KNN MCQ’S

Context 1-2:

Suppose, you have trained a k-NN model and now you want to get the prediction on test data. Before getting the prediction suppose you want to calculate the time taken by k-NN for predicting the class for test data.
Note: Calculating the distance between 2 observation will take D time.

1. What would be the relation between the time taken by 1-NN, 2-NN, 3-NN.

A) 1-NN >2-NN >3-NN
B) 1-NN < 2-NN < 3-NN
C) 1-NN ~ 2-NN ~ 3-NN
D) None of these

Solution: C

The training time for any value of k in kNN algorithm is the same.

2. What would be the time taken by 1-NN if there are N(Very large) observations in test data?

A) N*D
B) N*D*2
C) (N*D)/2
D) None of these

Solution: A

The value of N is very large, so option A is correct

3. Following are the two statements given for k-NN algorithm, which of the statement(s)

is/are true?

1. We can choose optimal value of k with the help of cross validation
2. Euclidean distance treats each feature as equally important

A) 1
B) 2
C) 1 and 2
D) None of these

Solution: C

Both the statements are true

4. In k-NN what will happen when you increase/decrease the value of k?

A) The boundary becomes smoother with increasing value of K
B) The boundary becomes smoother with decreasing value of K
C) Smoothness of boundary doesn’t dependent on value of K
D) None of these

Solution: A

The decision boundary would become smoother by increasing the value of K

5. True-False: It is possible to construct a 2-NN classifier by using the 1-NN classifier?

A) TRUE
B) FALSE

Solution: A

You can implement a 2-NN classifier by ensembling 1-NN classifiers

6. **Which of the following statements is true for k-NN classifiers?**

A) The classification accuracy is better with larger values of k
B) The decision boundary is smoother with smaller values of k
C) The decision boundary is linear
D) k-NN does not require an explicit training step

Solution: D

Option A: This is not always true. You have to ensure that the value of k is not too high or not too low.

Option B: This statement is not true. The decision boundary can be a bit jagged

Option C: Same as option B

Option D: This statement is true

7. **You have given the following 2 statements, find which of these options is/are true in case of k-NN?**

1. In case of very large value of k, we may include points from other classes into the neighbourhood.
2. In case of too small value of k the algorithm is very sensitive to noise

A) 1
B) 2
C) 1 and 2
D) None of these

Solution: C

Both the options are true and are self explanatory.

8. A company has built a kNN classifier that gets 100% accuracy on training data. When they deployed this model on client side it has been found that the model is not at all accurate. Which of the following thing might go wrong?

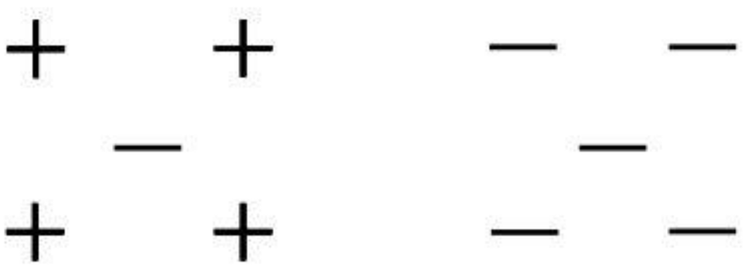
Note: Model has successfully deployed and no technical issues are found at client side except the model performance

A) It is probably a over fitted model
B) It is probably a under fitted model
C) Can't say
D) None of these

Solution: A

In an over fitted module, it seems to be performing well on training data, but it is not generalized enough to give the same results on a new data.

9. **Which of the following value of k in the following graph would you give least leave one out cross validation accuracy?**



A) 1
B) 2
C) 3
D) 5
Solution: B

If you keep the value of k as 2, it gives the lowest cross validation accuracy. You can try this out yourself.

