

A decorative graphic on the left side of the slide, consisting of a network of white lines and small circles on a blue gradient background, resembling a circuit board or data flow diagram.

05. CATEGORY PREDICTION

The background is a dark blue gradient. In the corners, there are white line-art illustrations of circuit boards or neural network connections, consisting of lines and small circles.

WHAT IS A CATEGORY PREDICTION?

CATEGORY PREDICTION

- Is the prediction of information that is not numerical from another set of related data
- For example:
 - Will it rain or not
 - Favorite food
 - Favorite drink
 - Music type

CATEGORY PREDICTION

- The main prediction concept relies on using historical data that has been already categorized
- In order to predict a new set of unknown data we can use many different techniques
- In this example we will learn about finding the K-nearest neighbor

Level of fizziness

Level of sweetness

Carbonated drink or juice



K-NEAREST NEIGHBOR

- The main idea is to compare the new unclassified data to the original, categorized data
- The comparison is the data with the least distance
- We compare it with the data that is spaced around the uncategorized data until the K is complete
- Then we use the result to categorize our new information

EXAMPLE 2: PREDICTION OF BEVERAGE TYPE

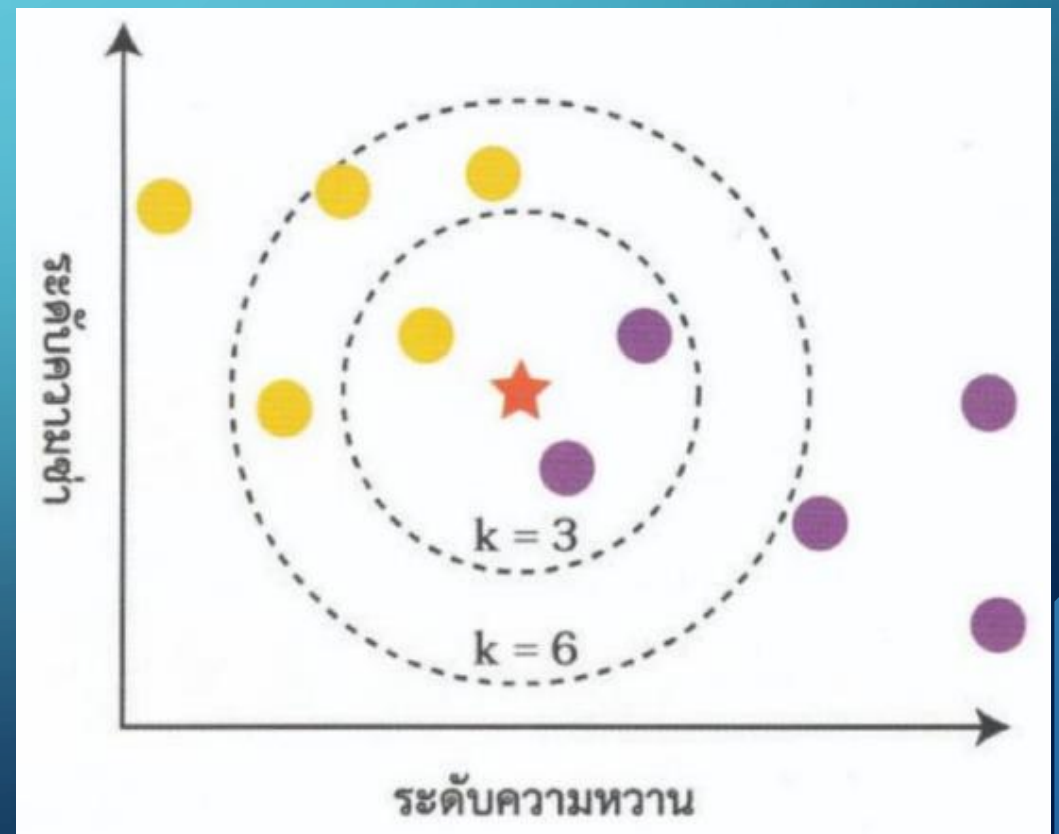
- You are working in a store and there is a new type of drink coming to the store (A)
- You have to organize the chiller/fridge to put the new drink in
- The drink belongs into the same group and has similar properties, it will be either juice or carbonated drink
- Where will you put the drink?

