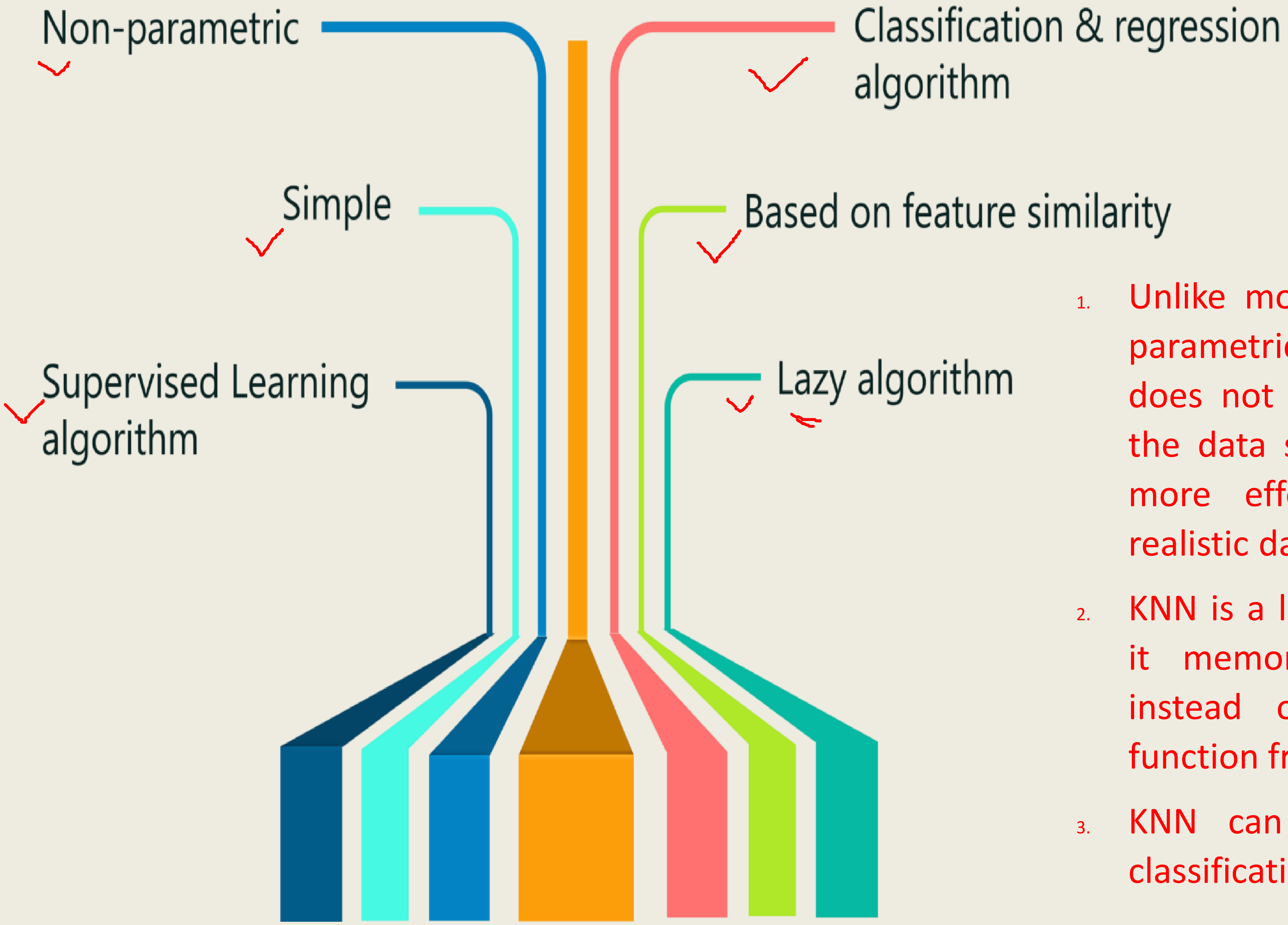
Features of KNN Algorithm

- The KNN algorithm has the following features:

1. ✓ KNN is a Supervised Learning algorithm that uses labeled input data set to predict the output of the data points.
2. ✓ It is one of the simplest Machine learning algorithms and it can be easily implemented for a varied set of problems.
3. ✓ It is mainly based on feature similarity. KNN checks how similar a data point is to its neighbor and classifies the data point into the class it is most similar to.

K-value

K = 1, 2, 3, 4, - - N



✓ Non-parametric

✓ Simple

✓ Supervised Learning algorithm

✓ Classification & regression algorithm

✓ Based on feature similarity

✓ Lazy algorithm

1. Unlike most algorithms, KNN is a non-parametric model which means that it does not make any assumptions about the data set. This makes the algorithm more effective since it can handle realistic data.
2. KNN is a lazy algorithm, this means that it memorizes the training data set instead of learning a discriminative function from the training data.
3. KNN can be used for solving both classification and regression problems.

KNN Algorithm Example

VIT UNIVERSITY
VIT BHOPAL
www.vitbhopal.ac.in

K? NN ?