10. **Suppose you have given the following images(1 left, 2 middle and 3 right), Now your task is to find out the value of k in k-NN in each image where k1 is for 1st, k2 is for 2nd and k3 is for 3rd figure.**

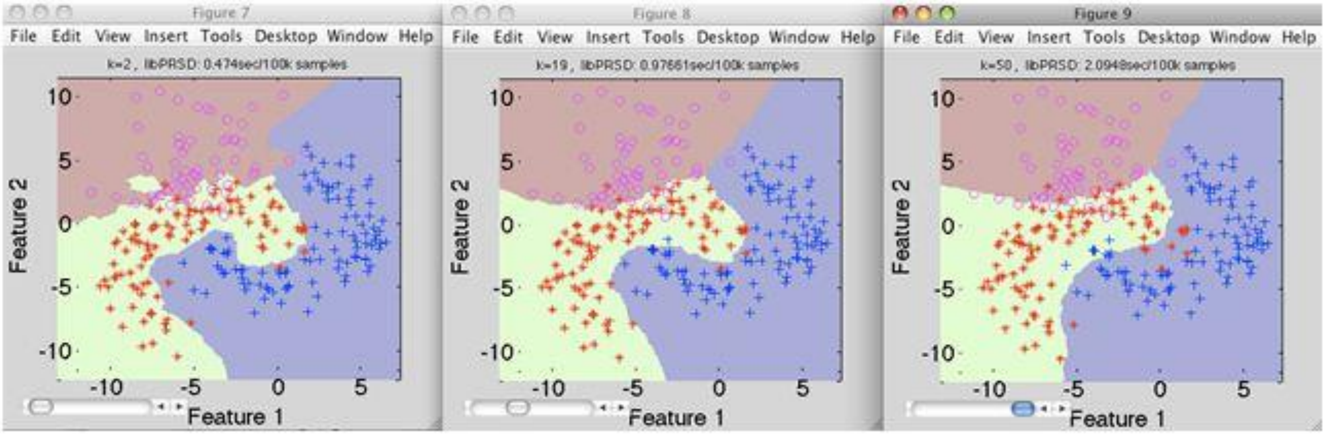


Figure 7
File Edit View Insert Tools Desktop Window Help
k=2 , libPRSD: 0.474sec/100k samples
Feature 2
Feature 1
10
5
0
-5
-10
-10
-5
0
5
Figure 8
File Edit View Insert Tools Desktop Window Help
k=19 , libPRSD: 0.97661sec/100k samples
Feature 2
Feature 1
Figure 9
File Edit View Insert Tools Desktop Window Help
k=50 , libPRSD: 2.0948sec/100k samples
Feature 2
Feature 1

A) k1 > k2> k3
B) k1<k2
C) k1 = k2 = k3
D) None of these
Solution: D

Value of k is highest in k3, whereas in k1 it is lowest

11. **Below are two statements given? Which of the following will be true both statements?**

1. k-NN is a memory-based approach is that the classifier immediately adapts as we collect new training data.

2. The computational complexity for classifying new samples grows linearly with the number of samples in the training dataset in the worst-case scenario.

A) 1
B) 2
C) 1 and 2
D) None of these

Solution: C

Both are true and self explanatory

12. **In k-NN it is very likely to over fit due to the curse of dimensionality. Which of the following option would you consider to handle such problem?**

1. Dimensionality Reduction
2. Feature selection

A) 1
B) 2
C) 1 and 2
D) None of these

Solution: C

In such case you can use either dimensionality reduction algorithm or the feature selection algorithm

13. **When you find noise in data which of the following option would you consider in k-NN?**

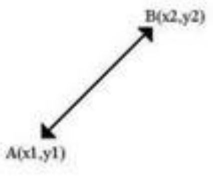
A) I will increase the value of k
B) I will decrease the value of k
C) Noise cannot be dependent on value of k
D) None of these

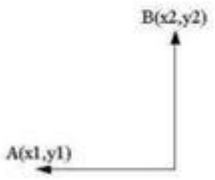
Solution: A

To be more sure of which classifications you make, you can try increasing the value of k.

14. The following two distances (Euclidean Distance and Manhattan Distance) have given to you which generally we used in K-NN algorithm. These distance are between two points A(x1,y1) and B(x2,Y2).

