

K-Nearest Neighbor Classifier

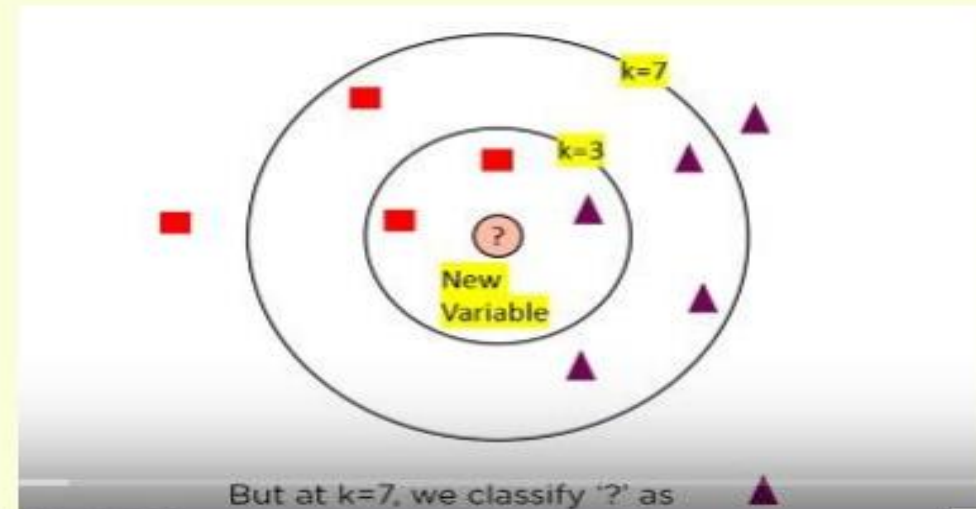
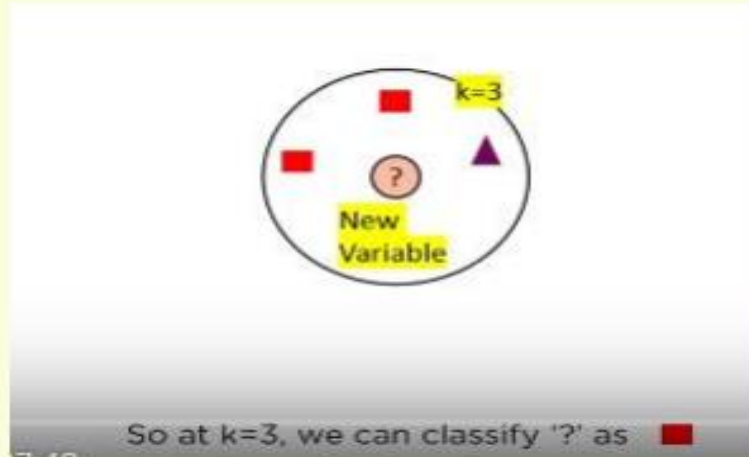
KNN - Classifier

- It classifies data points based on how its neighbors are classified
- KNN stores all available cases and classifies new cases based on similarity measure
- 'K' in KNN is a parameter that refers to the number of nearest neighbors to include in the majority voting process
- Choosing the right value of 'K' would determine the accuracy of the classifier

KNN Algorithm

1. Determine parameter K = Number of nearest neighbours.
2. Calculate the distance between the query instance and all the training samples.
3. Sort the distance and determine nearest neighbours based on the K^{th} minimum distance
4. Gather the values of Y of the nearest neighbours.
5. Use average of nearest neighbours as the prediction value of the query instance.

K value decides the classification of the data



How to choose the value of k

- ▶ K is usually calculated depending on the total number of records in the dataset.
- ▶ **$K = \sqrt{n}$** , where n is the size of the dataset
- ▶ Odd value for K is recommended to avoid confusion when we have equal number of classes around the new record.