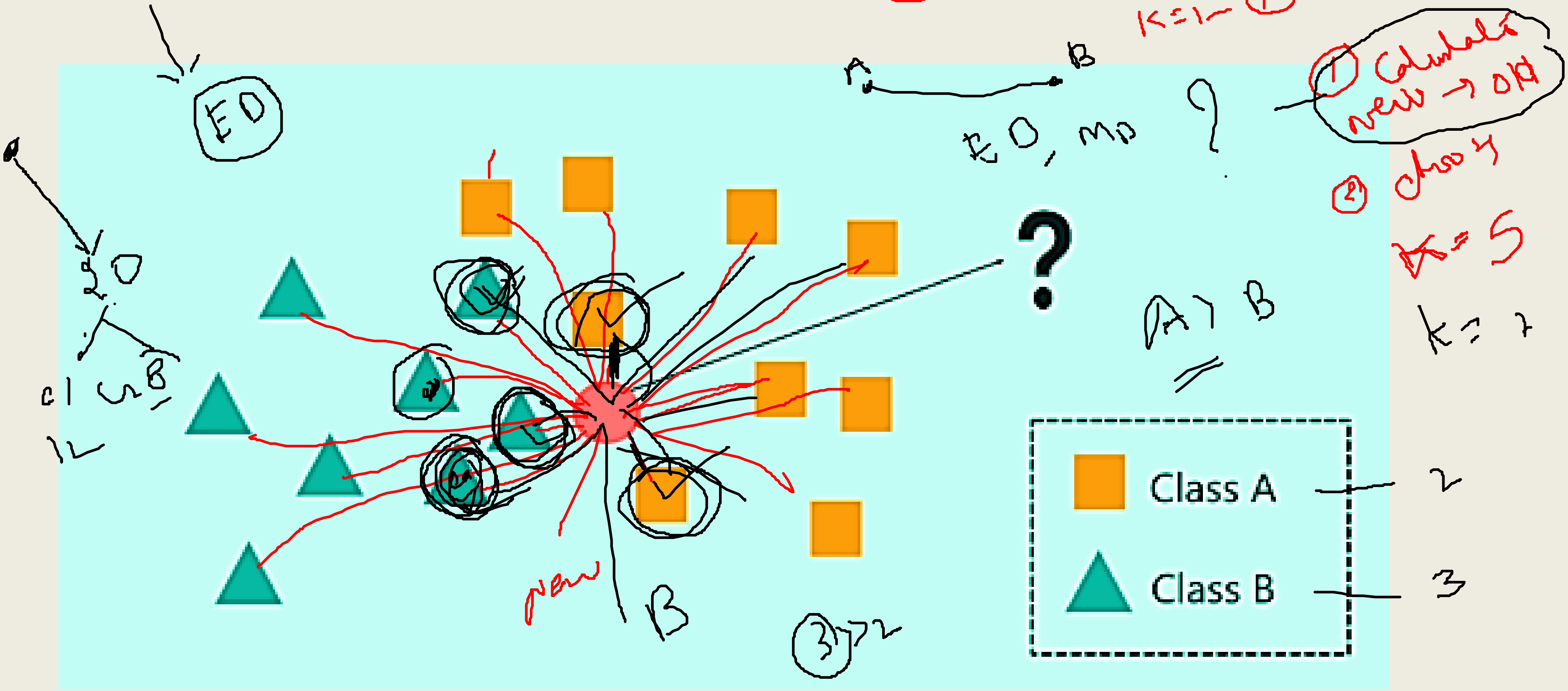
K=1 — NN

① Calculate new → old

② choose

K=5

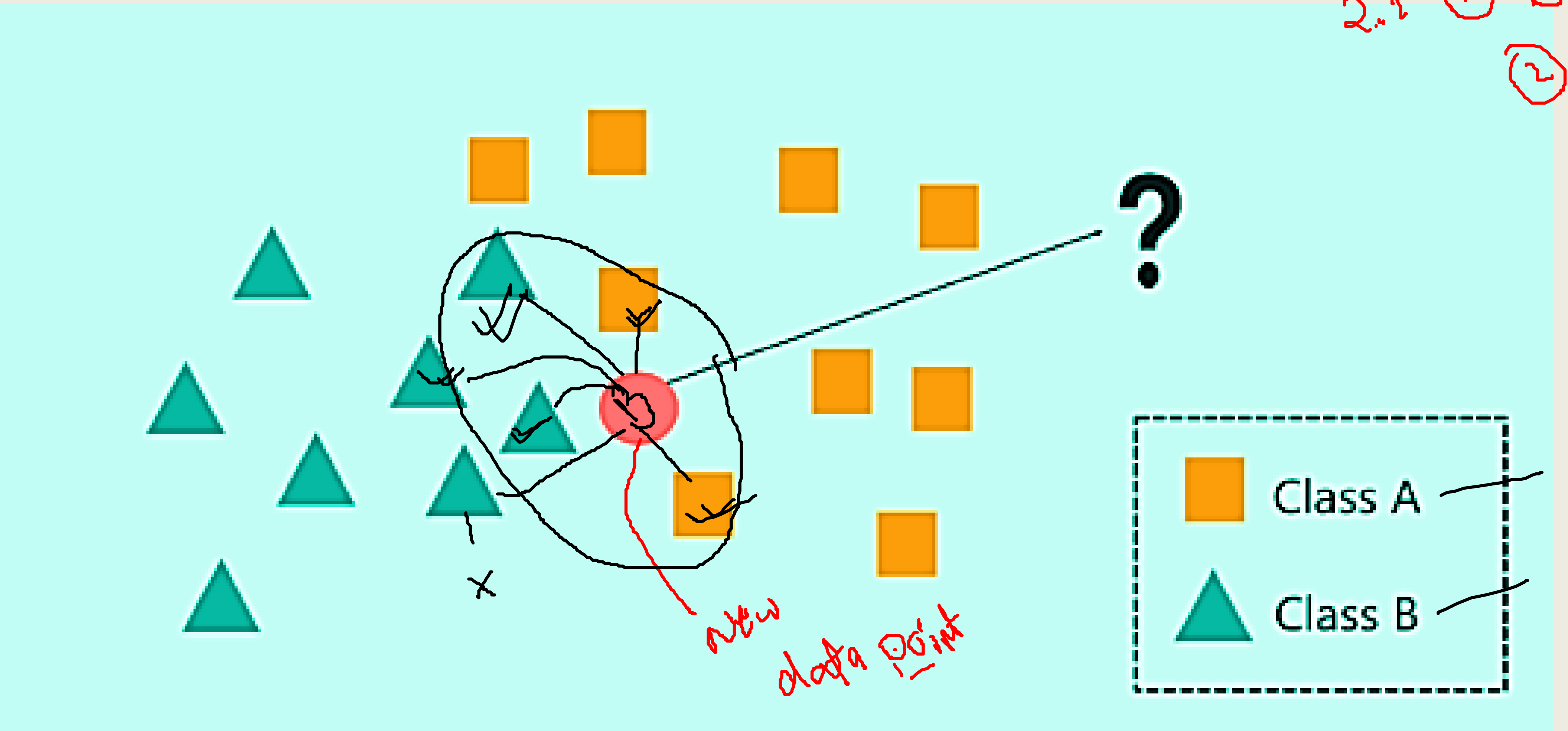
E D, MD ?