Your task is to tag the both distance by seeing the following two graphs. Which of the following option is true about below graph ?





A) Left is Manhattan Distance and right is euclidean Distance
B) Left is Euclidean Distance and right is Manhattan Distance
C) Neither left or right are a Manhattan Distance
D) Neither left or right are a Euclidian Distance
Solution: B

Left is the graphical depiction of how Euclidean distance works, whereas right one is of Manhattan distance.

15. **Which of the following will be true about k in k-NN in terms of variance?**

A) When you increase the k the variance will increases
B) When you decrease the k the variance will increases
C) Can’t say
D) None of these

Solution: B

Simple model will be consider as less variance model

16. **Which of the following will be true about k in k-NN in terms of Bias?**

A) When you increase the k the bias will be increases
B) When you decrease the k the bias will be increases
C) Can’t say
D) None of these

Solution: A

large K means simple model, simple model always consider as high bias

17. **Which of the following distance metric cannot be used in k-NN?**

A) Manhattan
B) Minkowski
C) Mahalanobis
D) All can be used

Solution: D

All of this distance metric can be used as a distance metric for k-NN.

18. **Which of the following option is true about k-NN algorithm?**

A) It can be used for classification
B) It can be used for regression
C) It can be used in both classification and regression

D) None

Solution: C

We can also use k-NN for regression problems. In this case the prediction can be based on the mean or the median of the k-most similar instances.

19. **Which of the following statement is true about k-NN algorithm?**

1. k-NN performs much better if all of the data have the same scale
2. k-NN works well with a small number of input variables (p), but struggles when the number of inputs is very large
3. k-NN makes no assumptions about the functional form of the problem being solved

A) 1 and 2
B) 1 and 3
C) Only 1
D) All of the above

Solution: D

The above mentioned statements are assumptions of kNN algorithm

20. Which of the following machine learning algorithm can be used for imputing missing values of both categorical and continuous variables?

A) K-NN
B) Linear Regression
C) Logistic Regression

D) All the above

Solution: A

k-NN algorithm can be used for imputing missing value of both categorical and continuous variables.

21. **Which of the following is true about Manhattan distance?**

A) It can be used for continuous variables
B) It can be used for categorical variables
C) It can be used for categorical as well as continuous
D) None of these

Solution: A

Manhattan Distance is designed for calculating the distance between real valued features.

22. **Which of the following distance measure do we use in case of categorical variables in k-NN?**

1. Hamming Distance
2. Euclidean Distance
3. Manhattan Distance

A) 1
B) 2
C) 3
D) 1 and 2
E) 2 and 3
F) 1, 2 and 3

Solution: A

Both Euclidean and Manhattan distances are used in case of continuous variables, whereas hamming distance is used in case of categorical variable.

23. **Which of the following will be Euclidean Distance between the two data point A (1,3) and B(2,3)?**

A) 1
B) 2
C) 4
D) 8

Solution: A

sqrt((1-2)^2 + (3-3)^2) = sqrt(1^2 + 0^2) = 1

24. **Which of the following will be Manhattan Distance between the two data point A (1,3) and B(2,3)?**