When do we use KNN Classifier

- Data is labelled
- Data set is small
 - Because 'KNN' is a lazy learner. It does not learn a discriminative function from the training data
- Data set is noise free

About KNN

- ▶ The k-nearest-neighbor (knn) method was first described in the early 1950s.
- ▶ It has been widely used in the area of pattern recognition.
- ▶ Nearest-neighbor classifiers are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it.
- ▶ When given an unknown tuple, a k-nearest-neighbor classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. These k training tuples are the k “nearest neighbors” of the unknown tuple.

Euclidean distance

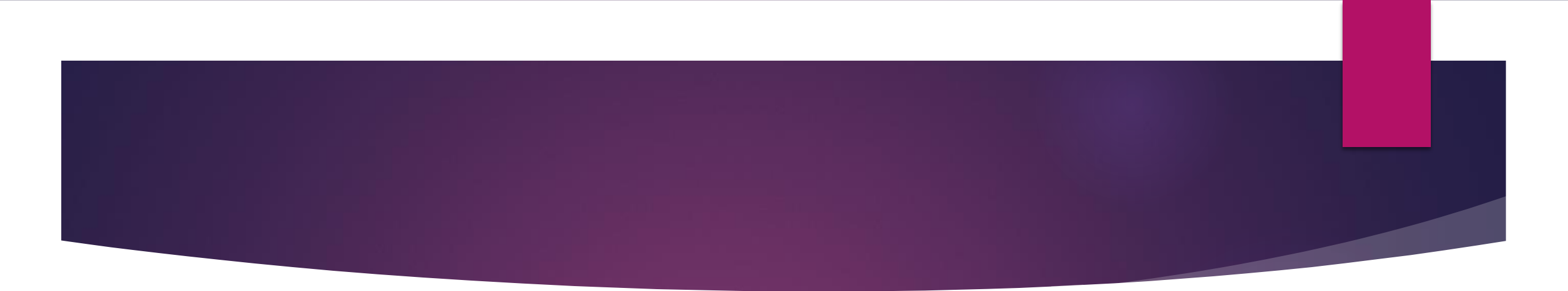
- ▶ “Closeness” is defined in terms of a distance metric, such as Euclidean distance. The Euclidean distance between two points or tuples, say, $X_1 = (x_{11}, x_{12}, \dots, x_{1n})$ and $X_2 = (x_{21}, x_{22}, \dots, x_{2n})$, is

$$\text{dist}(X_1, X_2) = \sqrt{\sum_{i=1}^n (x_{1i} - x_{2i})^2}.$$

Normalize the data

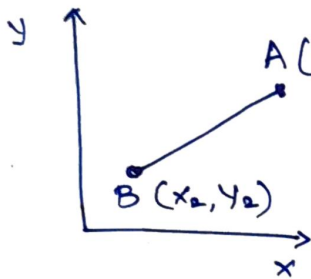
- ▶ Normalize the values of each attribute before using Equation. This helps to prevent attributes with initially large ranges (such as income) from outweighing attributes with initially smaller ranges (such as binary attributes).
- ▶ Min-max normalization, for example, can be used to transform a value v of a numeric attribute A to v' in the range $[0, 1]$ by computing

$$v' = \frac{v - \min_A}{\max_A - \min_A}$$

- 
- ▶ “But how can distance be computed for attributes that are not numeric, but categorical, such as color?”
 - ▶ For categorical attributes, a simple method is to compare the corresponding value of the attribute in tuple X1 with that in tuple X2.
 - ▶ If the two are identical (e.g., tuples X1 and X2 both have the color blue), then the difference between the two is taken as 0. If the two are different (e.g., tuple X1 is blue but tuple X2 is red), then the difference is considered to be 1.

KNN classifier.

How to find the distance between the two records?



Euclidean distance (d)

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

Weight (x_2)	Height (y_2)	Class	Euclidean Distance
51	167	Underweight	6.7
62	182	Normal	13
69	176	Normal	13.4
64	173	Normal	7.6
65	172	Normal	8.2
56	174	Underweight	4.1
58	169	Normal	1.4
57	173	Normal	3
55	170	Normal	2

Calculate the Euclidean Distance d , between the given new record and the existing records.