KNN Algorithm Example

k = 5
data point

2..2 ① Distance
② kth
1,3,5,7,9 ,



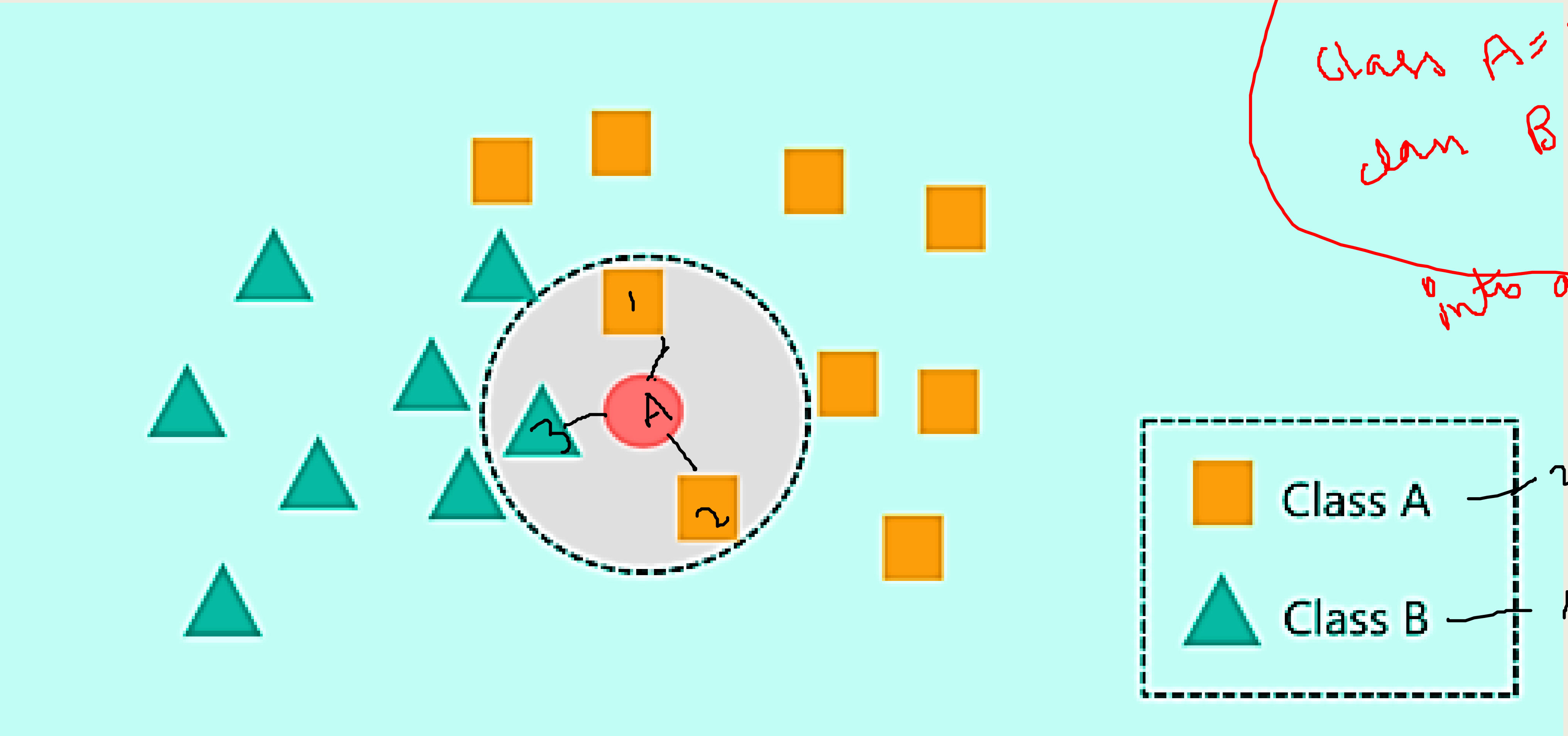
2 < ③

2

3

We have to choose minimum 3 distance based values from N → 10

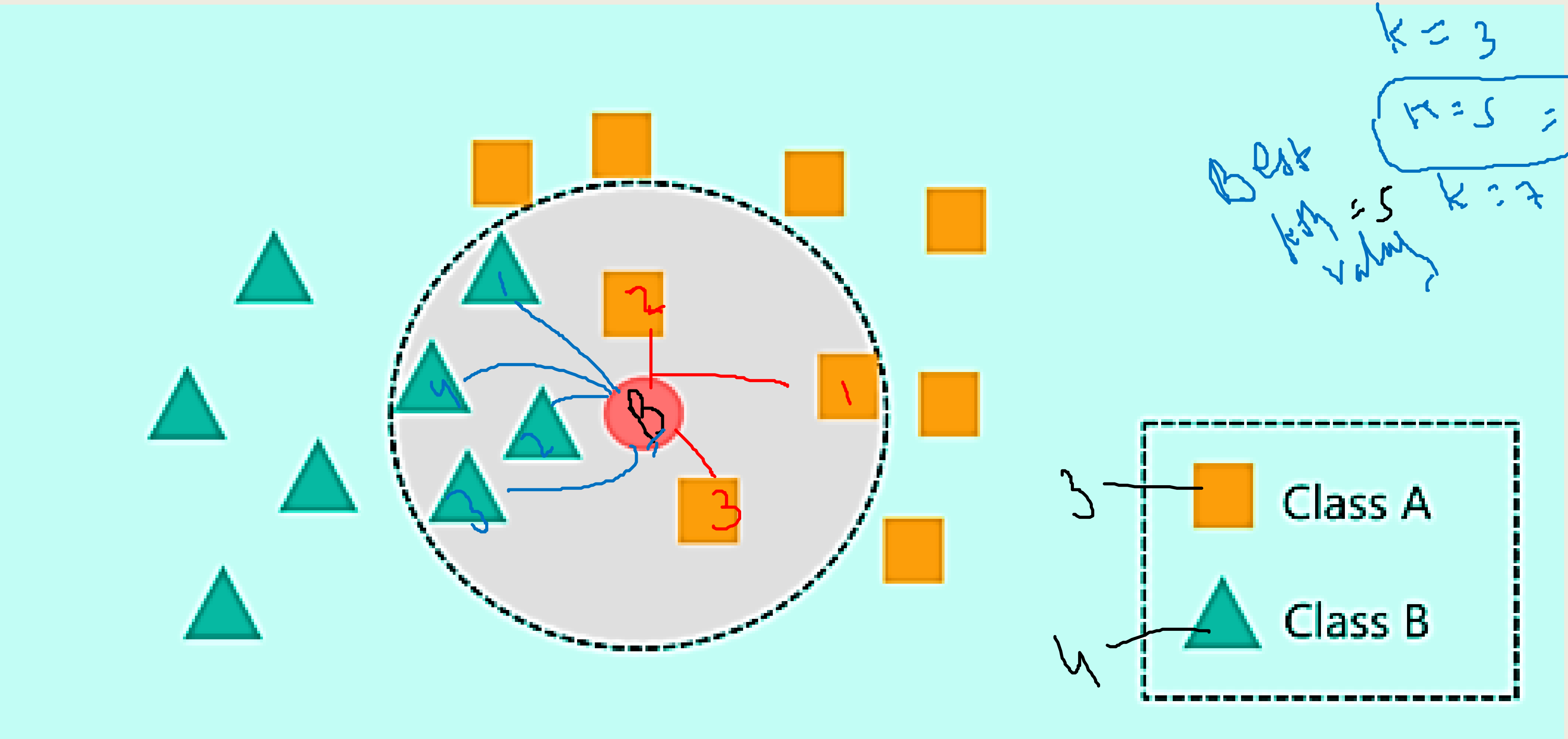
K = 3



Class A = 2
class B = 1

into a specific distance

K = 3 to 5, 7



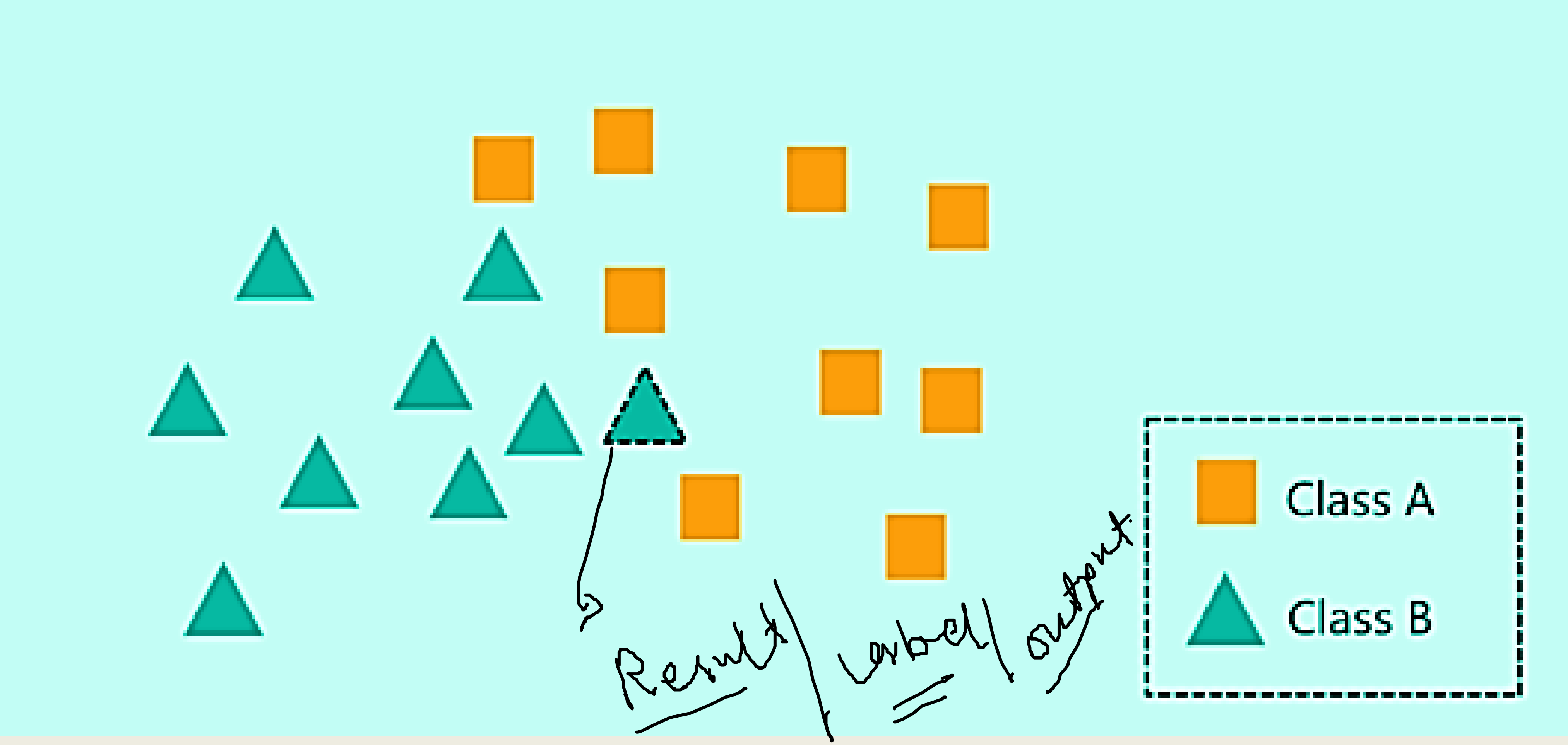
K = 3 82%. ✓

K = 5 = 88%. ✓

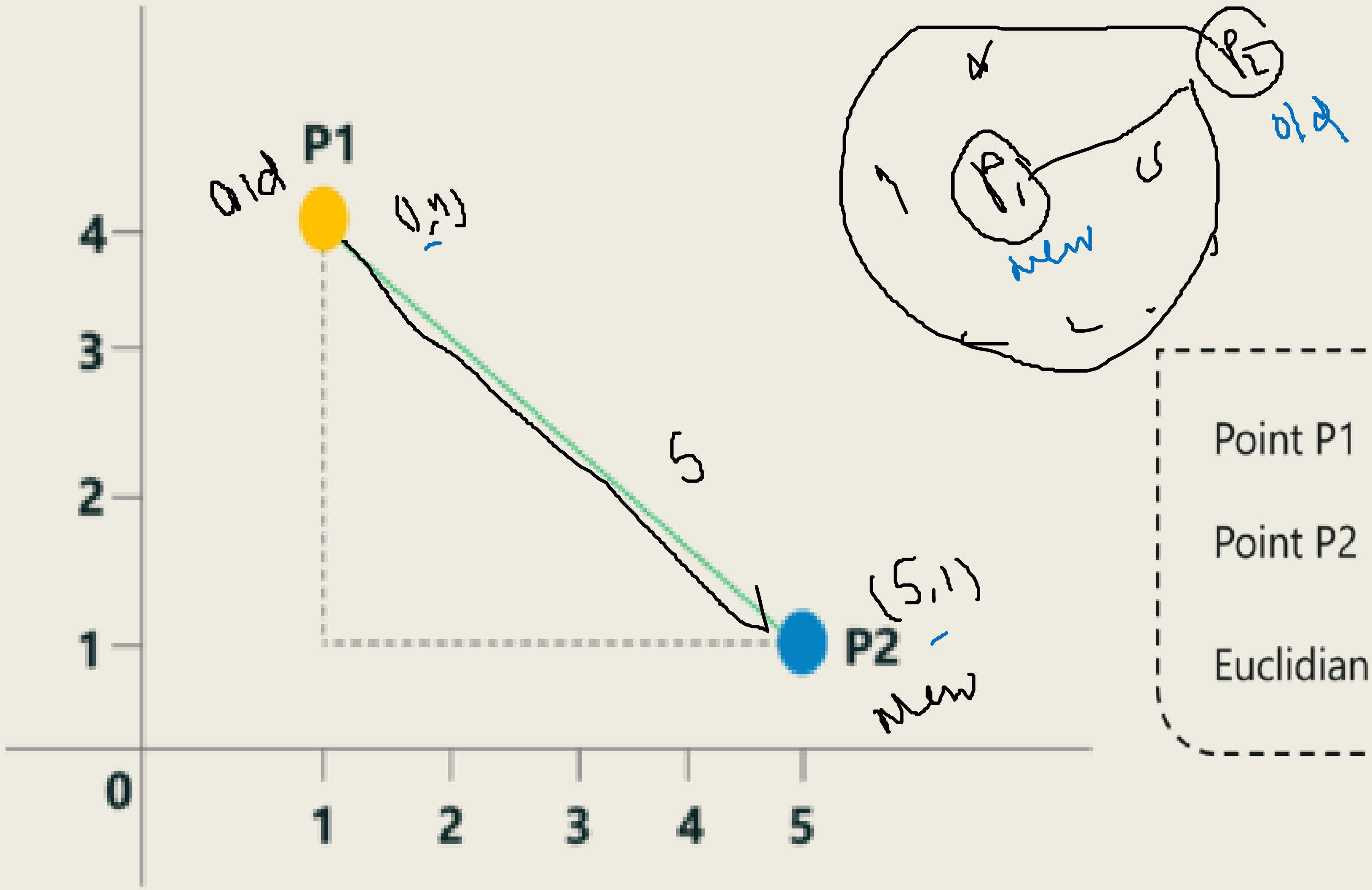
K = 7 = 83%. ✓

Best K = 5 value

K = 1 to 10



Result / label / output



Point P1 = (1,4) ✓

Point P2 = (5,1) ✓

Euclidian distance = $\sqrt{(5-1)^2 + (4-1)^2} = 5$

$= \sqrt{4^2 + 3^2}$

$= \sqrt{16+9} = \sqrt{25}$

$= \boxed{5}$

The KNN Algorithm

1. Load the data set or training dataset
2. Initialize K to your chosen number of neighbors K=1-n
3. For each data point in the data set
    - 3.1 Calculate the distance between the new data point and the existing data point from the data.
    - 3.2 Add the distance and the index of the datapoint to an ordered collection
4. Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances → min (Top) max (Bottom)
5. Pick the first K entries from the sorted collection
6. Get the labels of the selected K entries
7. If regression, return the mean of the K labels
8. If classification, return the mode of the K labels

Let us understand an example in detail

Let us consider another example

I.V

- Suppose we have height and weight and its corresponding T-shirt size of several customers. Your task is to predict the T-shirt size of Virendra, whose height is 161cm and his weight is 61kg.