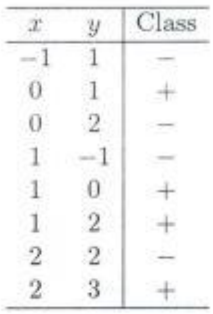
A) 1
B) 2
C) 4
D) 8

Solution: A

sqrt(mod((1-2)) + mod((3-3))) = sqrt(1 + 0) = 1

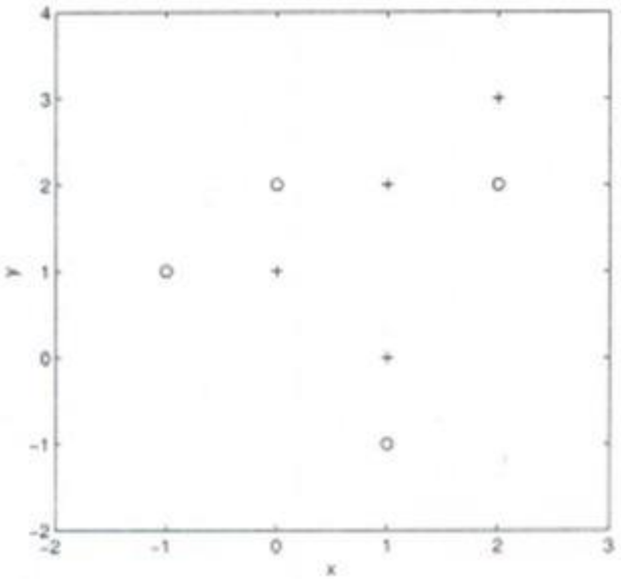
Q 25 -26

Suppose, you have given the following data where x and y are the 2 input variables and Class is the dependent variable.



x	y	Class
-1	1	$-$
0	1	$+$
0	2	$-$
1	-1	$-$
1	0	$+$
1	2	$+$
2	2	$-$
2	3	$+$

Below is a scatter plot which shows the above data in 2D space.



25) Suppose, you want to predict the class of new data point x=1 and y=1 using Euclidian distance in 3-NN. In which class these data points belong to?

A) + Class
B) – Class

C) Can’t say

D) None of these

Solution: A

All three nearest point are of +class so this point will be classified as +class.

26) In the previous question, you are now want use 7-NN instead of 3-KNN which of the following x=1 and y=1 will belong to?

A) + Class
B) – Class

C) Can’t say

Solution: B

Now this point will be classified as – class because there are 4 – class and 3 +class point are in nearest circle.

Assignment Questions

Question 1)Data of a questionnaire survey and objective testing for two attributes acid durability and strength to test if special paper tissue is good or not. The four training samples are given below

X1 acid durability (secs)	X2 strength Kg/sq meter	Y= classification
7	7	bad
7	4	bad
3	4	good
1	4	good

Now the factory produces a new paper tissue that pass laboratory test with x1= 3 and x2=7.

Without any expensive survey, can we classify what the classification of new tissue is?

Ans) 1. Determine the parameter k= number of nearest neighbours

Suppose we use k=3

2. Calculate the distance between the query-instance and the training samples

X1 acid durability (secs)	X2 strength Kg/sq meter	square of distance to query-instance
7	7	$(7-3)^2 + (7-7)^2 = 16$
7	4	$(7-3)^2 + (4-7)^2 = 25$
3	4	$(3-3)^2 + (4-7)^2 = 9$
1	4	$(1-3)2 + (4-7)^2 = 13$

3.Sort the distance and find the nearest neighbour based on the k-th minimum distance

X1 acid durability (secs)	X2 strength Kg/sq meter	square of distance to query-instance	Rank Min dist	Is it incl In 3 NN
7	7	$(7-3)^2 + (7-7)^2 = 16$	3	Yes
7	4	$(7-3)^2 + (4-7)^2 = 25$	4	No
3	4	$(3-3)^2 + (4-7)^2 = 9$	1	Yes
1	4	$(1-3)^2 + (4-7)^2 = 13$	2	Yes

.4. Gather the category of the nearest neighbours. The second row will not be included as the rank is more than 3(=k)

