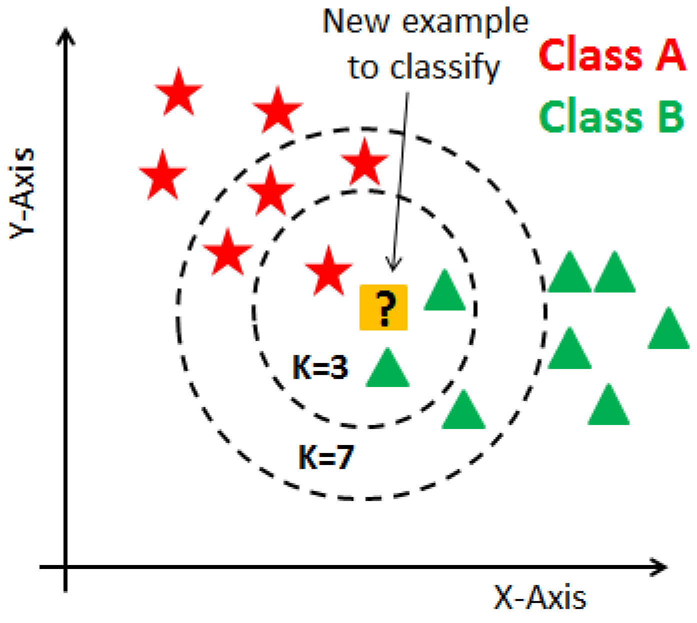
Do It Skills
do it. enjoy it.

K-NEAREST NEIGHBORS

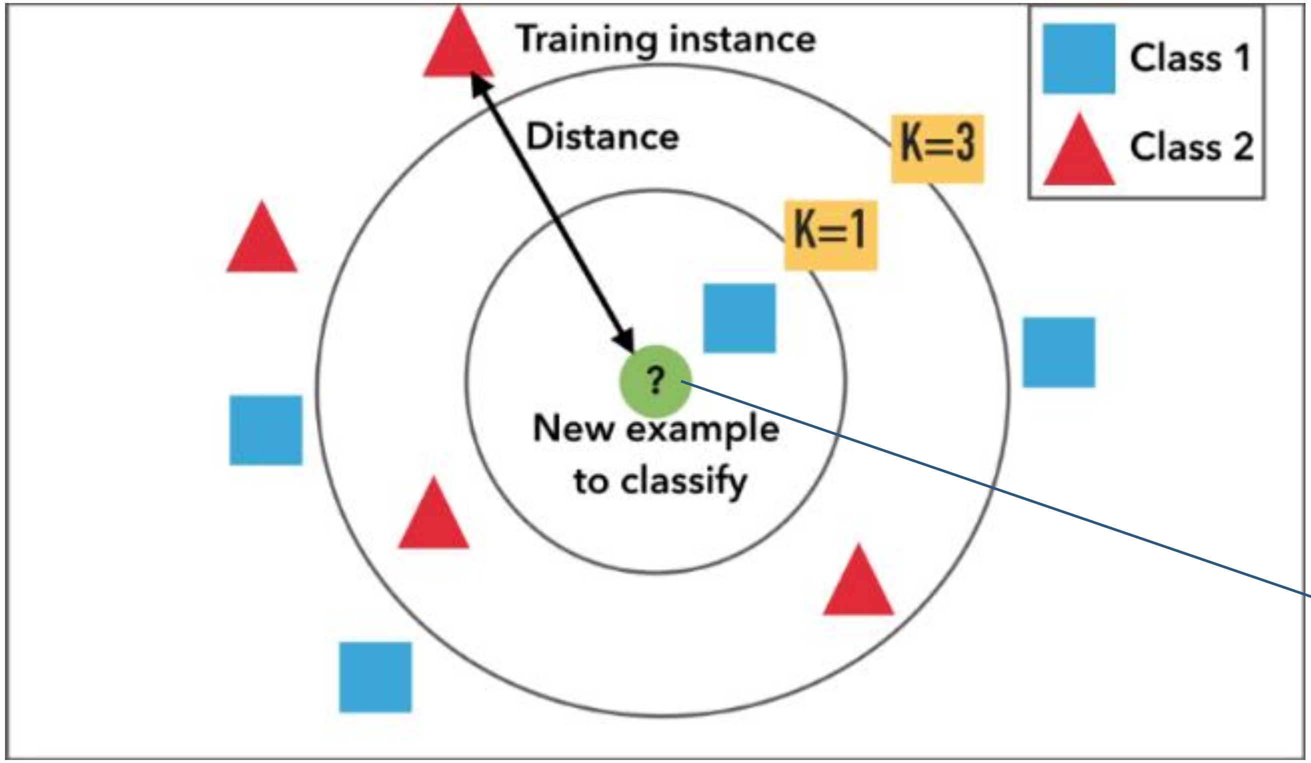
CLASSIFICATION ALGORITHM MACHINE LEARNING

WHAT IS KNN ?

Knn is a non-parametric supervised learning technique in which we try to classify the data point to a given category with the help of training set. In simple words, it captures information of all training cases and classifies new cases based on a similarity.



HOW DOES KNN WORK?



KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression).