D.V

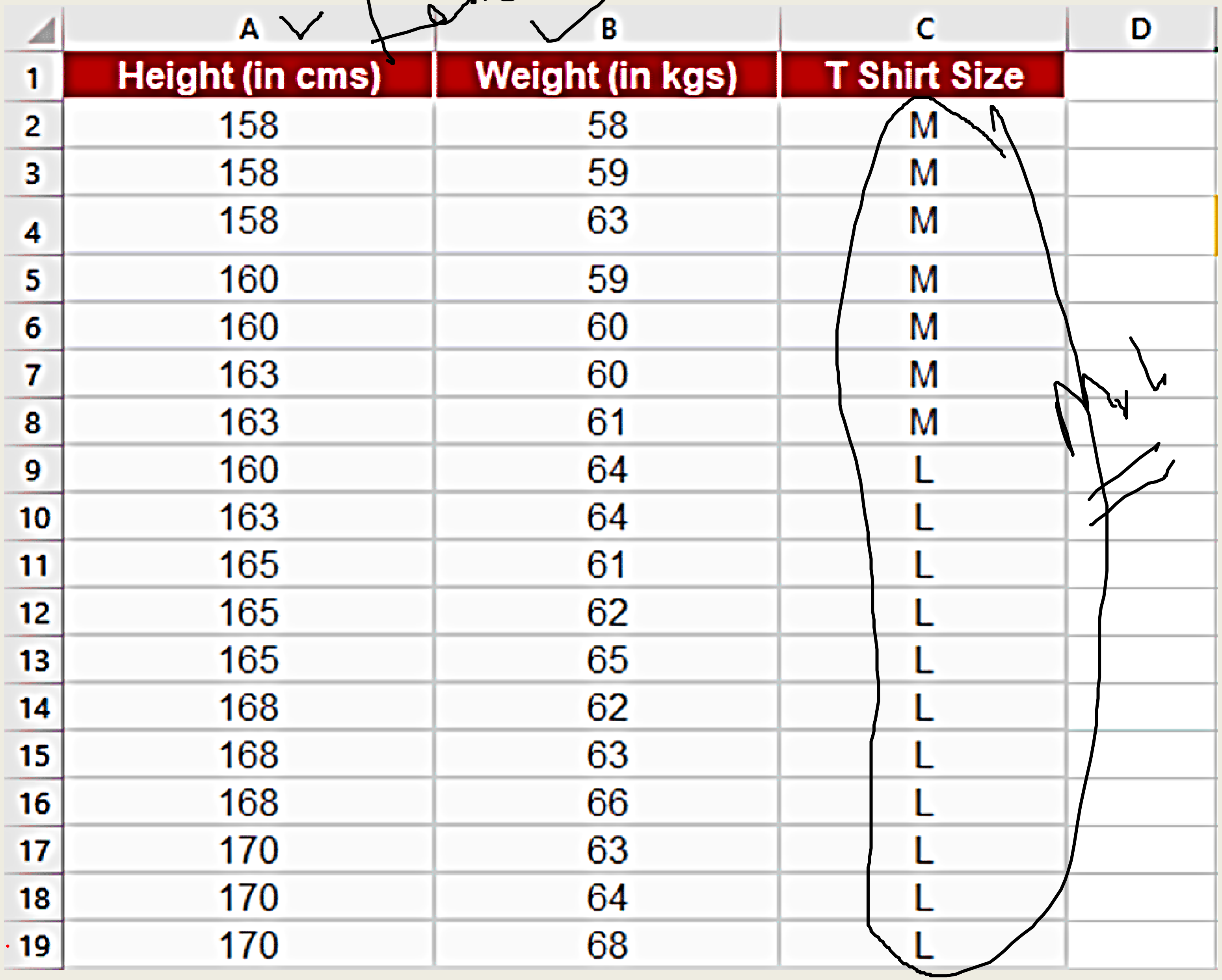
H	W	T.S

dependent

D M

features values

features



	A	B	C	D
1	Height (in cms)	Weight (in kgs)	T Shirt Size	
2	158	58	M	
3	158	59	M	
4	158	63	M	
5	160	59	M	
6	160	60	M	
7	163	60	M	
8	163	61	M	
9	160	64	L	
10	163	64	L	
11	165	61	L	
12	165	62	L	
13	165	65	L	
14	168	62	L	
15	168	63	L	
16	168	66	L	
17	170	63	L	
18	170	64	L	
19	170	68	L	

M/L

Binary classification

	Height (in cms)	Weight (in kgs)	T Shirt Size	Distance
1.	158	58	M	4.24
	158	59	M	✓
	158	63	M	✓
	160	59	M	✓
	160	60	M	✓
	163	60	M	✓
	163	61	M	✓
	160	64	L	✓
	163	64	L	✓
	165	61	L	✓
	165	62	L	✓
	165	65	L	✓
	168	62	L	✓
	168	63	L	✓
	168	66	L	✓
	170	63	L	✓
	170	64	L	✓
	170	68	L	✓

New data

(161, 61)

$$E.D: \sqrt{(161-158)^2 + (61-58)^2}$$

$$= \sqrt{3^2 + 3^2}$$

$$= \sqrt{18}$$

$$= \boxed{4.24}$$

Height (in cms)	Weight (in kgs)	T Shirt Size	Distance
158	58	M	4.24
158	59	M ✓	
158	63	M ✓	
160	59	M ✓	
160	60	M ✓	
163	60	M	
163	61	M	
160	64	L	
163	64	L	
165	61	L	
165	62	L	
165	65	L	
168	62	L	
168	63	L	
168	66	L	
170	63	L	
170	64	L	
170	68	L ✓	

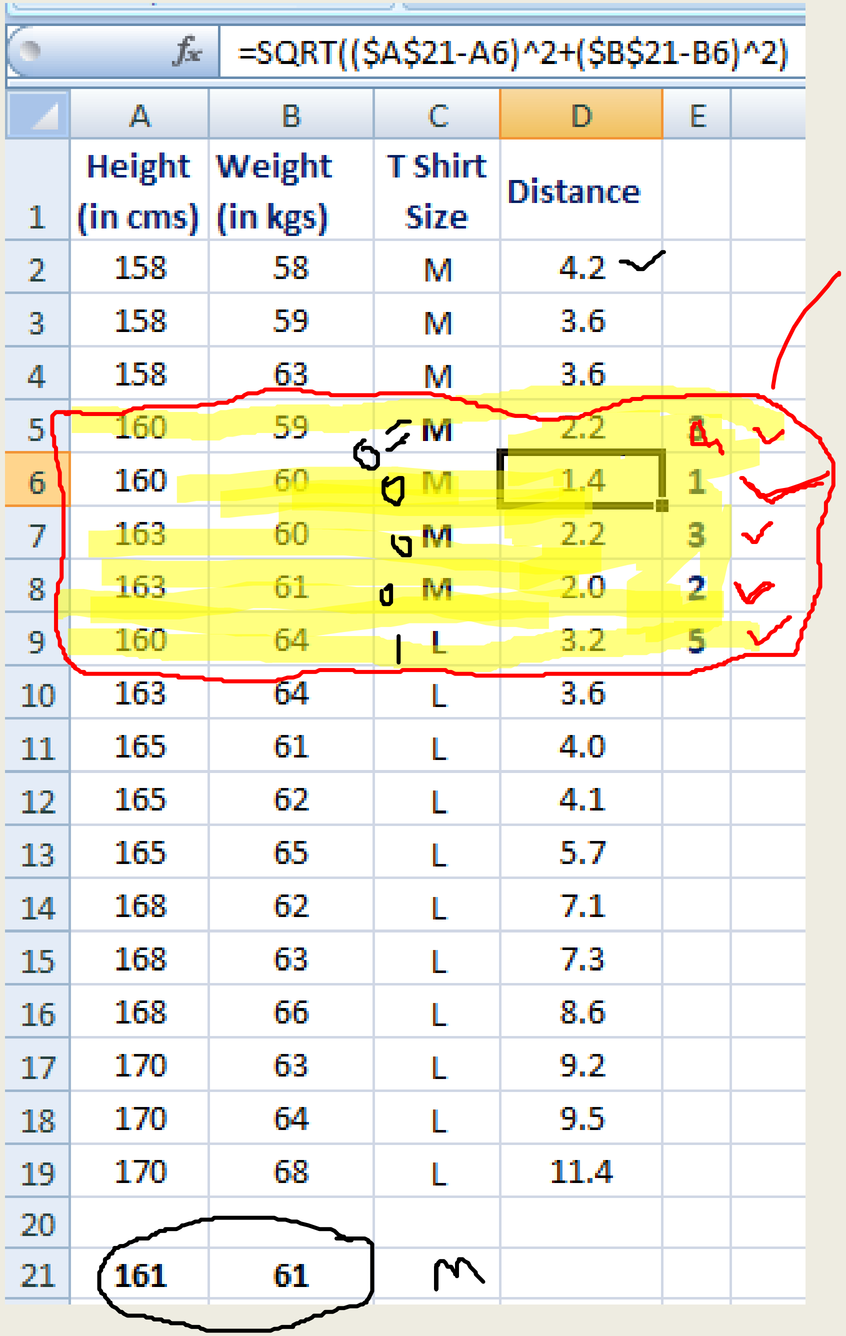
New data point

$(\underbrace{161}_{H}, \underbrace{61}_{W})$

$$= \sqrt{(161-158)^2 + (61-58)^2}$$

$$= \sqrt{3^2 + 3^2} = \sqrt{18}$$

$$= 4.2426$$



=SQRT((A21-A6)^2+(B21-B6)^2)

	A	B	C	D	E
1	Height (in cms)	Weight (in kgs)	T Shirt Size	Distance	
2	158	58	M	4.2 ✓	
3	158	59	M	3.6	
4	158	63	M	3.6	
5	160	59	0 = M	2.2	4 ✓
6	160	60	0 M	1.4	1 ✓
7	163	60	0 M	2.2	3 ✓
8	163	61	0 M	2.0	2 ✓
9	160	64	1 L	3.2	5 ✓
10	163	64	L	3.6	
11	165	61	L	4.0	
12	165	62	L	4.1	
13	165	65	L	5.7	
14	168	62	L	7.1	
15	168	63	L	7.3	
16	168	66	L	8.6	
17	170	63	L	9.2	
18	170	64	L	9.5	
19	170	68	L	11.4	
20					
21	161	61	M		