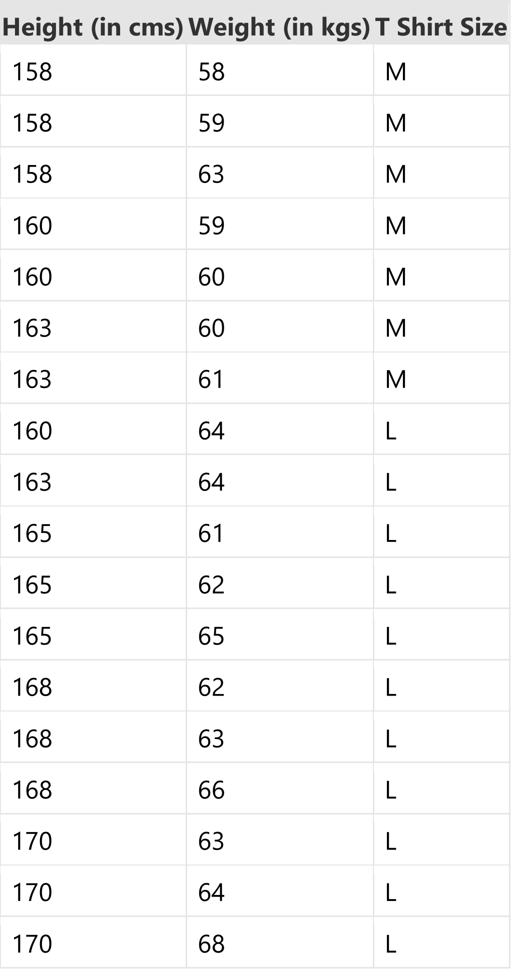
X1 acid durability (secs)	X2 strength Kg/sq meter	square of distance to query-instance	Rank Min dist	Is it incl In 3 NN	Y= category of NN
7	7	$(7-3)^2 + (7-7)^2 = 16$	3	Yes	Bad
7	4	$(7-3)^2 + (4-7)^2 = 25$	4	No	--
3	4	$(3-3)^2 + (4-7)^2 = 9$	1	Yes	Good

1	4	$(1-3)^2 + (4-7)^2 = 13$	2	Yes	Good

. 5.Use majority of the category of the nearest neighbours as prediction value of the query instance

We have 3 Good and 1 bad and so we conclude that the new paper tissue that pass the laboratory test with x1=3 and x2=7 is included in the Good category

Question 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. Data including height, weight and T-shirt size information is shown below -

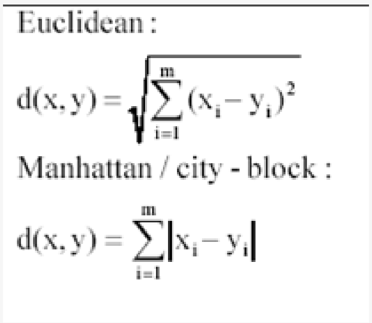


Height (in cms)	Weight (in kgs)	T Shirt Size
158	58	M
158	59	M
158	63	M
160	59	M
160	60	M
163	60	M
163	61	M
160	64	L
163	64	L
165	61	L
165	62	L
165	65	L
168	62	L
168	63	L
168	66	L
170	63	L
170	64	L
170	68	L

Ans)

Step 1 : **Calculate Similarity based on distance function**

There are many distance functions but Euclidean is the most commonly used measure. It is mainly used when data is continuous. Manhattan distance is also very common for continuous variables.



Euclidean :

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

Manhattan / city - block :

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

Distance Functions

The idea to use distance measure is to find the distance (similarity) between new sample and training cases and then finds the k-closest customers to new customer in terms of height and weight.