EXAMPLE 2: PREDICTION OF BEVERAGE TYPE

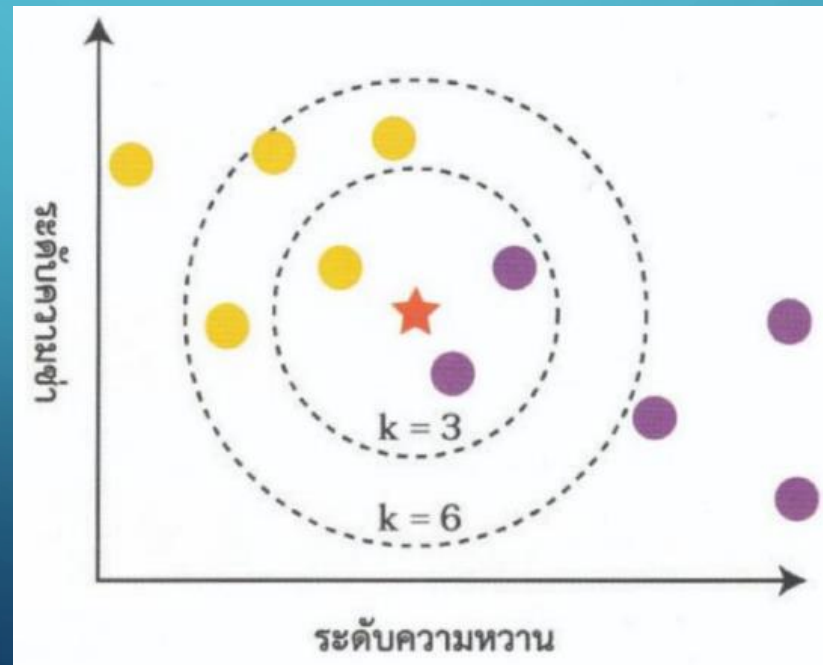
- First we need to consider the main ingredient of the drink
- Depending on if the drink is carbonated or sweet it will belong to one or the other group
- That is why we have to check its neighbors

EXAMPLE 2: PREDICTION OF BEVERAGE TYPE

- To understand the technique of K-NN we will consider this distribution diagram
-  represents carbonated drinks
-  represents sweet drinks
-  represents the new type of drink



Number of points closest to the star (K)	Number of carbonated drinks	Number of juice	Classification result
K = 1	0	1	Juice drink
K = 3	1	2	Juice drink
K = 6	4	2	Carbonated drink

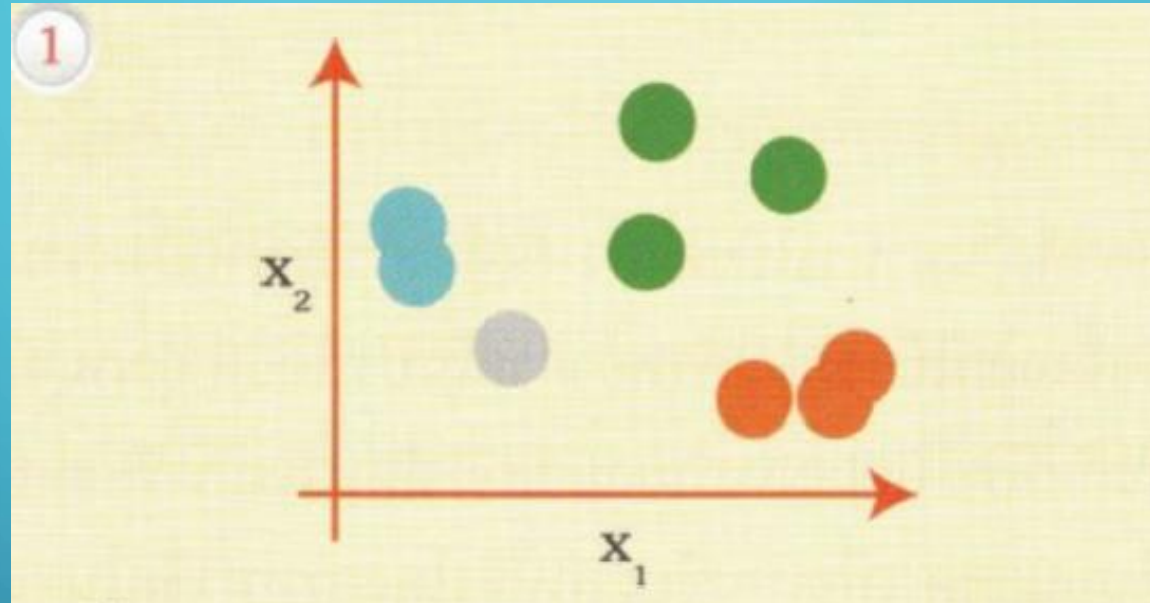


K-NN ALGORITHM

- 1 Set the number of nearest neighbors (K value) ตั้ง 1
- 2 Receive information for the new item to be classified ตั้ง 1
- 3 Loop through each data item ตั้ง 2
 - 3.1 Calculate the distance with new data
 - 3.2 Record the distance value of each item
- 4 Sort a list of data by ascending distance ตั้ง 3
- 5 Select the K data in order of the sorted data ตั้ง 4
- 6 Count each group member in the K-data sequence ตั้ง 5
- 7 Classify the new item in the group that has the most members ตั้ง 5