New Data point

EUCLIDEAN , MANHATTAN DISTANCE

Distance functions

Euclidean

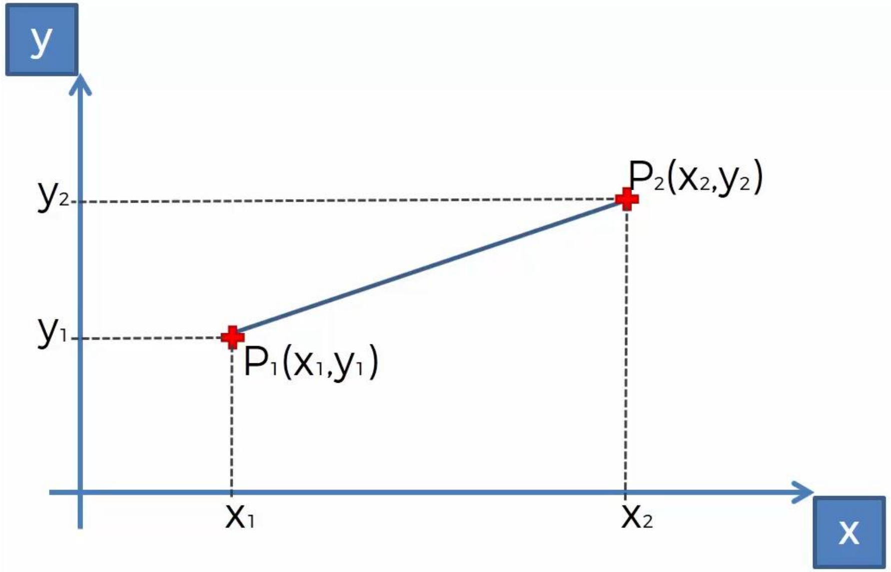
$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Manhattan

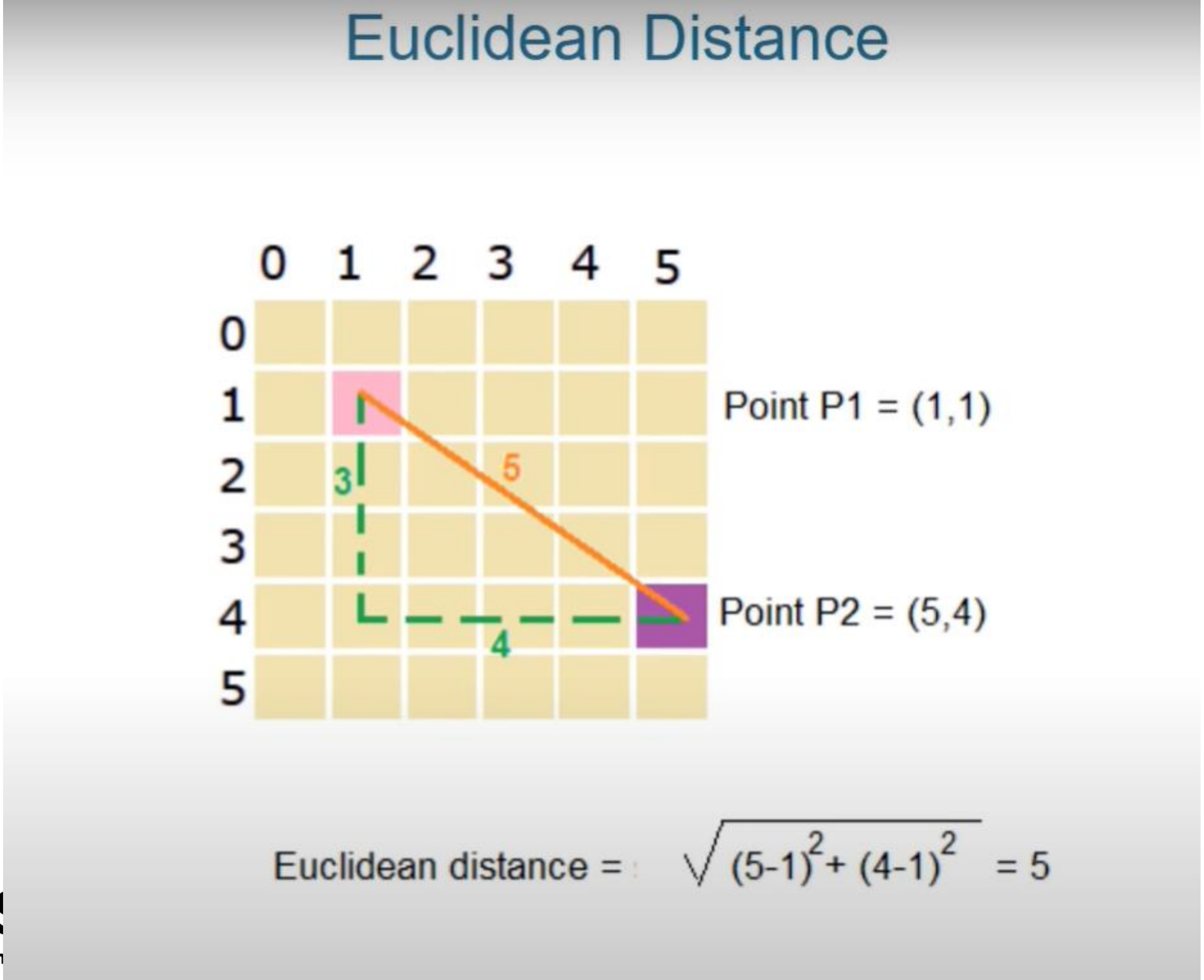
$$\sum_{i=1}^{k}\left|x_i - y_i\right|$$

CALCULATING DISTANCE:

To calculate the distance between two points (your new sample and all the data you have in your dataset), we will use the **Euclidean distance.**



$$d(x, x') = \sqrt{\left(x_1 - x'_1\right)^2 + \ldots + \left(x_n - x'_n\right)^2}$$