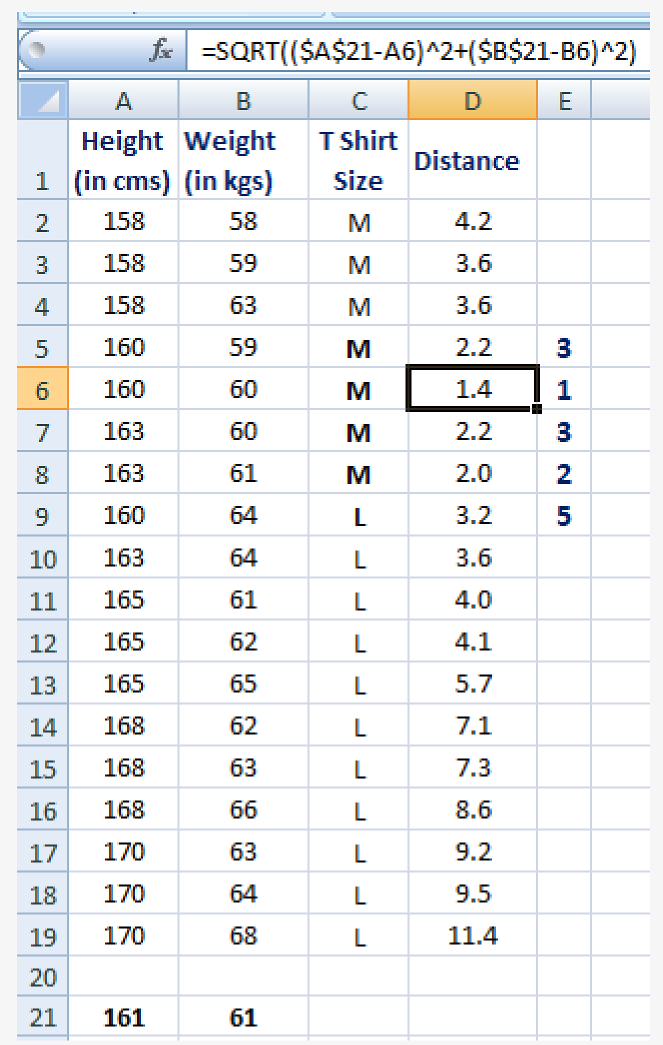
New customer named 'Monica' has height 161cm and weight 61kg.

Euclidean distance between first observation and new observation (monica) is as follows -

=SQRT((161-158)^2+(61-58)^2)
Similarly, we will calculate distance of all the training cases with new case and calculates the rank in terms of distance. The smallest distance value will be ranked 1 and considered as nearest neighbor.

Step 2 : **Find K-Nearest Neighbors**

Let k be 5. Then the algorithm searches for the 5 customers closest to Monica, i.e. most similar to Monica in terms of attributes, and see what categories those 5 customers were in. If 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. See the calculation shown in the snapshot below -



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

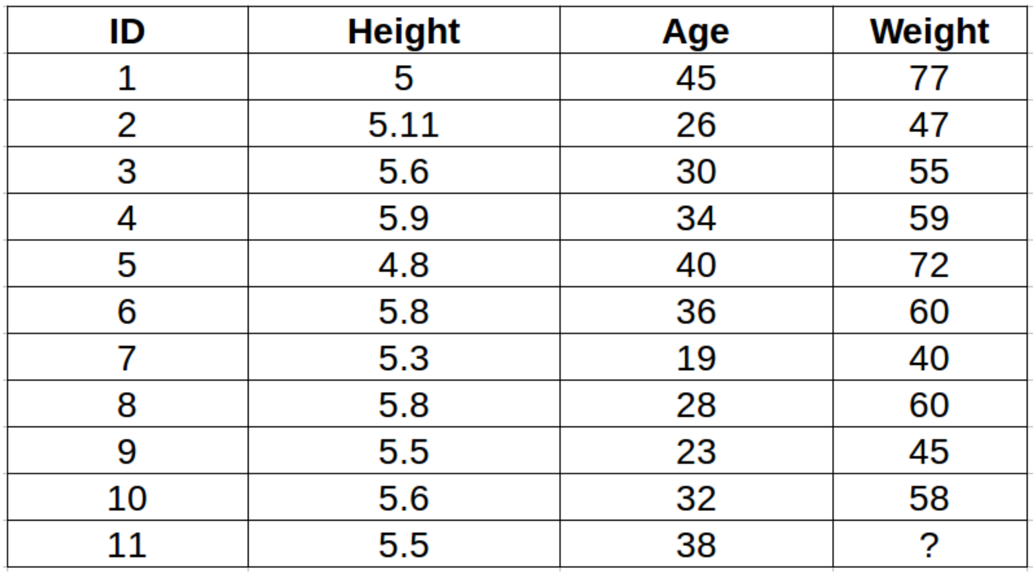
	A	B	C	D	E
1	Height (in cms)	Weight (in kgs)	T Shirt Size	Distance	
2	158	58	M	4.2	
3	158	59	M	3.6	
4	158	63	M	3.6	
5	160	59	M	2.2	3
6	160	60	M	1.4	1
7	163	60	M	2.2	3
8	163	61	M	2.0	2
9	160	64	L	3.2	5
10	163	64	L	3.6	
11	165	61	L	4.0	
12	165	62	L	4.1	
13	165	65	L	5.7	
14	168	62	L	7.1	
15	168	63	L	7.3	
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**			