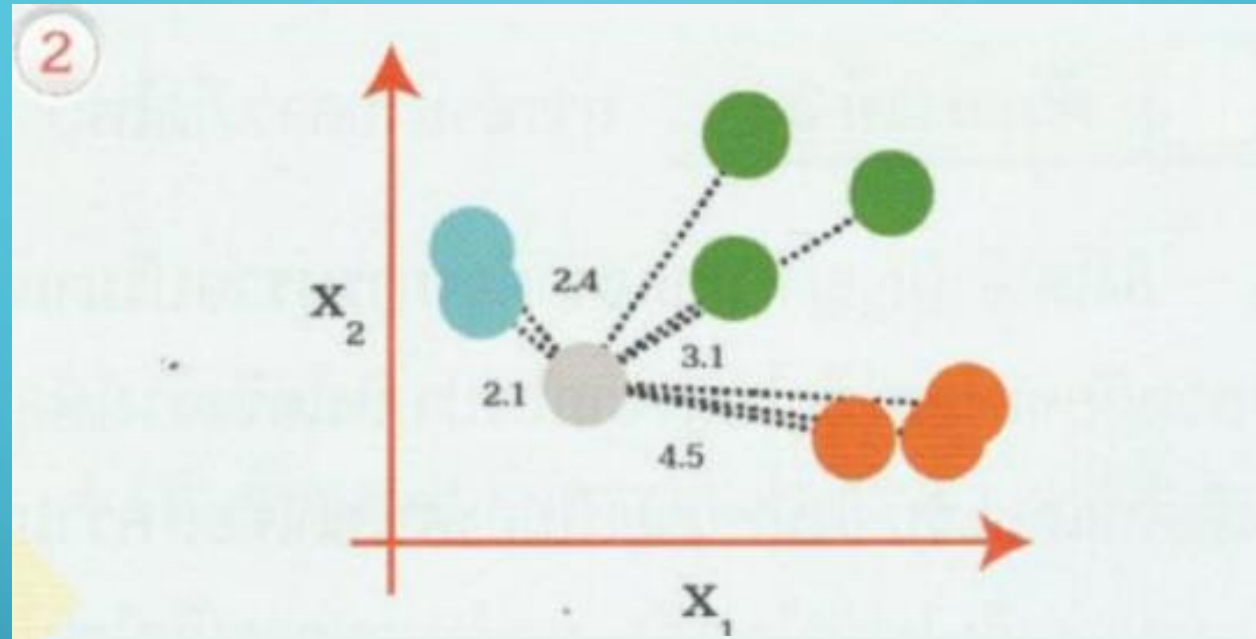
- | | | | |
|---|---|------|---|
| 1 | กำหนดจำนวนเพื่อนบ้านใกล้เคียงที่สุด (ค่า K) | ตั้ง | 1 |
| 2 | รับข้อมูลรายการใหม่ที่ต้องการจำแนก | ตั้ง | 1 |
| 3 | วนซ้ำในแต่ละรายการข้อมูล | ตั้ง | 2 |
| | 3.1 คำนวณระยะทาง (distance) กับข้อมูลรายการใหม่ | | |
| | 3.2 บันทึกค่าระยะทางของแต่ละรายการ | | |
| 4 | จัดเรียงรายการข้อมูลตามระยะทางจากน้อยไปมาก | ตั้ง | 3 |
| 5 | เลือกข้อมูล K ลำดับในข้อมูลที่จัดเรียงแล้ว | ตั้ง | 4 |
| 6 | นับสมาชิกแต่ละกลุ่มในข้อมูล K ลำดับ | ตั้ง | 5 |
| 7 | จำแนกข้อมูลรายการใหม่ ให้อยู่ในกลุ่มที่มีจำนวนสมาชิกมากที่สุด | ตั้ง | 5 |

STEP 1:



Step 1-2: Set the value of $K = 3$, and the gray point will be representing new data that has to be categorized into one of three groups (blue, green, orange)

STEP 2:



Step 3: Calculate the distance between the gray point and other objects and record their distance

STEP 3.

3

ระยะทางระหว่างจุด









	- -		2.1	→	K ลำดับที่ 1
	- -		2.4	→	K ลำดับที่ 2
	- -		3.1	→	K ลำดับที่ 3
	- -		4.5	→	K ลำดับที่ 4

Step 4: Sort the data list by ascending distance

STEP 4

4

ระยะทางระหว่างจุด

	- -		2.1	→	K ลำดับที่ 1
	- -		2.4	→	K ลำดับที่ 2
	- -		3.1	→	K ลำดับที่ 3
	- -		4.5	→	K ลำดับที่ 4

Step 5: Select three values that have the lowest distance

STEP 5:

5

กลุ่ม	จำนวน โหนด
	2
	1



กลุ่ม  มีจำนวนสมาชิก
มากกว่า ดังนั้นจะจำแนก 
ให้อยู่ในกลุ่ม 

Step 6-7: Count the members of each group in the first 3 datasets. We can see that 2/3 of data is in the blue group while 1/3 belongs to the green group. Therefore, our data will be classified as blue category.

FOREST FIRE PREDICTION

- Open this link:

bit.ly/CatPredYEP

FOREST FIRE PREDICTION

- We want to know if there will be a forest fire when the temperature is 19.7°C and the relative humidity is 28%
- How can we do it?