ordered collection of min distances

K = 5

M = 0, 1

L = 1, 0

0 = 4

1 = 1

4 = M ✓

1 = L ✓

4 > 1

M > L

TS = M

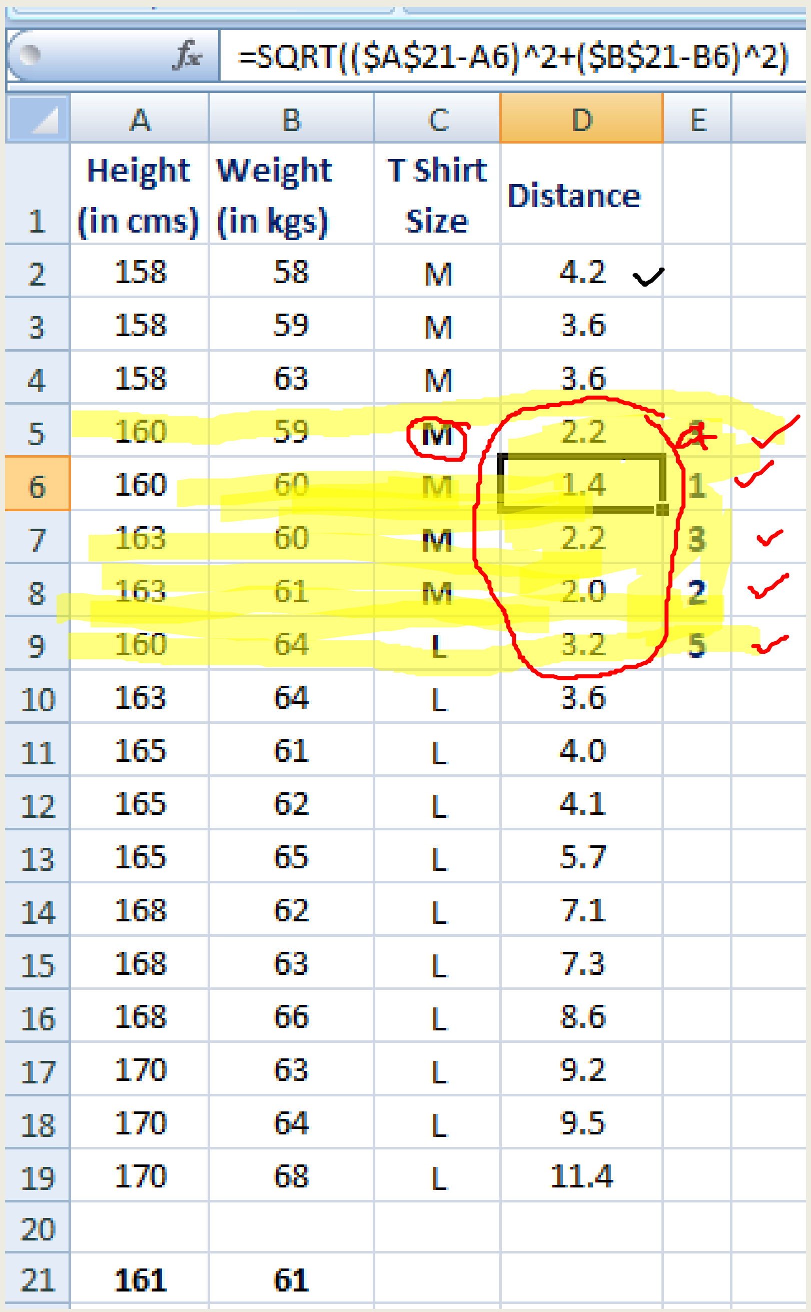
0 L

M

→ Regression → mean

→ classification → mode

$$\frac{(m+m+m+m+L)}{5}$$



=SQRT((A21-A6)^2+(B21-B6)^2)

	A	B	C	D	E	
1	Height (in cms)	Weight (in kgs)	T Shirt Size	Distance		
2	158	58	M	4.2	✓	
3	158	59	M	3.6		
4	158	63	M	3.6		
5	160	59	M	2.2	4	✓
6	160	60	M	1.4	1	✓
7	163	60	M	2.2	3	✓
8	163	61	M	2.0	2	✓
9	160	64	L	3.2	5	✓
10	163	64	L	3.6		
11	165	61	L	4.0		
12	165	62	L	4.1		
13	165	65	L	5.7		
14	168	62	L	7.1		
15	168	63	L	7.3		
16	168	66	L	8.6		
17	170	63	L	9.2		
18	170	64	L	9.5		
19	170	68	L	11.4		
20						
21	161	61				

why 5 ?

k = 5

new

(161, 61)

M

Regression = mean

classification = mode → Distance

H = ?
W = ?

2.2
→ M

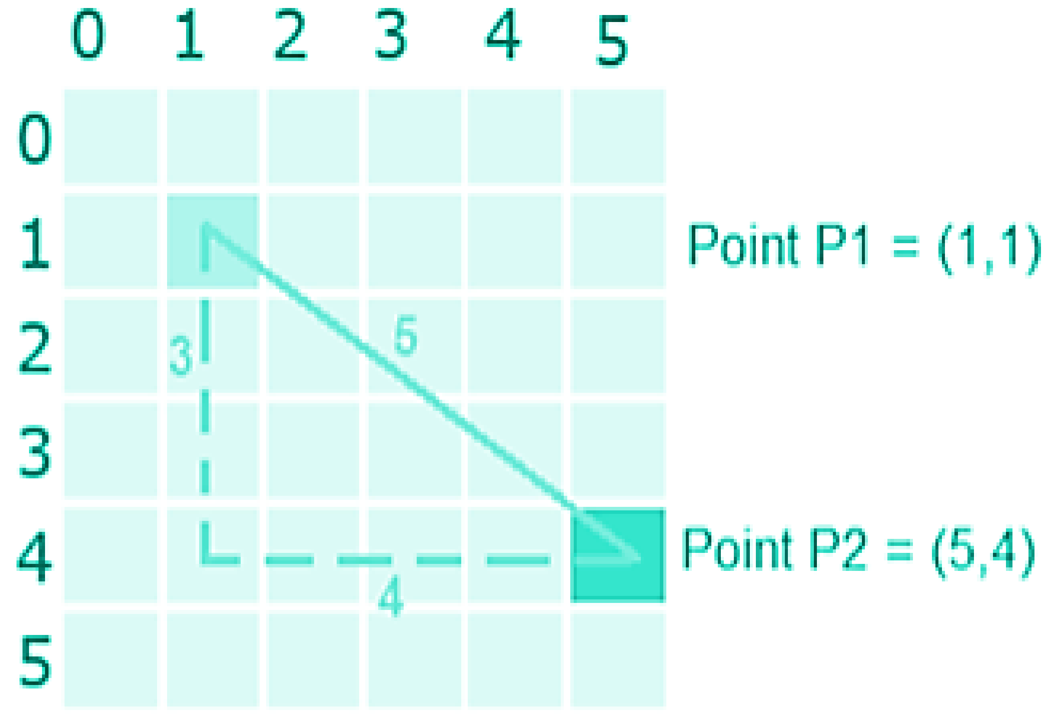
M = 4 ✓
L = 1
→ M > L
4 > 1

- Step1: Calculate the Euclidean distance between the new point and the existing points

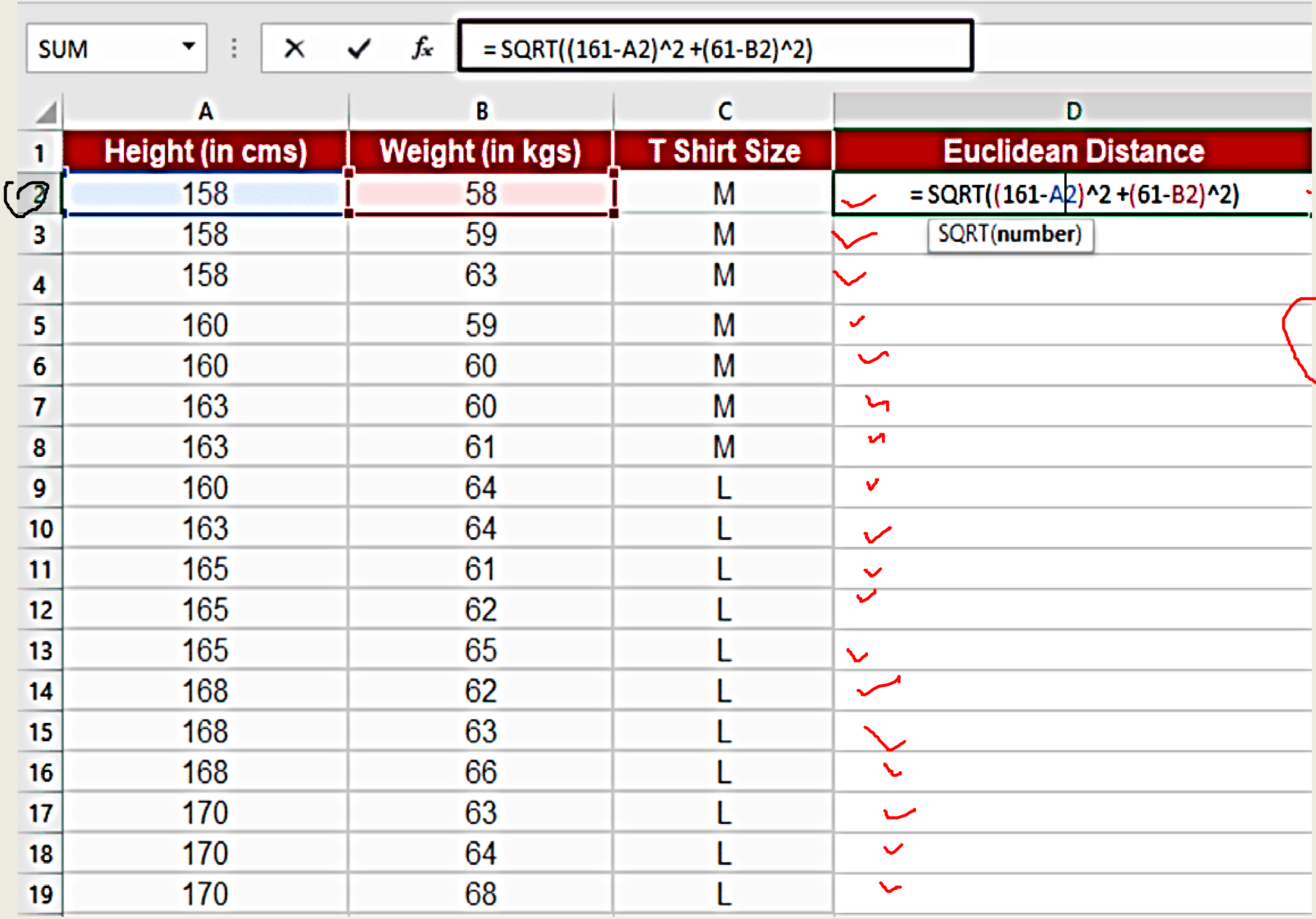
- For example, Euclidean distance between point P1(1,1) and P2(5,4) is:

Euclidean :

$$d(x,y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$



Euclidean distance = $\sqrt{\underline{(5-1)}^2+\underline{(4-1)}^2} = \boxed{5}$



= SQRT((161-A2)^2 +(61-B2)^2)

	A	B	C	D
1	Height (in cms)	Weight (in kgs)	T Shirt Size	Euclidean Distance
2	158	58	M	✓ = SQRT((161-A2)^2 +(61-B2)^2)
3	158	59	M	✓ SQRT(number)
4	158	63	M	✓
5	160	59	M	✓
6	160	60	M	✓
7	163	60	M	✓
8	163	61	M	✓
9	160	64	L	✓
10	163	64	L	✓
11	165	61	L	✓
12	165	62	L	✓
13	165	65	L	✓
14	168	62	L	✓
15	168	63	L	✓
16	168	66	L	✓
17	170	63	L	✓
18	170	64	L	✓
19	170	68	L	✓

$= \sqrt{(161-158)^2 + (61-58)^2}$

$\boxed{4.24}$

$= 3^2 + 3^2 = \sqrt{9+9}$

$\boxed{18}$?