(i) Given: $(57, 170)$ $x_1 = 57$
 $y_1 = 170$

First record: $x_2 = 51$ $y_2 = 167$

$$\text{Euclidean distance: } d = \sqrt{(51 - 57)^2 + (167 - 170)^2}$$

$$d = \sqrt{6^2 + 3^2} = \sqrt{36+9} = \sqrt{45} = 6.7.$$

$$(ii) X_1 = 57 \quad Y_1 = 170, \quad X_2 = 62 \quad Y_2 = 182$$

$$d = \sqrt{(62-57)^2 + (182-170)^2} = \sqrt{5^2 + 12^2} = 13$$

$$(iii) X_1 = 57 \quad Y_1 = 170, \quad X_2 = 69 \quad Y_2 = 176$$

$$d = \sqrt{(69-57)^2 + (176-170)^2} = \sqrt{12^2 + 6^2} = 13.4$$

$$(iv) X_1 = 57 \quad Y_1 = 170, \quad X_2 = 64 \quad Y_2 = 173$$

$$d = \sqrt{(64-57)^2 + (173-170)^2} = \sqrt{7^2 + 3^2} = 7.6$$

$$(v) X_1 = 57 \quad Y_1 = 170, \quad X_2 = 65 \quad Y_2 = 172$$

$$d = \sqrt{(65-57)^2 + (172-170)^2} = \sqrt{8^2 + 2^2} = 8.2$$

$$(vi) X_1 = 57 \quad Y_1 = 170, \quad X_2 = 56 \quad Y_2 = 174$$

$$d = \sqrt{(56-57)^2 + (174-170)^2} = \sqrt{(-1)^2 + (4)^2} = 4.1$$

$$(vii) X_1 = 57 \quad Y_1 = 170, \quad X_2 = 58 \quad Y_2 = 169$$

$$d = \sqrt{(58-57)^2 + (169-170)^2} = \sqrt{1^2 + (-1)^2} = 1.4$$

$$(viii) X_1 = 57 \quad Y_1 = 170, \quad X_2 = 57 \quad Y_2 = 173$$

$$d = \sqrt{(57-57)^2 + (173-170)^2} = \sqrt{3^2} = 3$$

$$(ix) X_1 = 57 \quad Y_1 = 170, \quad X_2 = 55 \quad Y_2 = 170$$

$$d = \sqrt{(55-57)^2 + (170-170)^2} = \sqrt{2^2} = 2$$

Assume K as 3. Find 3 records with minimum (d)

records with maximum class label count are chosen.

The class label of the new record is the class label that has maximum count. Here the CL is Normal



Topic :

1. K-Nearest Neighbour Classifier

Description:

K-Nearest Neighbour Classifier

The k -nearest-neighbour method was first described in the early 1950s. The method is labour intensive when given large training sets, and did not gain popularity until the 1960s when increased computing power became available. It has since been widely used in the area of pattern recognition.

Nearest-neighbour classifiers are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it. The training tuples are described by n attributes. Each tuple represents a point in an n -dimensional space. In this way, all the training tuples are stored in an n -dimensional pattern space.

When given an unknown tuple, a **k -nearest-neighbour classifier** searches the pattern space for the k training tuples that are closest to the unknown tuple. These k training tuples are the k “nearest neighbours” of the unknown tuple.

“Closeness” is defined in terms of a distance metric, such as Euclidean distance. The Euclidean distance between two points or tuples, say, $X_1 = (x_{11}, x_{12}, \dots, x_{1n})$ and $X_2 = (x_{21}, x_{22}, \dots, x_{2n})$, is

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^n (x_{1i} - x_{2i})^2}.$$

Some points about the K-NN classification:

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

K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

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.