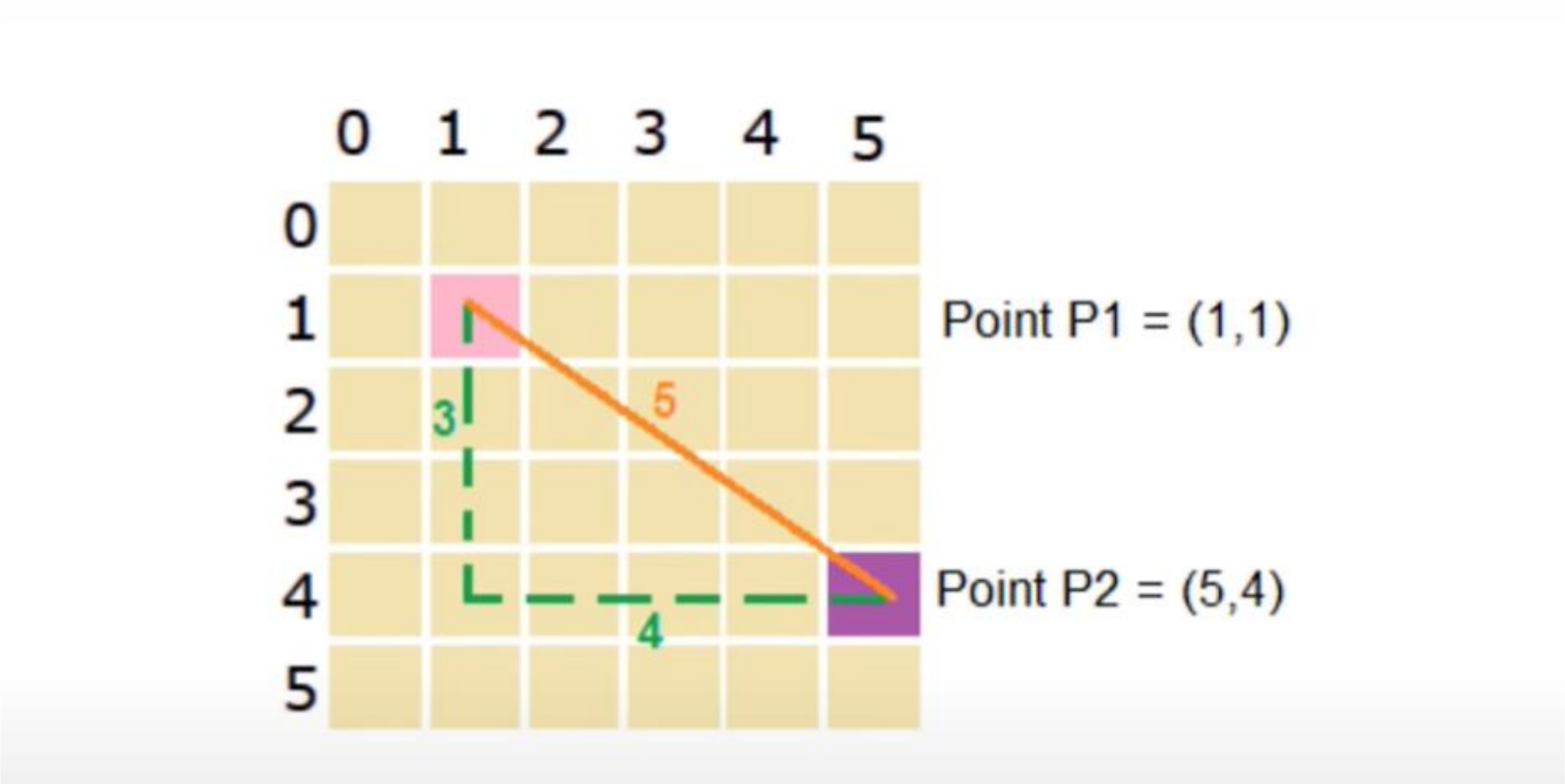
Euclidean Distance

Point P1 = (1,1)

Point P2 = (5,4)

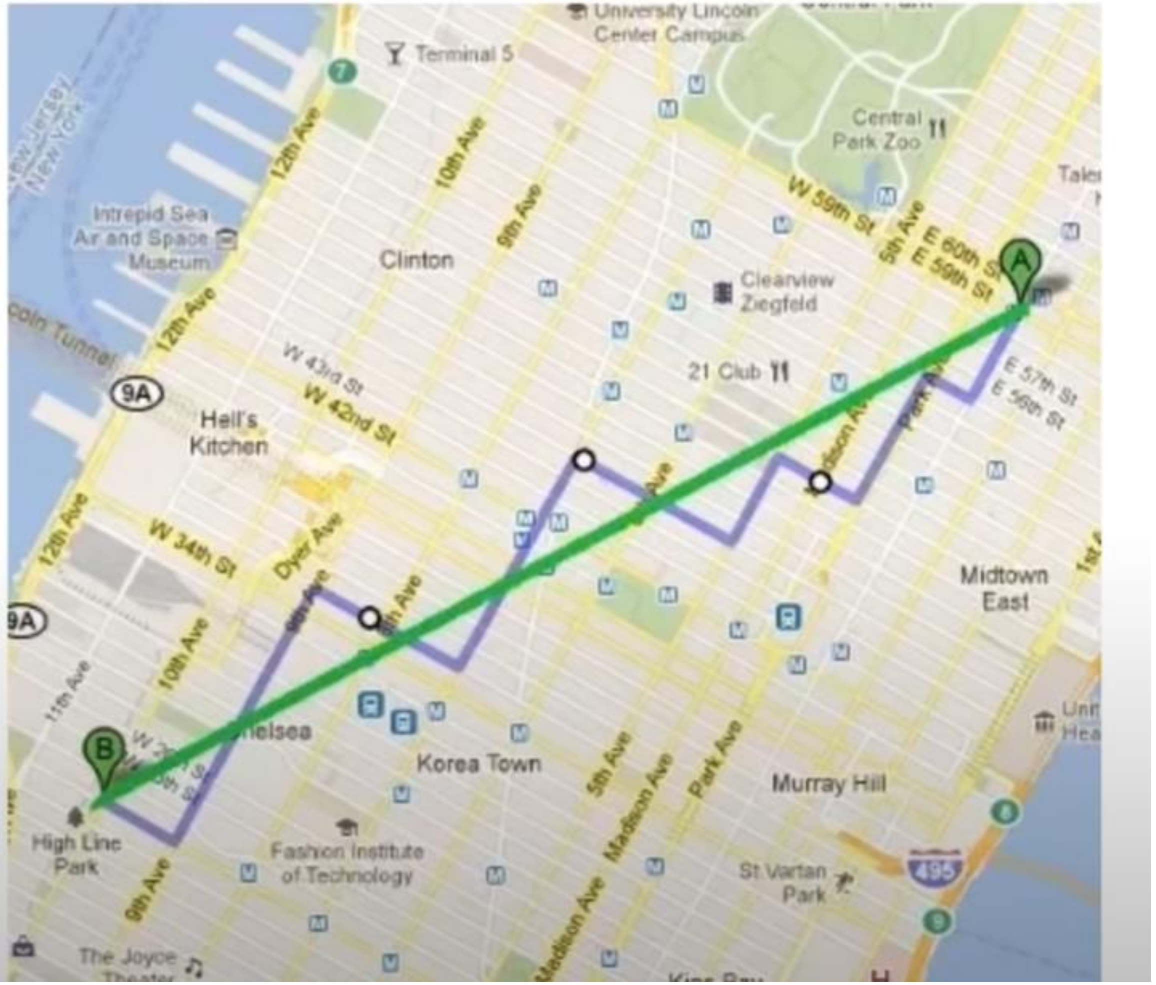
Euclidean distance = $\sqrt{(5-1)^2 + (4-1)^2} = 5$