STEP 1

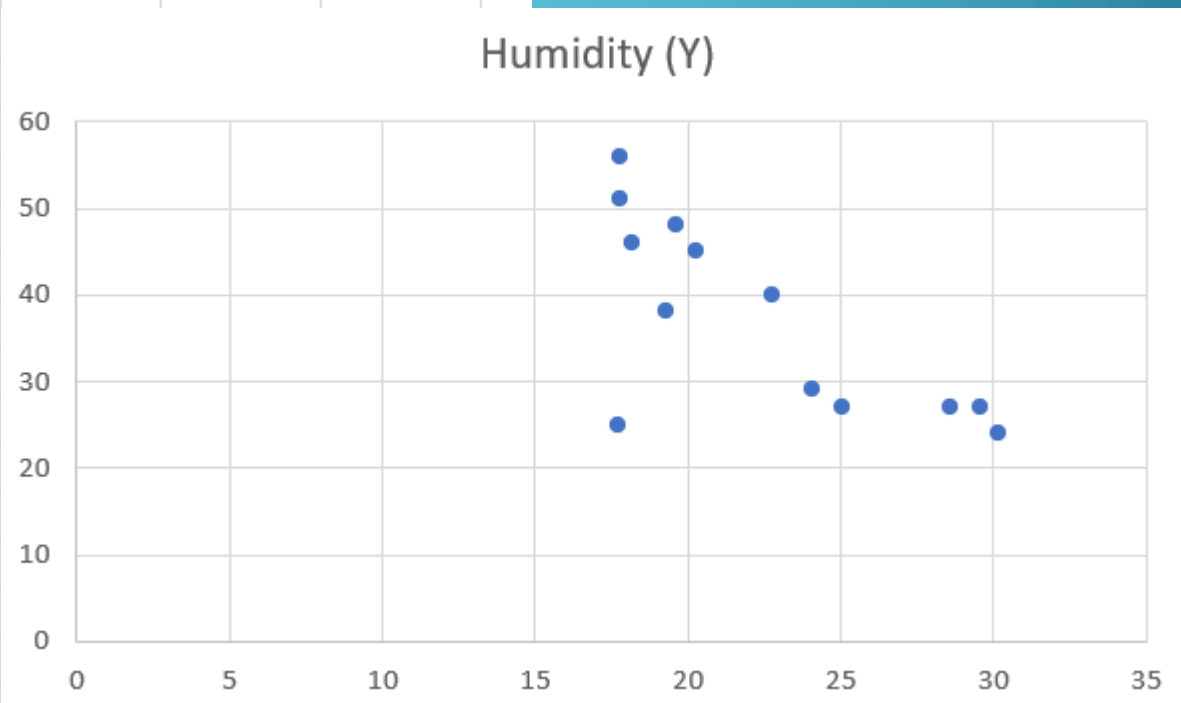
- Create a scatter chart

Forest Fire Prediction			
No	Temp. (X)	Humidity (Y)	Fire (Class)
1	22.8	40	No
2	29.6	27	Yes
3	17.8	51	No
4	19.3	38	No
5	28.6	27	Yes
6	17.8	56	Yes
7	17.7	25	Yes
8	18.2	46	Yes
9	25.1	27	No
10	30.2	24	No
11	19.6	48	No
12	24.1	29	No
13	20.3	45	No

STEP 1

- Create a scatter chart

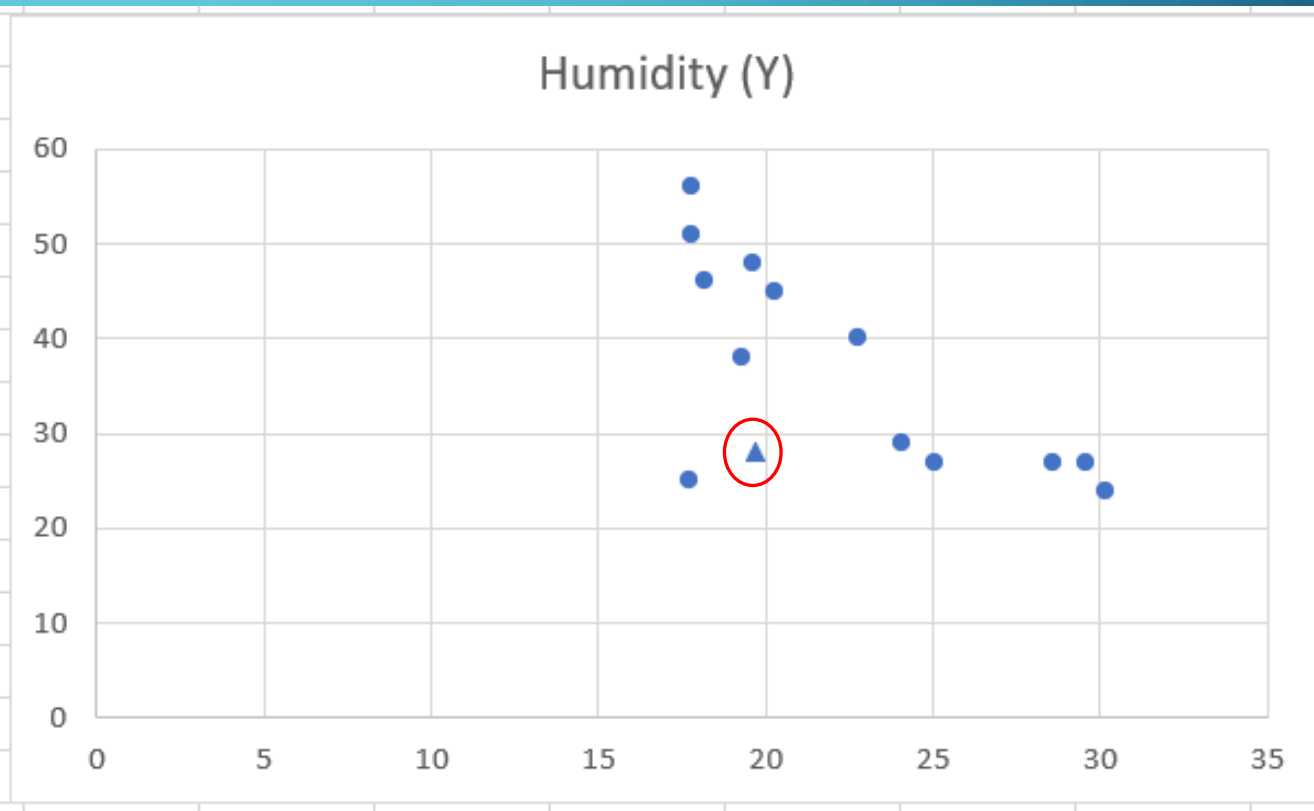
No	Temp. (X)	Humidity (Y)	Fire (Class)
1	22.8	40	No
2	29.6	27	Yes
3	17.8	51	No
4	19.3	38	No
5	28.6	27	Yes
6	17.8	56	Yes
7	17.7	25	Yes
8	18.2	46	Yes
9	25.1	27	No
10	30.2	24	No
11	19.6	48	No
12	24.1	29	No
13	20.3	45	No