(158, 58)

E.D. ?

new (161, 61)

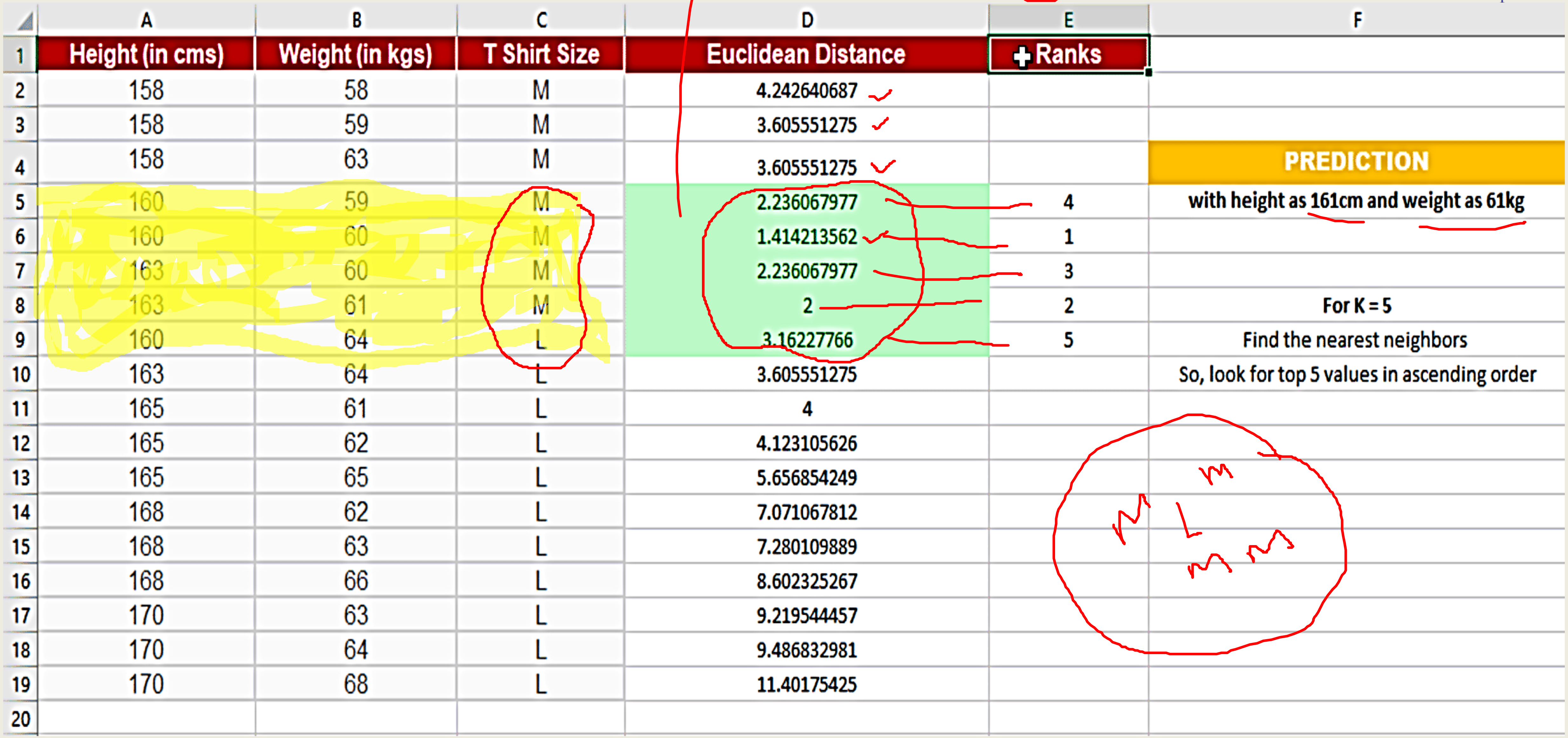
= 5

- Step 2: Choose the value of K and select K neighbors' closet to the new point.
- In this case, select the top 5 parameters having least Euclidean distance

K=1,3,5,7,9 n — m=4/2
L=1 (3)

K=(5) 6

M=3
L=3

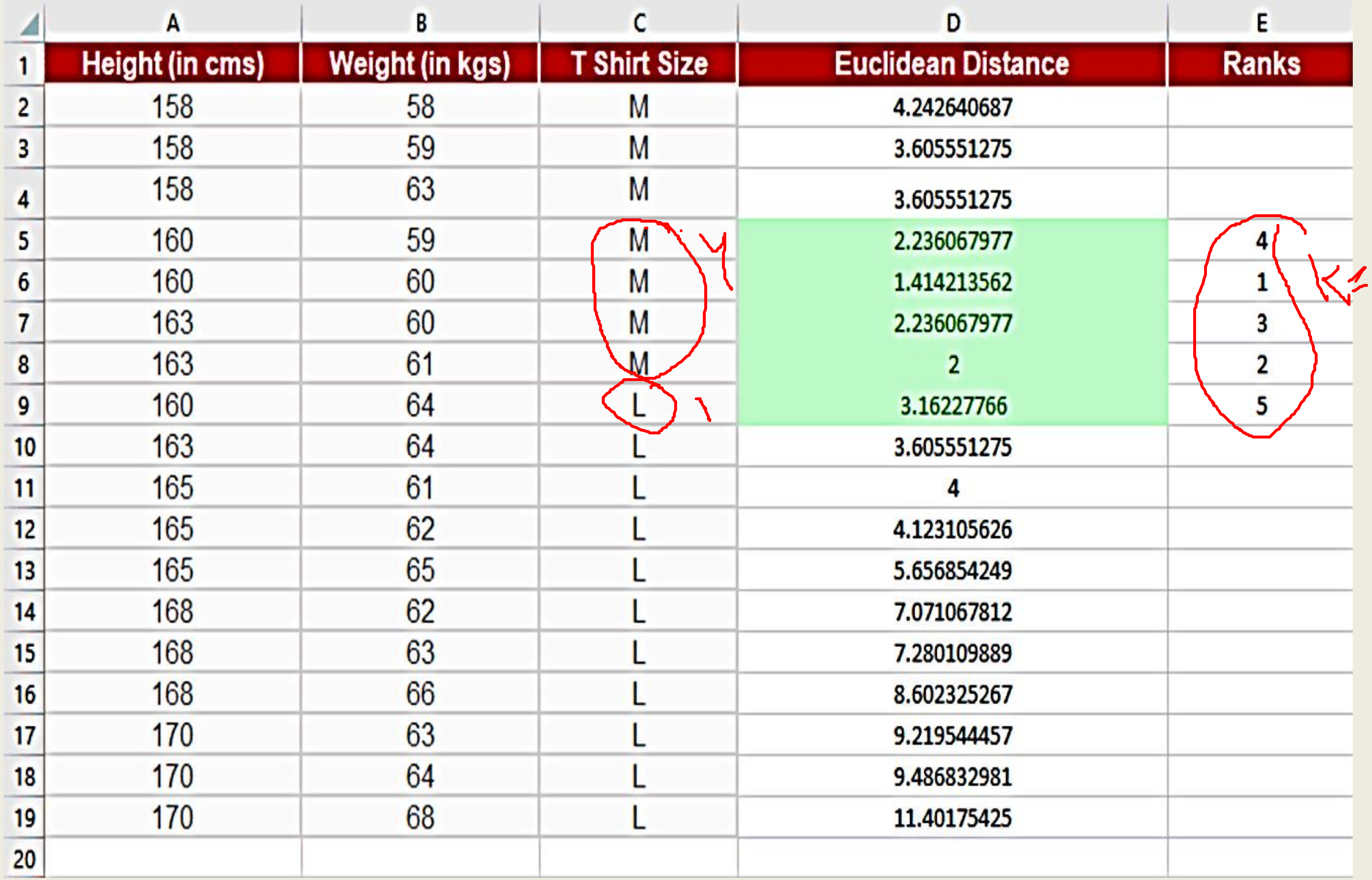


	A	B	C	D	E	F
1	Height (in cms)	Weight (in kgs)	T Shirt Size	Euclidean Distance	Ranks	
2	158	58	M	4.242640687 ✓		
3	158	59	M	3.605551275 ✓		
4	158	63	M	3.605551275 ✓		PREDICTION
5	160	59	M	2.236067977	4	with height as 161cm and weight as 61kg
6	160	60	M	1.414213562 ✓	1	
7	163	60	M	2.236067977	3	
8	163	61	M	2	2	For K = 5
9	160	64	L	3.16227766	5	Find the nearest neighbors
10	163	64	L	3.605551275		So, look for top 5 values in ascending order
11	165	61	L	4		
12	165	62	L	4.123105626		
13	165	65	L	5.656854249		
14	168	62	L	7.071067812		
15	168	63	L	7.280109889		
16	168	66	L	8.602325267		
17	170	63	L	9.219544457		
18	170	64	L	9.486832981		
19	170	68	L	11.40175425		
20						

m
M L
m m

- Step 3: Count the votes of all the K neighbors / Predicting Values
- Since for K = 5, we have 4 T-shirts of size M, therefore according to the k-NN Algorithm, Virendra of height 161 cm and weight, 61kg will fit into a T-shirt of size M.

always choose odd values K=1, 3, 5, 7, 9, 11, 13 – —



	A	B	C	D	E
1	Height (in cms)	Weight (in kgs)	T Shirt Size	Euclidean Distance	Ranks
2	158	58	M	4.242640687	
3	158	59	M	3.605551275	
4	158	63	M	3.605551275	
5	160	59	M	2.236067977	4
6	160	60	M	1.414213562	1
7	163	60	M	2.236067977	3
8	163	61	M	2	2
9	160	64	L	3.16227766	5
10	163	64	L	3.605551275	
11	165	61	L	4	
12	165	62	L	4.123105626	
13	165	65	L	5.656854249	
14	168	62	L	7.071067812	
15	168	63	L	7.280109889	
16	168	66	L	8.602325267	
17	170	63	L	9.219544457	
18	170	64	L	9.486832981	
19	170	68	L	11.40175425	
20					