MANHATTAN DISTANCE



Manhattan Distance = |5-1| + |4-1| = 7

EUCLIDEAN VS MANHATTAN DISTANCE



FEW IDEAS ON PICKING A VALUE FOR 'K'

1)There is no structured method to find the best value for "K". We need to find out with various values by trial and error and assuming that training data is unknown.

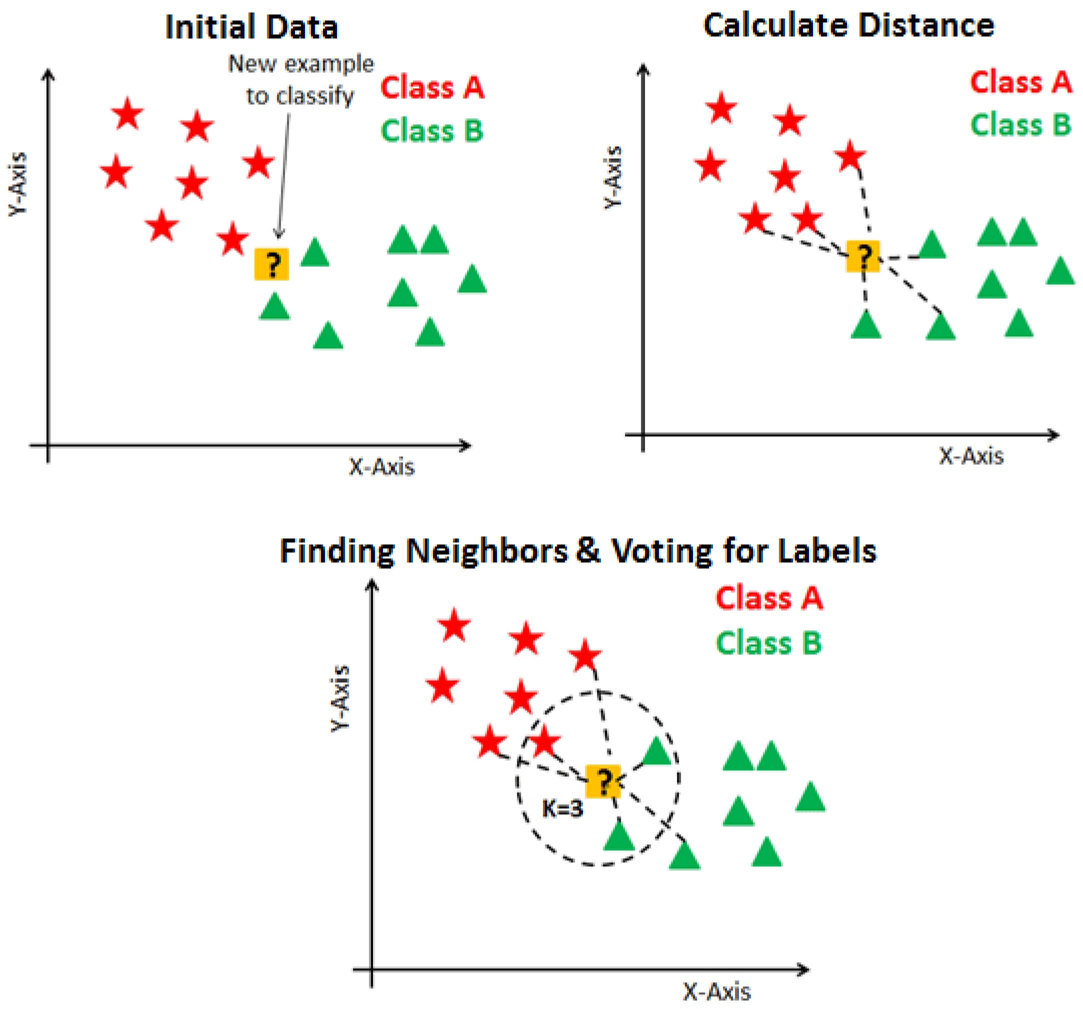
2)Choosing smaller values for K can be noisy and will have a higher influence on the result.

3) Larger values of K will have smoother decision boundaries which mean lower variance but increased bias. Also, computationally expensive.

Do It Skills
do it. enjoy it.