STEP 2

- Add the new data to our table and the chart. Change the new data to a triangle in the chart for better visualization

No	Temp. (X)	Humidity (Y)	Fire (Class)
1	22.8	40	No
2	29.6	27	Yes
3	17.8	51	No
4	19.3	38	No
5	28.6	27	Yes
6	17.8	56	Yes
7	17.7	25	Yes
8	18.2	46	Yes
9	25.1	27	Yes
10	30.2	24	No
11	19.6	48	No
12	24.1	29	No
13	20.3	45	No
14	19.7	28	????



STEP 4:

- Calculate the distance between the known data and the new data that has to be categorized

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

STEP 4:

No	Temp. (X)	Humidity (Y)	Fire (Class)
1	22.8	40	No

$$\sqrt{(22.8 - 19.7)^2 + (40 - 28)^2} =$$

STEP 4:

No	Temp. (X)	Humidity (Y)	Fire (Class)
1	22.8	40	No

$$\begin{aligned}\sqrt{(22.8 - 19.7)^2 + (40 - 28)^2} &= \\ \sqrt{9.61 + 144} &= \end{aligned}$$

STEP 4:

No	Temp. (X)	Humidity (Y)	Fire (Class)
1	22.8	40	No

$$\sqrt{(22.8 - 19.7)^2 + (40 - 28)^2} =$$

$$\sqrt{9.61 + 144} =$$

$$\sqrt{153.61} =$$

STEP 4:

No	Temp. (X)	Humidity (Y)	Fire (Class)
1	22.8	40	No

$$\sqrt{(22.8 - 19.7)^2 + (40 - 28)^2} =$$

$$\sqrt{9.61 + 144} =$$

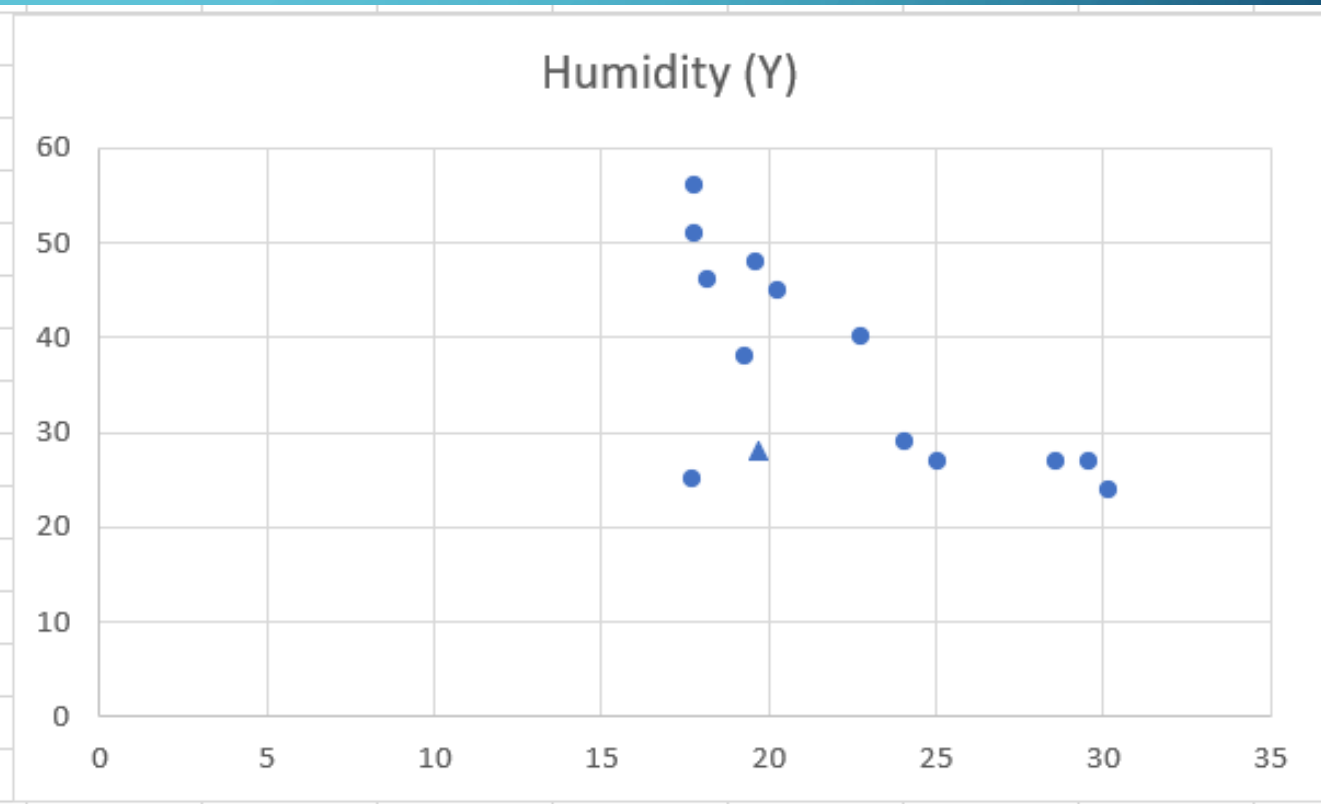
$$\sqrt{153.61} =$$

$$12.39$$

STEP 5

- Calculate the distance for all other items and add it to the list

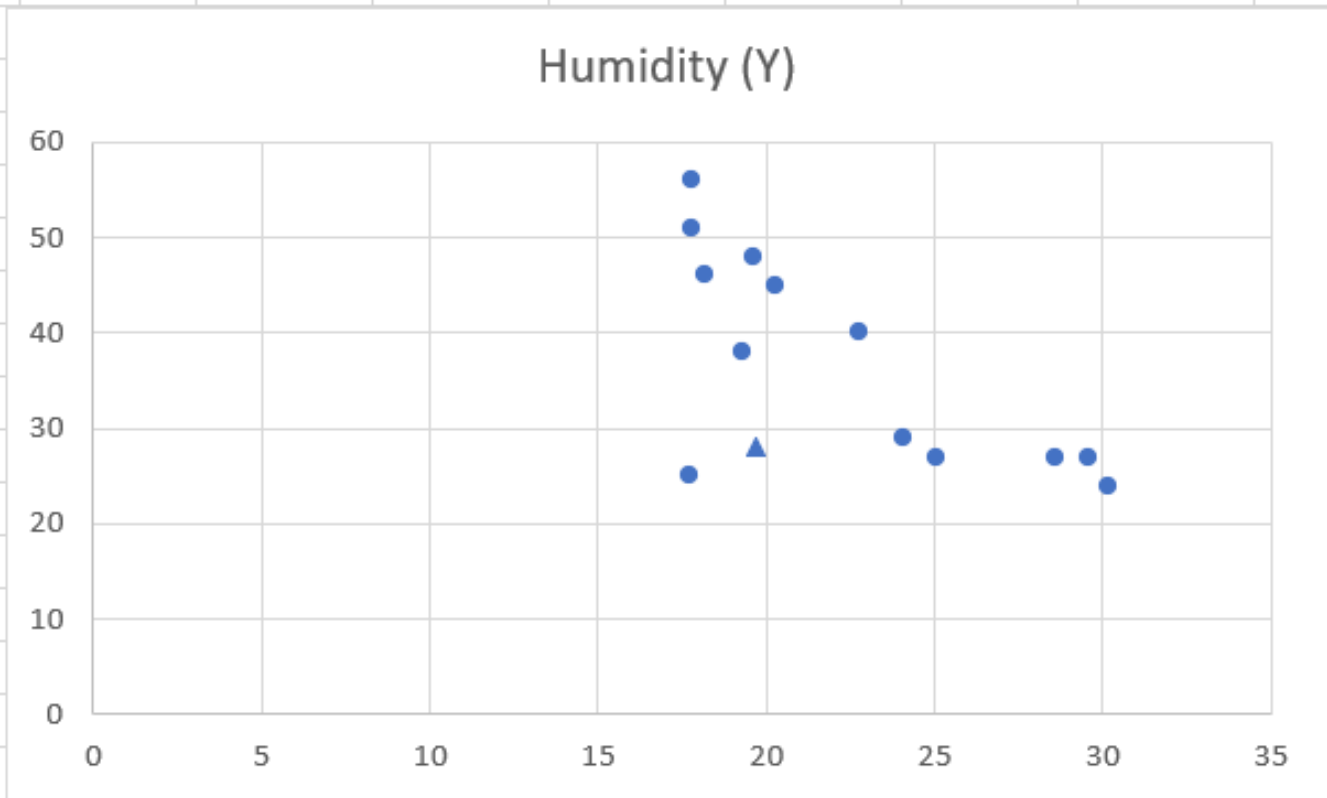
No	Temp. (X)	Humidity (Y)	Fire (Class)	Distance
1	22.8	40	No	12.39
2	29.6	27	Yes	9.95
3	17.8	51	No	
4	19.3	38	No	
5	28.6	27	Yes	
6	17.8	56	Yes	
7	17.7	25	Yes	
8	18.2	46	Yes	
9	25.1	27	Yes	
10	30.2	24	No	
11	19.6	48	No	
12	24.1	29	No	
13	20.3	45	No	
14	19.7	28	????	