4

1

K=5

suppose

K=6

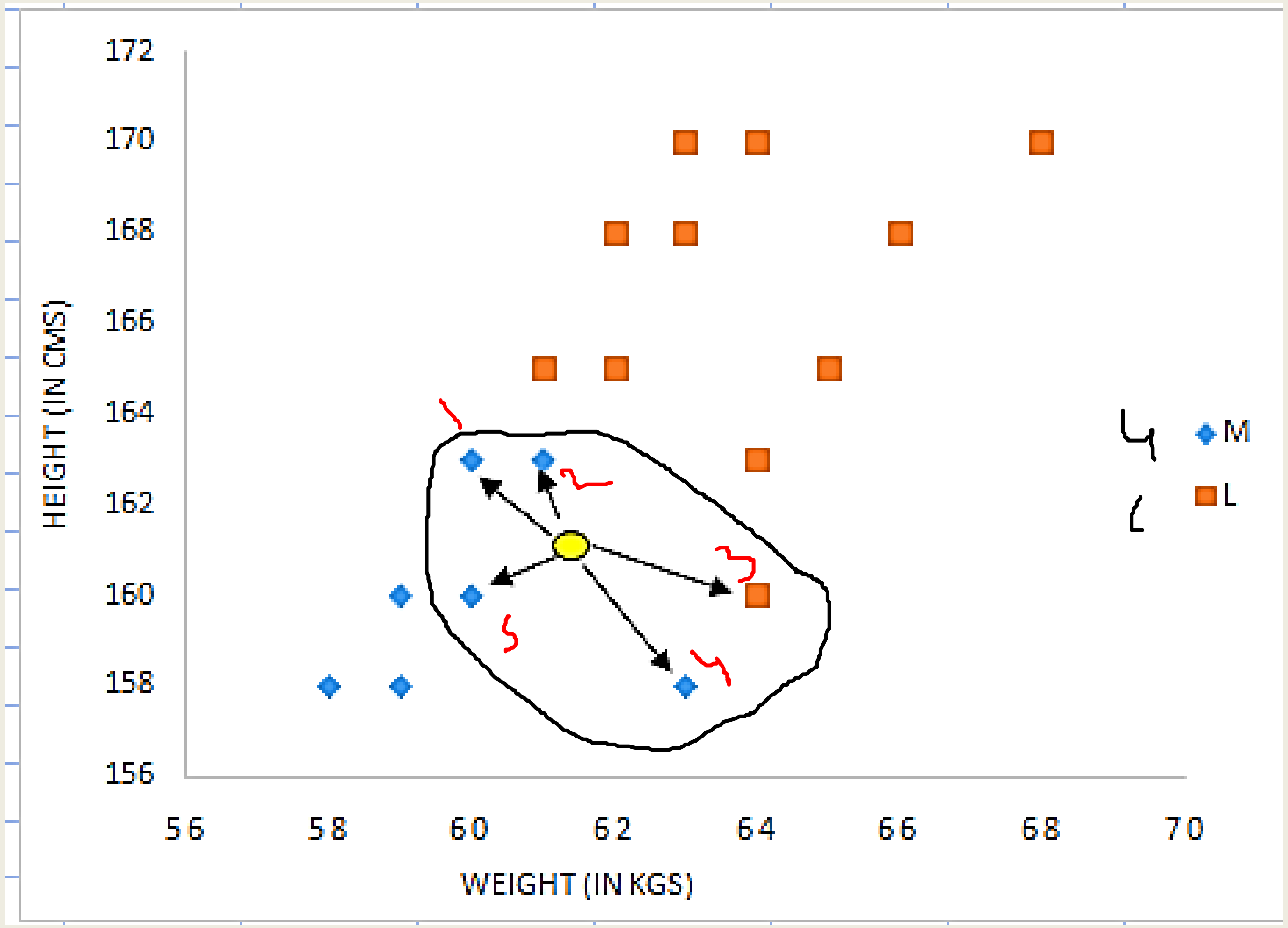
M=3 ✓

L=3 ✓

?

what T-shirt size?

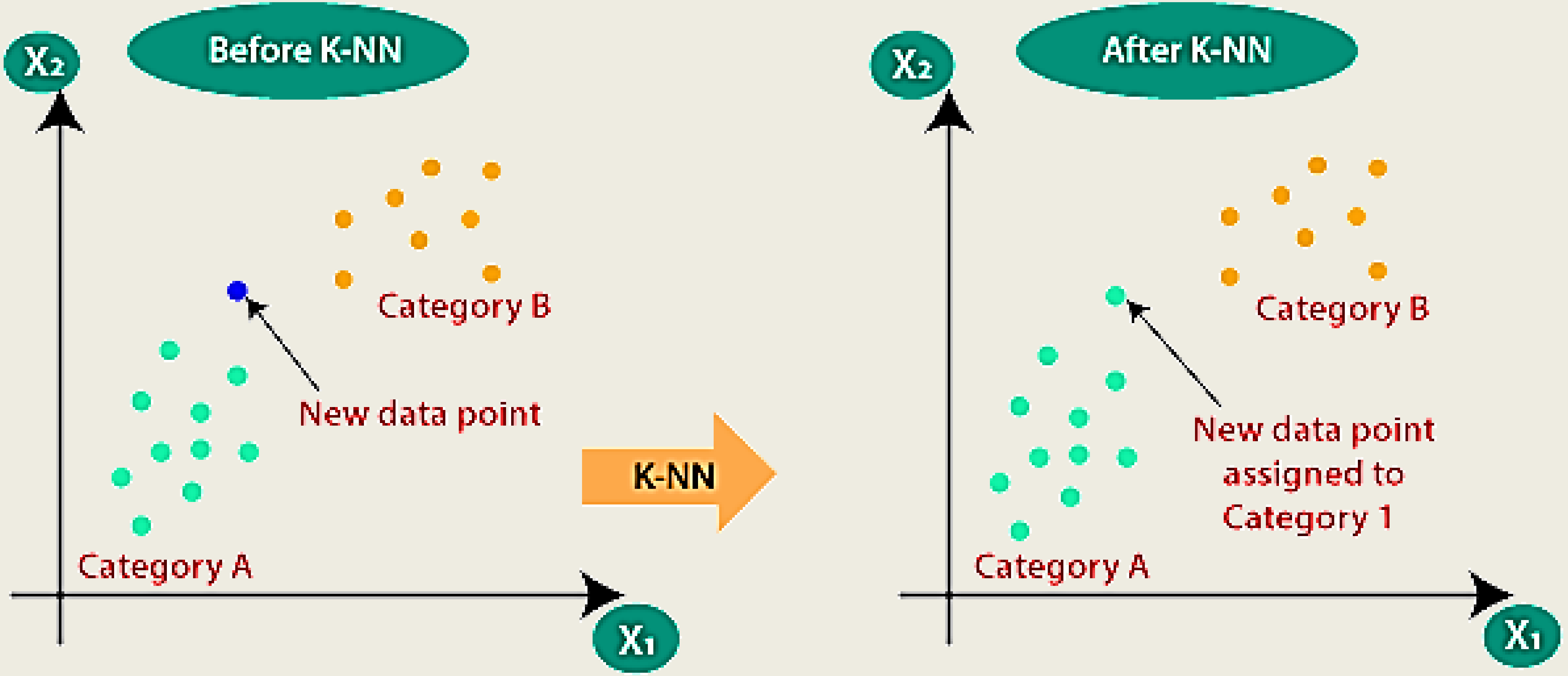
- In the graph, binary dependent variable (T-shirt size) is displayed in blue and orange color. 'Medium T-shirt size' is in blue color and 'Large T-shirt size' in orange color. New customer information is exhibited in yellow circle. Four blue highlighted data points and one orange highlighted data point are close to yellow circle. so the prediction for the new case is blue highlighted data point which is Medium T-shirt size.