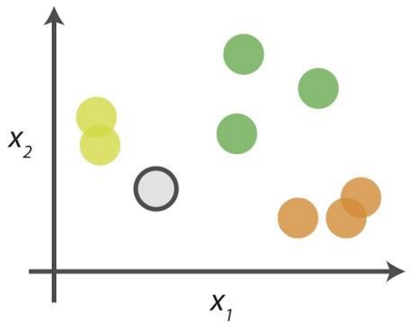
STEP FOR KNN

1 — Receive an unclassified data;
2 — Measure the distance (Euclidian, Manhattan, Minkowski or Weighted) from the new data to all others data that is already classified;
3 — Gets the K(K is a parameter that you define) smaller distances;
4 — Check the list of classes had the shortest distance and count the amount of each class that appears;
5 — Takes as correct class the class that appeared the most times;
6 —Classifies the new data with the class that you took in step 5;


Initial Data
New example to classify
Class A
Class B
?
Y-Axis
X-Axis
Calculate Distance
Class A
Class B
?
Y-Axis
X-Axis
Finding Neighbors & Voting for Labels
Class A
Class B
?
K=3
Y-Axis
X-Axis

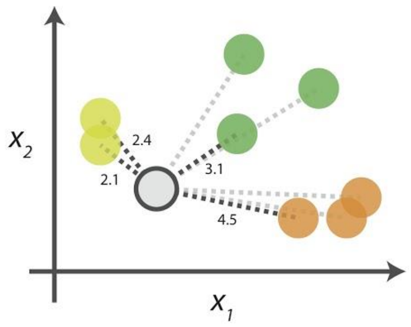
kNN Algorithm

0. Look at the data



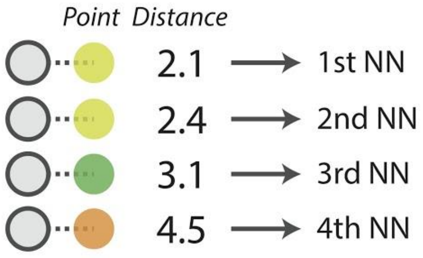
Say you want to classify the grey point into a class. Here, there are three potential classes - lime green, green and orange.

1. Calculate distances



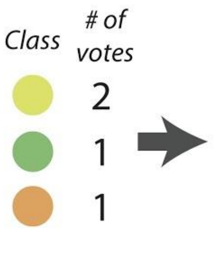
Start by calculating the distances between the grey point and all other points.

2. Find neighbours



Next, find the nearest neighbours by ranking points by increasing distance. The nearest neighbours (NNs) of the grey point are the ones closest in dataspace.

3. Vote on labels



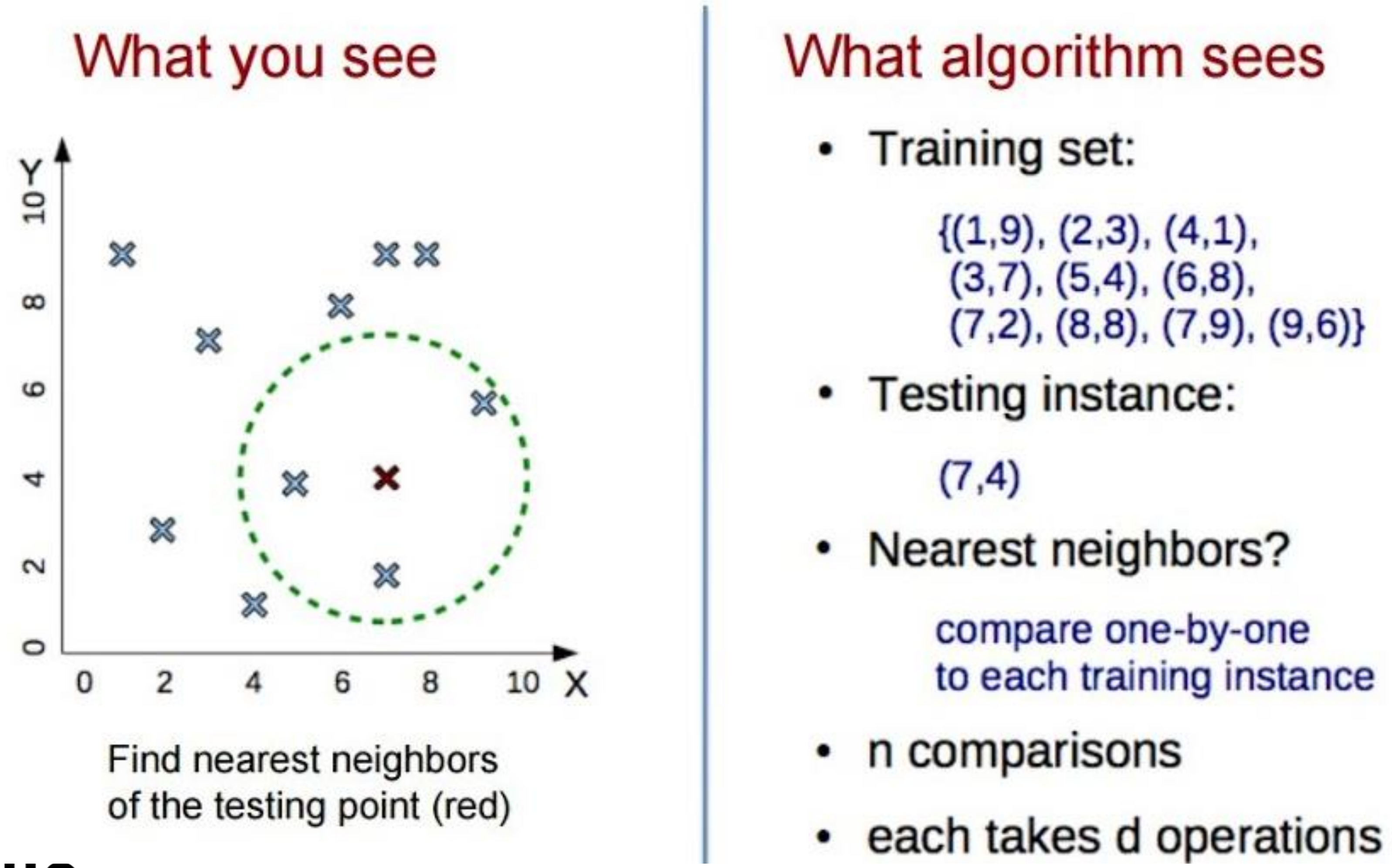
Class ● wins the vote!

Point ○ is therefore predicted to be of class ●.

Vote on the predicted class labels based on the classes of the k nearest neighbours. Here, the labels were predicted based on the k=3 nearest neighbours.