STEP 5

- This might take a while...

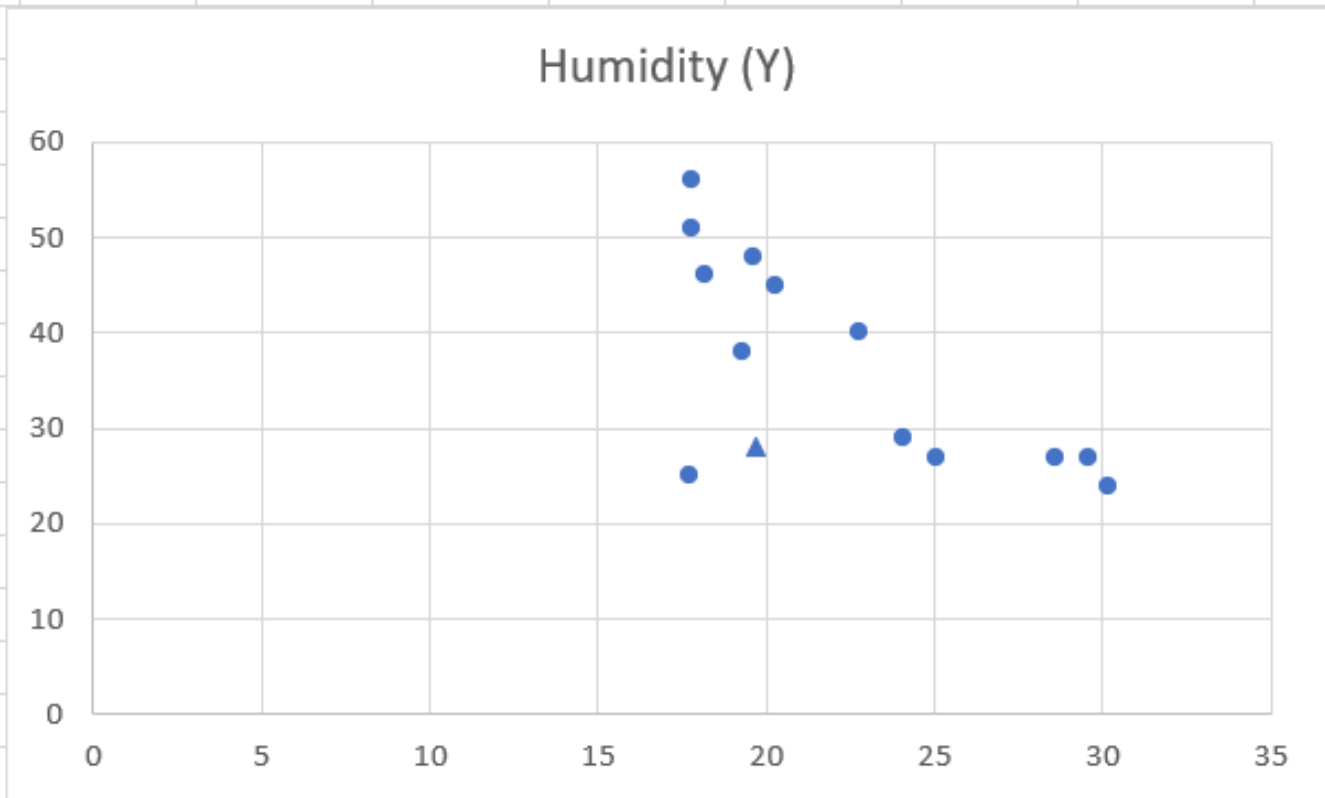
No	Temp. (X)	Humidity (Y)	Fire (Class)	Distance
1	22.8	40	No	12.39
2	29.6	27	Yes	9.95
3	17.8	51	No	
4	19.3	38	No	
5	28.6	27	Yes	
6	17.8	56	Yes	
7	17.7	25	Yes	
8	18.2	46	Yes	
9	25.1	27	Yes	
10	30.2	24	No	
11	19.6	48	No	
12	24.1	29	No	
13	20.3	45	No	
14	19.7	28	????	



STEP 5

- This might take a while...
- Because we are using excel we can easily make the program do the work for US

No	Temp. (X)	Humidity (Y)	Fire (Class)	Distance
1	22.8	40	No	12.39
2	29.6	27	Yes	9.95
3	17.8	51	No	
4	19.3	38	No	
5	28.6	27	Yes	
6	17.8	56	Yes	
7	17.7	25	Yes	
8	18.2	46	Yes	
9	25.1	27	Yes	
10	30.2	24	No	
11	19.6	48	No	
12	24.1	29	No	
13	20.3	45	No	
14	19.7	28	????	