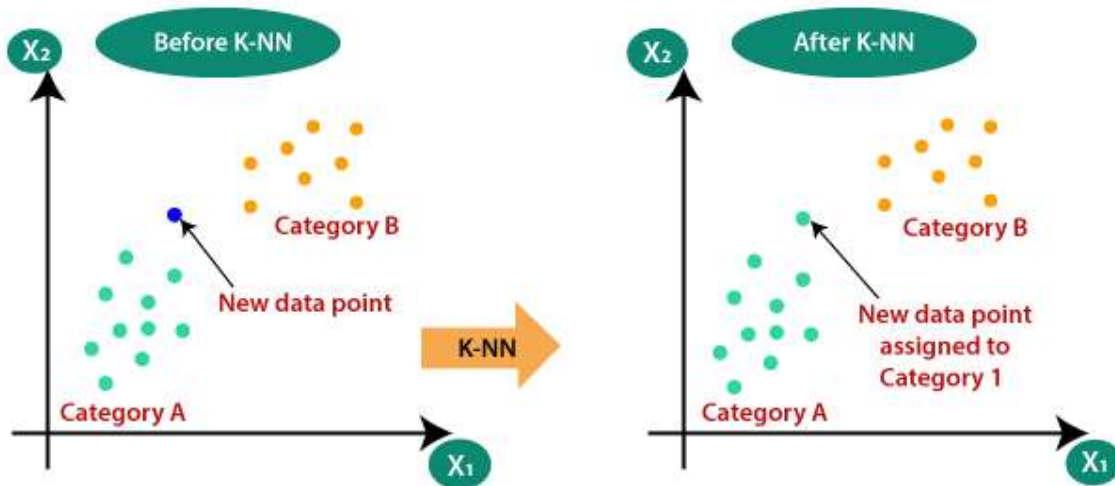
K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification it performs an action on the dataset.

KNN algorithm at the training phase just stores the dataset and when it gets new data, it classifies that data into a category that is much similar to the nearest data.

Diagrammatic View of KNN

Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x_1 , so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset. Consider the below diagram:



The K-NN working can be explained on the basis of the below algorithm:

Step-1: Select the number K of the neighbours

Step-2: Calculate the Euclidean distance of K number of neighbours

Step-3: Take the K nearest neighbours as per the calculated Euclidean distance.

Step-4: Among these k neighbours, 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 neighbour is maximum.

Step-6: K-NN model is ready.

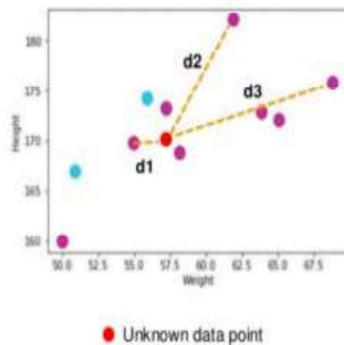
SCHOOL OF INFORMATION TECHNOLOGY AND ENGINEERING
ITA5007-Data Mining and Business Intelligence
MCA

Example 1:

Consider a dataset having two variables and they are height in centimeter & weight in kilogram. Each point is classified as normal or underweight. Number of nearest neighbours is 3 (i.e. $k=3$). Predict the class of a new customer given only height and weight information we have (i.e. weight = 57, height = 170, and class = ?) using KNN classification.

Weight	Height	Class
51	167	Underweight
62	182	Normal
69	176	Normal
64	173	Normal
65	172	Normal
56	174	Underweight
58	169	Normal
57	173	Normal
55	170	Normal

Apply the Euclidean distance formula between the known data values and the unknown data point:



$$\text{dist}(d1) = \sqrt{(170-167)^2 + (57-51)^2} \approx 6.7$$

$$\text{dist}(d2) = \sqrt{(170-182)^2 + (57-62)^2} \approx 13$$

$$\text{dist}(d3) = \sqrt{(170-176)^2 + (57-69)^2} \approx 13.4$$

Similarly, we will calculate Euclidean distance of unknown data point from all the points in the dataset

After calculating the Euclidean distance the table is updated with the distance measure:

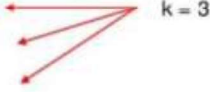
Where $(x1, y1) = (57, 170)$ whose class we have to classify

Weight(x2)	Height(y2)	Class	Euclidean Distance
51	167	Underweight	6.7
62	182	Normal	13
69	176	Normal	13.4
64	173	Normal	7.6
65	172	Normal	8.2
56	174	Underweight	4.1
58	169	Normal	1.4
57	173	Normal	3
55	170	Normal	2



SCHOOL OF INFORMATION TECHNOLOGY AND ENGINEERING
ITA5007-Data Mining and Business Intelligence
MCA

Class	Euclidean Distance
Underweight	6.7
Normal	13
Normal	13.4
Normal	7.6
Normal	8.2
Underweight	4.1
Normal	1.4
Normal	3
Normal	2



Here majority of neighbours are pointing to Normal. Hence as per KNN algorithm the class of the unknown data point is also Normal.

Choosing the right value for K

To select the K that's right for the data, run the KNN algorithm several times with different values of K and choose the K that reduces the number of errors that are encountered while maintaining the algorithm's ability to accurately make predictions.

Here are some things to keep in mind:

- As we decrease the value of K to 1, our predictions become less stable. Imagine K=1 and we have a query point surrounded by several reds and one green but the green is the single nearest neighbour. Reasonably, we would think the query point is most likely red, but because K=1, KNN incorrectly predicts that the query point is green.
- Inversely, as we increase the value of K, our predictions become more stable due to majority voting / averaging, and thus, more likely to make more accurate predictions (up to a certain point).
- Usually the K value is selected as an odd number to have a tiebreaker.

Distance measures:

Distance functions

Euclidean	$\sqrt{\sum_{i=1}^k (x_i - y_i)^2}$
Manhattan	$\sum_{i=1}^k x_i - y_i $
Minkowski	$\left(\sum_{i=1}^k (x_i - y_i)^q \right)^{1/q}$

Advantages

1. The algorithm is simple and easy to implement.
2. There's no need to build a model, tune several parameters, or make additional assumptions.
3. The algorithm is versatile. It can be used for classification, regression, and search.

Disadvantages

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



SCHOOL OF INFORMATION TECHNOLOGY AND ENGINEERING
ITA5007-Data Mining and Business Intelligence
MCA

Example 2:

Suppose 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 (i.e. height = 161, weight = 61 and T-shirt size = ?). Data including height, weight and T-shirt size information is shown below.

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
161	61	?

Step 1: Determine parameter K = Number of nearest neighbours.

K = 5

Step 2: Calculate the distance between the query instance and all the training samples.

S.NO.	Height (in cms)	Weight (in kgs)	T Shirt Size	Euclidean distance
1	158	58	M	4.2
2	158	59	M	3.6
3	158	63	M	3.6
4	160	59	M	2.2
5	160	60	M	1.4
6	163	60	M	2.2
7	163	61	M	2



SCHOOL OF INFORMATION TECHNOLOGY AND ENGINEERING
ITA5007-Data Mining and Business Intelligence
MCA

8	160	64	L	3.2
9	163	64	L	3.6
10	165	61	L	4
11	165	62	L	4.1
12	165	65	L	5.7
13	168	62	L	7.1
14	168	63	L	7.3
15	168	66	L	8.6
16	170	63	L	9.2
17	170	64	L	9.5
18	170	68	L	11.4

Step 3: Sort the distance and determine nearest neighbours based on the K^{th} minimum distance

S.NO.	Height (in cms)	Weight (in kgs)	T Shirt Size	Euclidean distance	Sorted Euclidean distance	Rank minimum distance	Is it included in 5 nearest neighbors
1	158	58	M	4.2	1.4		No
2	158	59	M	3.6	2		No
3	158	63	M	3.6	2.2		No
4	160	59	M	2.2	2.2	3	Yes
5	160	60	M	1.4	3.2	1	Yes
6	163	60	M	2.2	3.6	3	Yes
7	163	61	M	2	3.6	2	Yes
8	160	64	L	3.2	3.6	5	Yes
9	163	64	L	3.6	4		No
10	165	61	L	4	4.1		No
11	165	62	L	4.1	4.2		No
12	165	65	L	5.7	5.7		No
13	168	62	L	7.1	7.1		No