Do It Skills
do it. enjoy it.

WHY IS KNN SLOW



What you see

Find nearest neighbors of the testing point (red)

What algorithm sees

- Training set:

  {(1,9), (2,3), (4,1), (3,7), (5,4), (6,8), (7,2), (8,8), (7,9), (9,6)}

- Testing instance:

  (7,4)

- Nearest neighbors?

  compare one-by-one
  to each training instance

- n comparisons
- each takes d operations

T-SHIRT

EXAMPLE STEP 1

Height (in cms)	Weight (in kgs)	T Shirt Size
158	58	M
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

We want to classify the T shirt size for new person to class Medium or Large?

New customer named 'Monica' has height 161cm and weight 61kg.

We have height, weight and T-shirt size of some customers and we need to predict the T-shirt size of a new customer given only height and weight information we have. Data including height, weight and T-shirt size information is shown.

New customer named 'Monica' has height 161cm and weight 61kg.

STEP 2: CALCULATE DISTANCE

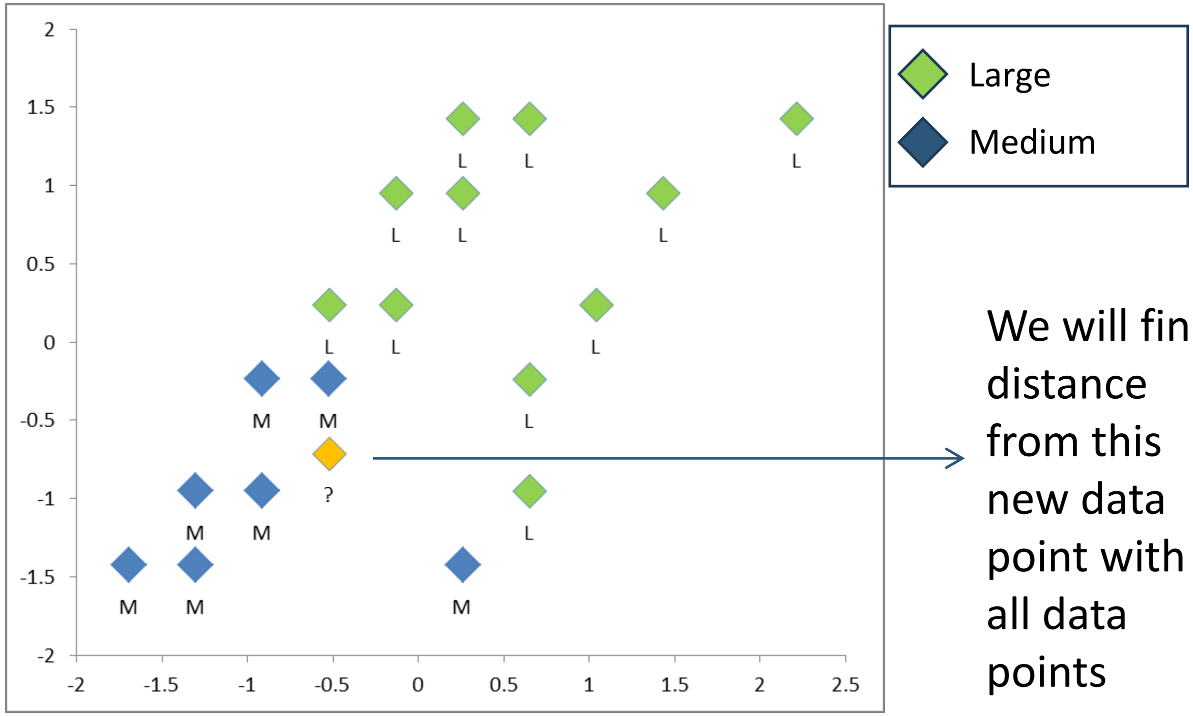
Height (in cms)	Weight (in kgs)	T Shirt Size	Distance
158	58	M	4.24
158	59	M	3.61
158	63	M	3.61
160	59	M	2.24
160	60	M	1.41
163	60	M	2.24
163	61	M	2.00
160	64	L	3.16
163	64	L	3.61
165	61	L	4.00
165	62	L	4.12
165	65	L	5.66
168	62	L	7.07
168	63	L	7.28
168	66	L	8.60
170	63	L	9.22
170	64	L	9.49
170	68	L	11.40
161	**61**		

Euclidian distance

=SQRT(New Obs. – old obs)^2 +(New Obs – Old obs)^2

=SQRT((161-158)^2+(61-58)^2)

New observations attributes



We will find distance from this new data point with all data points

STEP 3: FIND K-NEAREST NEIGHBORS (NON STANDARDIZED DATA)

Height (in cms)	Weight (in kgs)	T Shirt Size	Distance	Rank Minimum Distance
158	58	M	4.24	11
158	59	M	3.61	7
158	63	M	3.61	8
160	59	**M**	2.24	**3**
160	60	**M**	1.41	**1**
163	60	**M**	2.24	**3**
163	61	**M**	2.00	**2**
160	64	**L**	3.16	**5**
163	64	L	3.61	6
165	61	L	4.00	9
165	62	L	4.12	10
165	65	L	5.66	12
168	62	L	7.07	13
168	63	L	7.28	14
168	66	L	8.60	15
170	63	L	9.22	16
170	64	L	9.49	17
170	68	L	11.40	18

- **Let k be 5.** Then the algorithm searches for the 5 customers closest to Monica

- We will rank 5 distances as (k is 5) from lower to maximum.