No	Temp. (X)	Humidity (Y)	Fire (Class)	Distance		
1	22.8	40	No	12.39		
2	29.6	27	Yes	9.95	60	
3	17.8	51	=SQRT(POWER(B6-19.7,2)+POWER(C6-28,2))			
4	19.3	38	No		50	
5	28.6	27	Yes			

=SQRT(POWER(B6-19.7,2)+POWER(C6-28,2))

ASSIGN THE EQUATION TO ALL OTHER CELLS

	No	Temp. (X)	Humidity (Y)	Fire (Class)	Distance
4	1	22.8	40	No	12.39
5	2	29.6	27	Yes	9.95
6	3	17.8	51	No	23.07834
7	4	19.3	38	No	
8	5	28.6	27	Yes	
9	6	17.8	56	Yes	
10	7	17.7	25	Yes	
11	8	18.2	46	Yes	
12	9	25.1	27	Yes	
13	10	30.2	24	No	
14	11	19.6	48	No	
15	12	24.1	29	No	
16	13	20.3	45	No	
17	14	19.7	28	????	
18					
19					

Click on this dot and pull it down to the last cell



ASSIGN THE EQUATION TO ALL OTHER CELLS

No	Temp. (X)	Humidity (Y)	Fire (Class)	Distance
1	22.8	40	No	12.39
2	29.6	27	Yes	9.95
3	17.8	51	No	23.07834
4	19.3	38	No	
5	28.6	27	Yes	
6	17.8	56	Yes	
7	17.7	25	Yes	
8	18.2	46	Yes	
9	25.1	27	Yes	
10	30.2	24	No	
11	19.6	48	No	
12	24.1	29	No	
13	20.3	45	No	
14	19.7	28	????	

Click on this dot and pull it down to the last cell

And release



ASSIGN THE EQUATION TO ALL OTHER CELLS

	No	Temp. (X)	Humidity (Y)	Fire (Class)	Distance
3					
4	1	22.8	40	No	12.39
5	2	29.6	27	Yes	9.95
6	3	17.8	51	No	23.07834
7	4	19.3	38	No	10.008
8	5	28.6	27	Yes	8.956004
9	6	17.8	56	Yes	28.06439
0	7	17.7	25	Yes	3.605551
1	8	18.2	46	Yes	18.06239
2	9	25.1	27	Yes	5.491812
3	10	30.2	24	No	11.2361
4	11	19.6	48	No	20.00025
5	12	24.1	29	No	4.512206
6	13	20.3	45	No	17.01058
7	14	19.7	28	????	0
8					

STEP 6

- Sort the list by the distance from the smallest to the highest number

No	Temp. (X)	Humidity (Y)	Fire (Class)	Distance
1	22.8			
2	29.6			
3	17.8			
4	19.3			
5	28.6			
6	17.8			
7	17.7			
8	18.2			
9	25.1			
10	30.2			
11	19.6			
12	24.1			
13	20.3			
14	19.7			

Sort Smallest to Largest

Sort Largest to Smallest

Sort by Color

Sheet View

Clear Filter From "Distance"

Filter by Color

Number Filters

Search

- (Select All)
- 0
- 3.605551275
- 4.512205669
- 5.491812087
- 8.956003573
- 9.95
- 10.0079968
- 11.23610253
- 12.20

OK Cancel

STEP 7

- Our first 3 numbers should be our K nearest neighbors
- If we analyze it carefully we can see to what group our data will belong
- Will there be fire with temperature of 19.7C and relative humidity of 28%?

	No	Temp. (X)	Humidity (Y)	Fire (Class)	Distance
3					
4	14	19.7	28	????	0
5	7	17.7	25	Yes	3.605551
6	12	24.1	29	No	4.512206
7	9	25.1	27	Yes	5.491812
8	5	28.6	27	Yes	8.956004
9	2	29.6	27	Yes	9.95
10	4	19.3	38	No	10.008
11	10	30.2	24	No	11.2361
12	1	22.8	40	No	12.39
13	13	20.3	45	No	17.01058
14	8	18.2	46	Yes	18.06239
15	11	19.6	48	No	20.00025
16	3	17.8	51	No	23.07834
17	6	17.8	56	Yes	28.06439
18					

STEP 7

- What if K was 5?
- Will there be any change in our category?

	No	Temp. (X)	Humidity (Y)	Fire (Class)	Distance
3					
4	14	19.7	28	????	0
5	7	17.7	25	Yes	3.605551
6	12	24.1	29	No	4.512206
7	9	25.1	27	Yes	5.491812
8	5	28.6	27	Yes	8.956004
9	2	29.6	27	Yes	9.95
10	4	19.3	38	No	10.008
11	10	30.2	24	No	11.2361
12	1	22.8	40	No	12.39
13	13	20.3	45	No	17.01058
14	8	18.2	46	Yes	18.06239
15	11	19.6	48	No	20.00025
16	3	17.8	51	No	23.07834
17	6	17.8	56	Yes	28.06439
18					



ANY QUESTIONS?



THE END