SCHOOL OF INFORMATION TECHNOLOGY AND ENGINEERING
ITA5007-Data Mining and Business Intelligence
MCA

14	168	63	L	7.3	7.3	No
15	168	66	L	8.6	8.6	No
16	170	63	L	9.2	9.2	No
17	170	64	L	9.5	9.5	No
18	170	68	L	11.4	11.4	No

Step 4: Gather the values of Y of the nearest neighbours.

S.NO.	Height (in cms)	Weight (in kgs)	T Shirt Size	Euclidean distance	Sorted Euclidean distance	Rank minimum distance	Is it included in 5 nearest neighbours	Y = Category of the nearest neighbour
1	158	58	M	4.2	1.4		No	-
2	158	59	M	3.6	2		No	-
3	158	63	M	3.6	2.2		No	-
4	160	59	M	2.2	2.2	3	Yes	M
5	160	60	M	1.4	3.2	1	Yes	M
6	163	60	M	2.2	3.6	3	Yes	M
7	163	61	M	2	3.6	2	Yes	M
8	160	64	L	3.2	3.6	5	Yes	L
9	163	64	L	3.6	4		No	-
10	165	61	L	4	4.1		No	-
11	165	62	L	4.1	4.2		No	-
12	165	65	L	5.7	5.7		No	-
13	168	62	L	7.1	7.1		No	-
14	168	63	L	7.3	7.3		No	-
15	168	66	L	8.6	8.6		No	-
16	170	63	L	9.2	9.2		No	-
17	170	64	L	9.5	9.5		No	-
18	170	68	L	11.4	11.4		No	-

Step 5: Use average of nearest neighbours as the prediction value of the query instance.

We have 4 M and 1 L. Since $4 > 1$, 161, 61 is included in M. Therefore

Height (in cms)	Weight (in kgs)	T Shirt Size
161	61	M

Courtesy: <https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning>

Questions:



SCHOOL OF INFORMATION TECHNOLOGY AND ENGINEERING
ITA5007-Data Mining and Business Intelligence
MCA

Question 1:

Name	Acid Durability	Strength	Class
Type-1	7	7	Bad
Type-2	7	4	Bad
Type-3	3	4	Good
Type-4	1	4	Good

Test-Data → acid durability=3, and strength=7, class=?

Question 2:

User ID	Gender	Age	EstimatedSalary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	0
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	0
15728773	Male	27	58000	0
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	0
15727311	Female	35	65000	0
15570769	Female	26	80000	0
15606274	Female	26	52000	0
15746139	Male	20	86000	0
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1
15617482	Male	45	26000	1
15704583	Male	46	28000	1
15621083	Female	48	29000	1
15649487	Male	45	22000	1
15736760	Female	47	49000	1

New customer: Male, Age =28, Estimated salary =90000 Purchased =?

Question 3:

Sepal Length	Sepal Width	Species
5.3	3.7	Setosa
5.1	3.8	Setosa
7.2	3.0	Virginica
5.4	3.4	Setosa
5.1	3.3	Setosa
5.4	3.9	Setosa
7.4	2.8	Virginica
6.1	2.8	Versicolor
7.3	2.9	Virginica
6.0	2.7	Versicolor
5.8	2.8	Virginica
6.3	2.3	Versicolor
5.1	2.5	Versicolor
6.3	2.5	Versicolor
5.5	2.4	Versicolor

New species : Sepal Length =6.2, Sepal Width =3.1, Species =?

SCHOOL OF INFORMATION TECHNOLOGY AND ENGINEERING
ITA5007-Data Mining and Business Intelligence
MCA