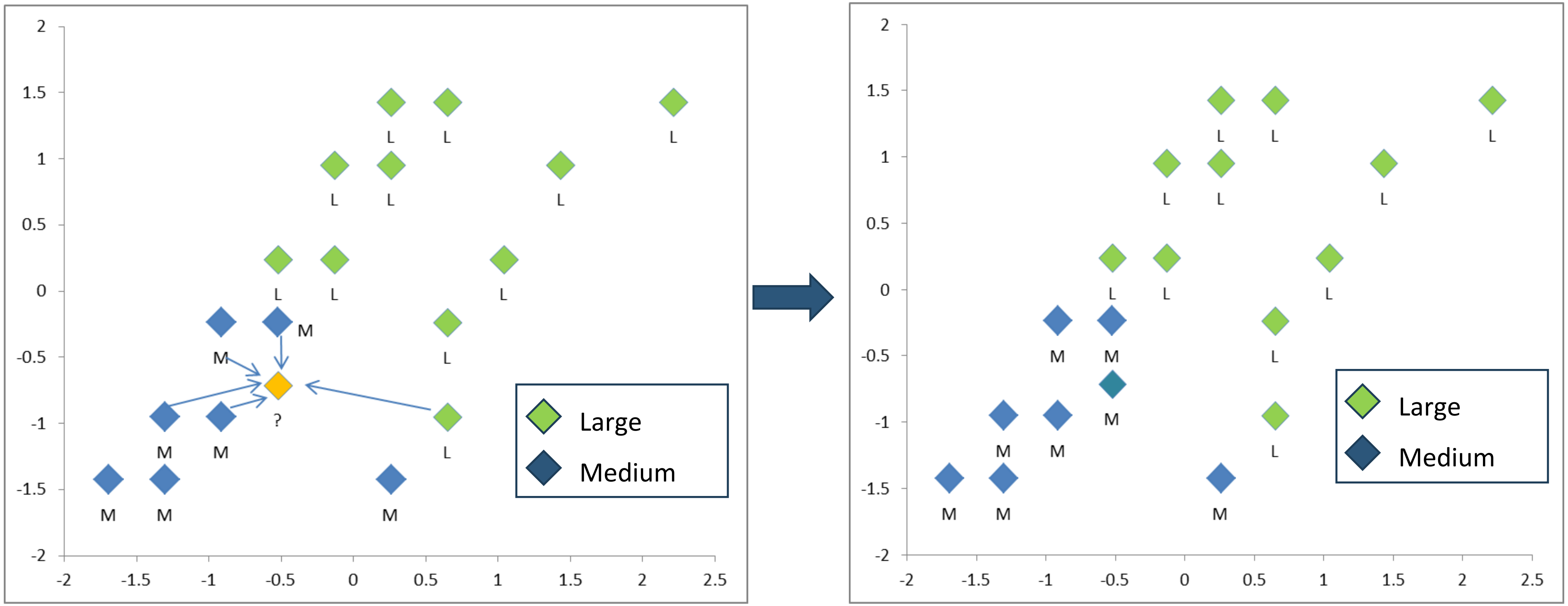
- As 4 of them had ‘Medium T shirt sizes’ and 1 had ‘Large T shirt size’ then your best guess for Monica is ‘Medium T shirt.

STEP 3: FIND K-NEAREST NEIGHBORS (STANDARDIZED DATA)

Height (in cms)	Weight (in kgs)	T Shirt Size	Distance	Rank Minimum Distance
-1.42749	-1.69246	M	1.37	11
-1.42749	-1.30189	M	1.06	7
-1.42749	0.260378	M	1.06	8
-0.95166	-1.30189	**M**	0.82	**4**
-0.95166	-0.91132	**M**	0.46	**1**
-0.23792	-0.91132	**M**	0.62	**3**
-0.23792	-0.52076	**M**	0.48	**2**
-0.95166	0.650945	L	1.20	9
-0.23792	0.650945	L	1.26	10
0.237915	-0.52076	**L**	0.95	**5**
0.237915	-0.13019	L	1.03	6
0.237915	1.041511	L	1.83	13
0.951662	-0.13019	L	1.71	12
0.951662	0.260378	L	1.84	14
0.951662	1.432078	L	2.57	17
1.427493	0.260378	L	2.28	15
1.427493	0.650945	L	2.44	16
1.427493	2.213211	L	3.47	18

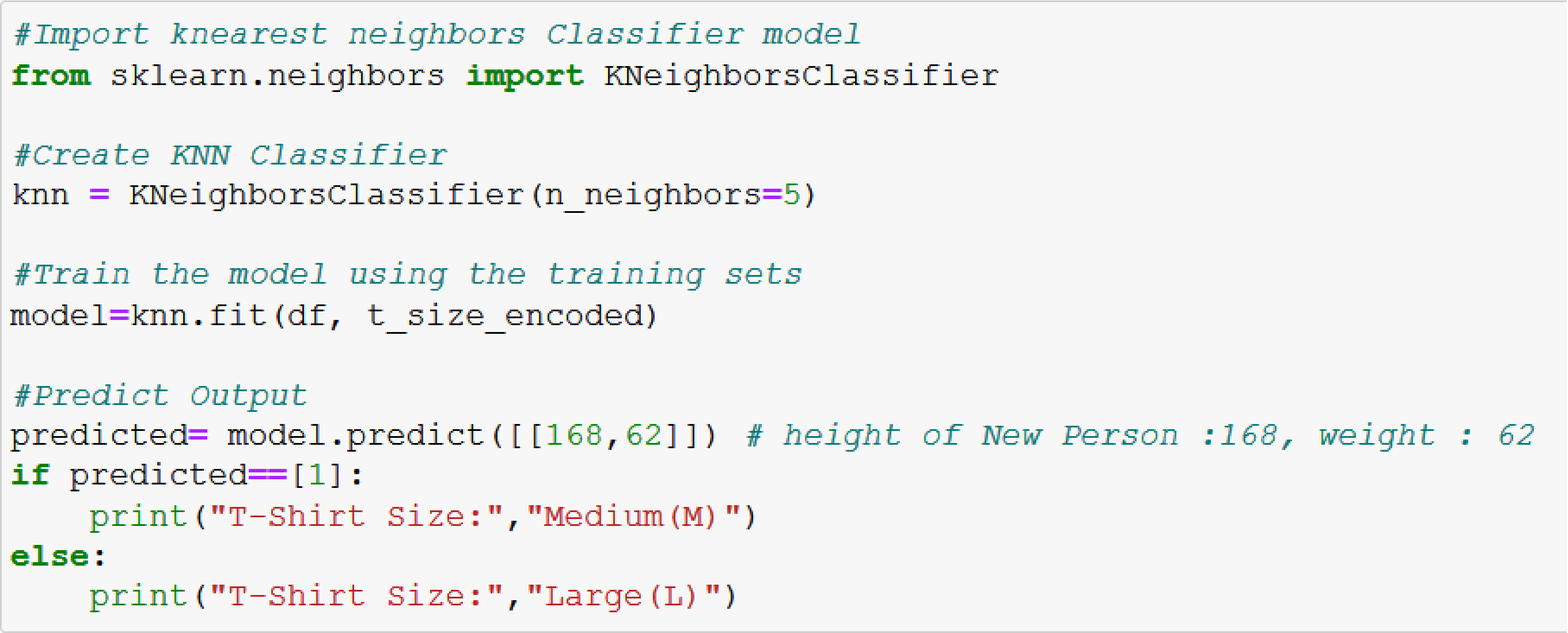
- Applying same procedure on standardized data.
- **Let k be 5.** Then the algorithm searches for the 5 customers closest to Monica
- We will rank 5 distances as (k is 5) from lower to maximum.
- The little difference we can see that position of one observation is change.
- As 4 of them had ‘Medium T shirt sizes’ and 1 had ‘Large T shirt size’ then your best guess for (New data point) Monica is ‘Medium T shirt.