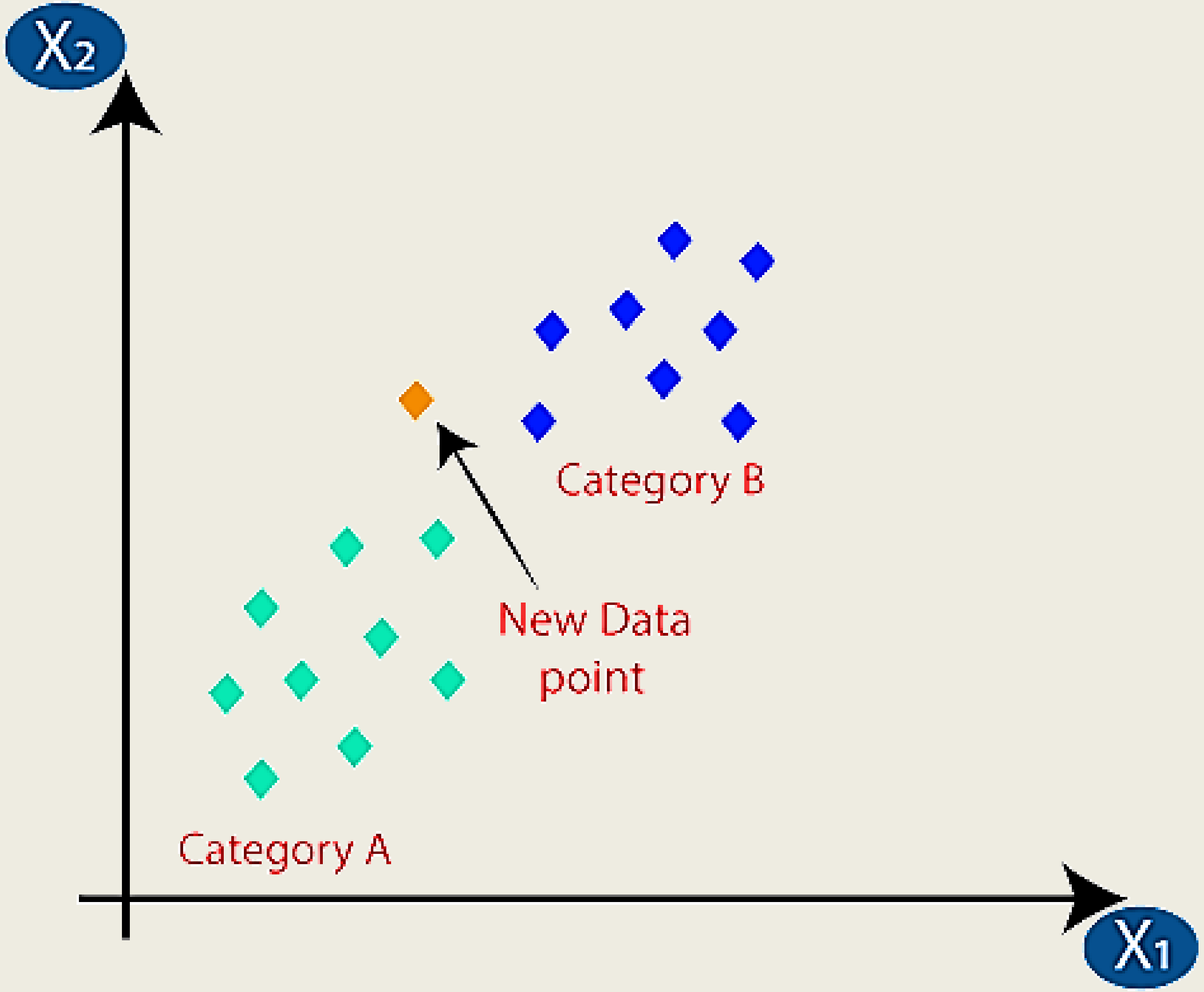


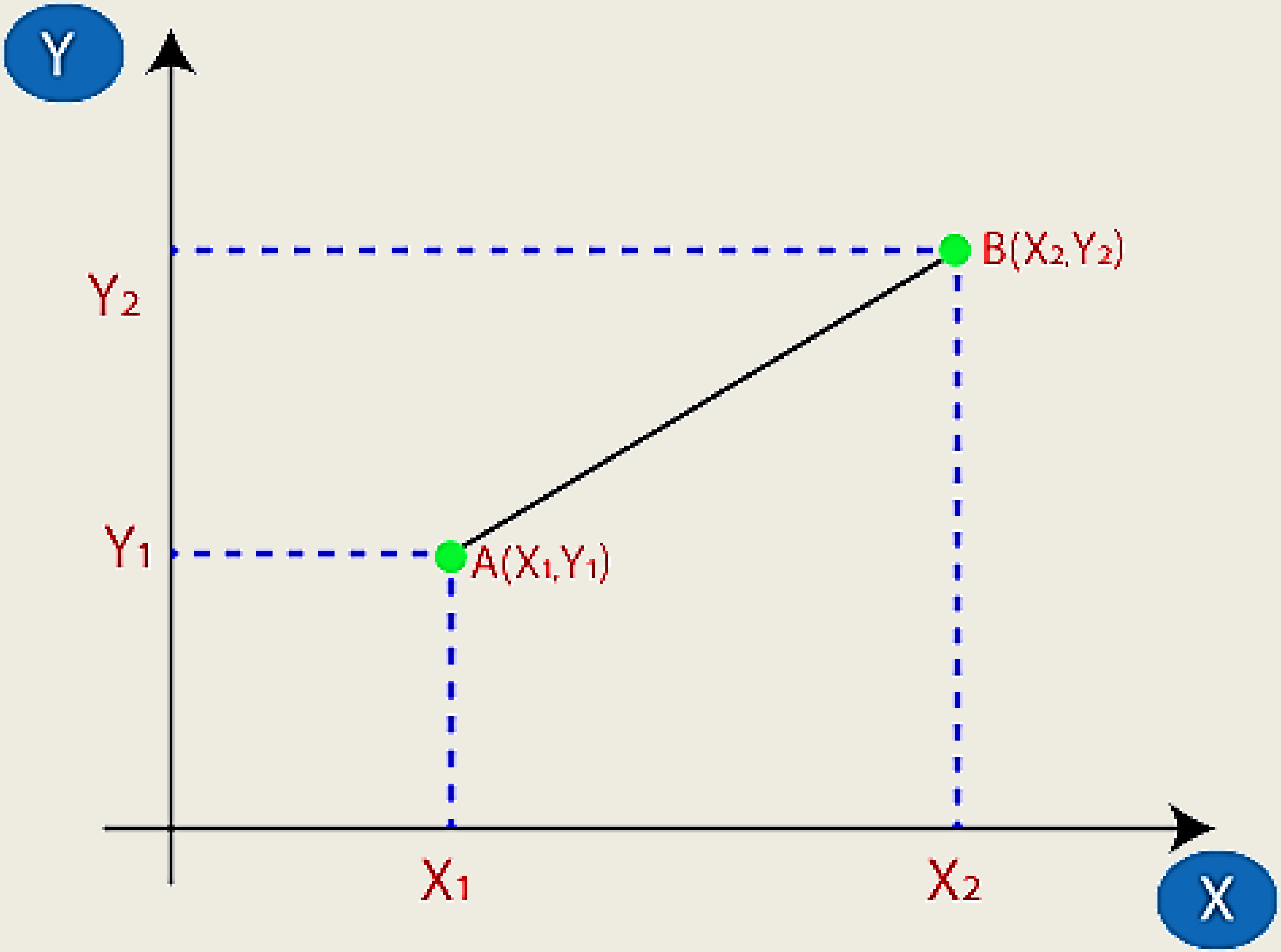
KNN Algorithm Pseudocode

1. Calculate D (x, xi), where 'i' =1, 2, ….., n and 'D' is the Euclidean measure between the data points.
2. The calculated Euclidean distances must be arranged in ascending order.
3. Initialize k and take the first k distances from the sorted list.
4. Figure out the k points for the respective k distances.
5. Calculate ki, which indicates the number of data points belonging to the ith class among k points i.e. k ≥ 0
6. If ki >kj ∀ i ≠ j; put x in class i.

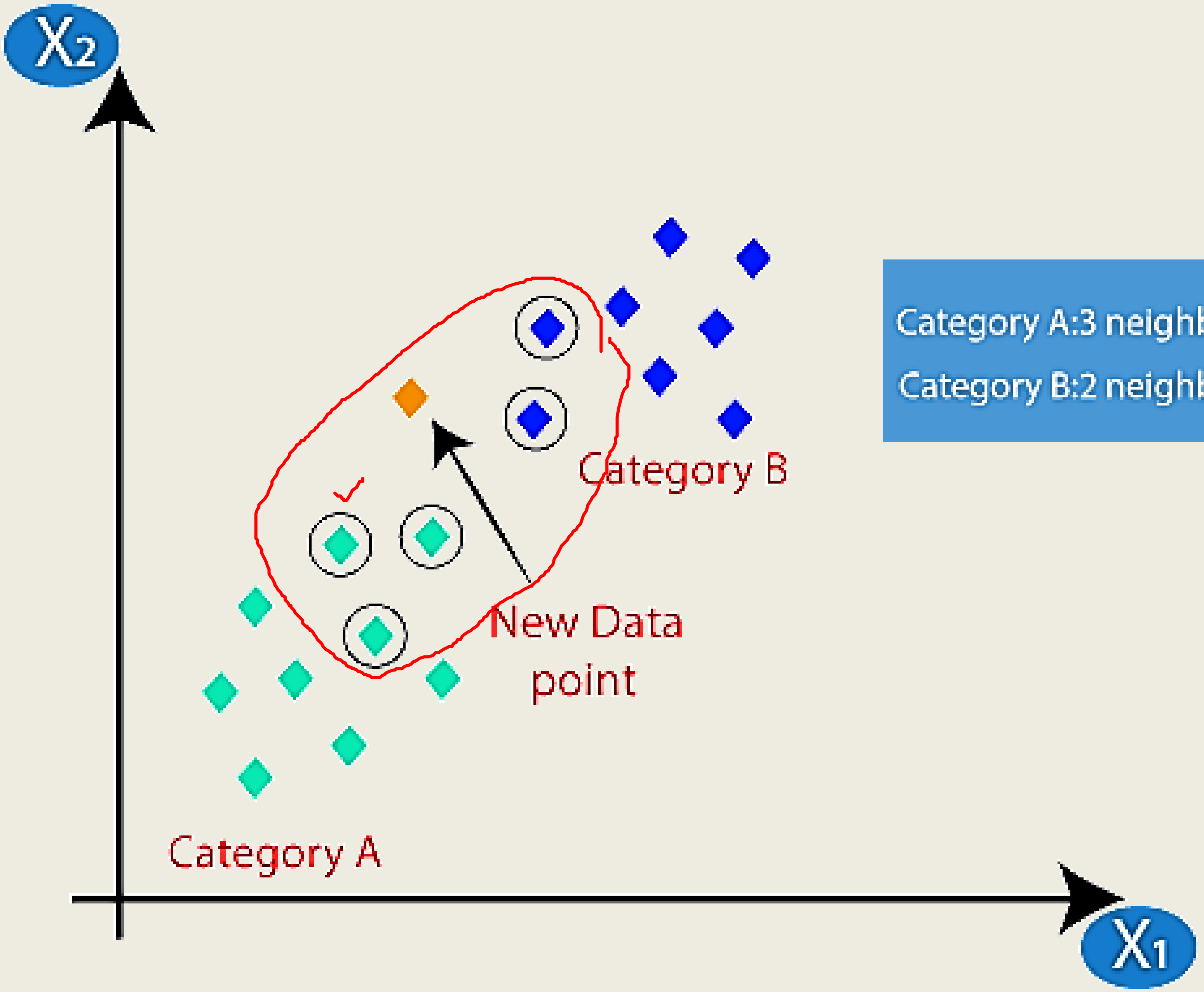
Example


Before K-NN
X₂
X₁
Category B
New data point
Category A
K-NN
After K-NN
X₂
X₁
Category B
New data point assigned to Category 1
Category A


X2
Category B
New Data point
Category A
X1



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



Category A:3 neighbors ✓

Category B:2 neighbors ✓

K = 7 79
K = 5 83
K = 9 92
K = 7
100 %

Pros and Cons

- Easy to use. ✓
- Quick calculation time. ✓ Ep
- Does not make assumptions about the data.

Training

Actual data

K = 5, 7, 9

K = 7

- Accuracy depends on the quality of the data.
- 'Must find an optimal k value (number of nearest neighbors).
- Poor at classifying data points in a boundary where they can be classified one way or another.

K = 6

3 = m
3 = L

Where to use KNN

- KNN is often used in simple recommendation systems, image recognition technology, and decision-making models. It is the algorithm companies like Netflix or Amazon use in order to recommend different movies to watch or books to buy.

- KNN is often used in simple recommendation systems, image recognition technology, and decision-making models. It is the algorithm companies like Netflix or Amazon use in order to recommend different movies to watch or books to buy.