FINAL RESULT



From 5 K nearest neighbors 4 are of class Medium so we will classify the new dataset to class Medium

PYTHON CODE

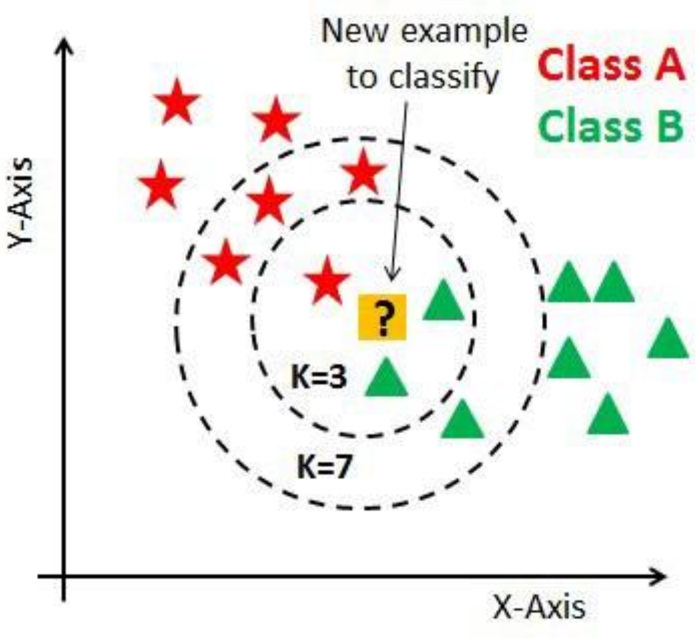


```python
#Import knearest neighbors Classifier model
from sklearn.neighbors import KNeighborsClassifier

#Create KNN Classifier
knn = KNeighborsClassifier(n_neighbors=5)

#Train the model using the training sets
model=knn.fit(df, t_size_encoded)

#Predict Output
predicted= model.predict([[168,62]]) # height of New Person :168, weight : 62
if predicted==[1]:
    print("T-Shirt Size:","Medium(M)")
else:
    print("T-Shirt Size:","Large(L)")
```

Output :

```
T-Shirt Size: Medium(M)
```

HOW TO CALCULATE THE VALUE OF K



IMPROVEMENTS

- An easy and mild approach to change skewed class distributions is by implementing weighted voting.
- Changing the distance metric (i.e. Hamming distance for text classification)
- Dimensionality reduction techniques like PCA should be executed prior to applying KNN and help make the distance metric more meaningful.

PROS AND CONS

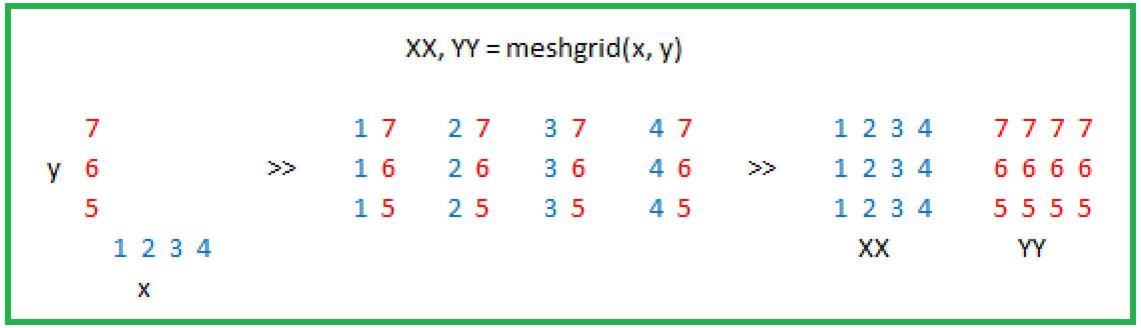
PROS

- The algorithm is simple and easy to implement.
- There's no need to build a model, tune several parameters, or make additional assumptions.

CONS

- The algorithm gets significantly slower as the number of examples and/or predictors/independent variables increase.

MESHGRID



XX, YY = meshgrid(x, y)

```
     7                  1 7   2 7   3 7   4 7        1 2 3 4   7 7 7 7
y    6           >>     1 6   2 6   3 6   4 6   >>   1 2 3 4   6 6 6 6
     5                  1 5   2 5   3 5   4 5        1 2 3 4   5 5 5 5
       1 2 3 4                                         XX        YY
         x
```

Do It Skills
do it. enjoy it.

THANK YOU

✉ ARUNKG99@GMAIL.COM

WWW.DOITSKILLS.COM