Calculate KNN manually

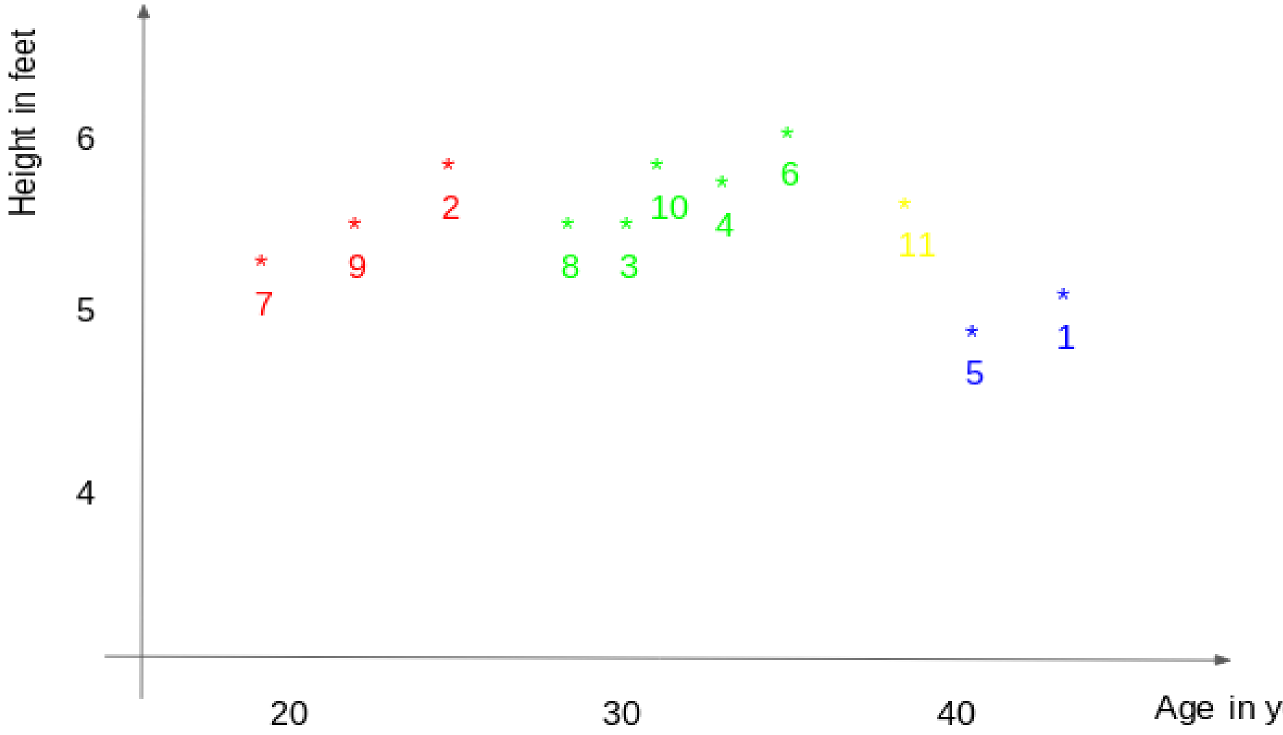
In the graph below, binary dependent variable (T-shirt size) is displayed in blue and orange color. 'Medium T-shirt size' is in blue color and 'Large T-shirt size' in orange color. New customer information is exhibited in yellow circle. Four blue highlighted data points and one orange highlighted data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

So the prediction for the new case is blue highlighted data point which is Medium T-shirt size.

Question 3) . Consider the following table – it consists of the height, age and weight (target) value for 10 people. As you can see, the weight value of ID11 is missing. Predict the weight of this person based on their height and age.

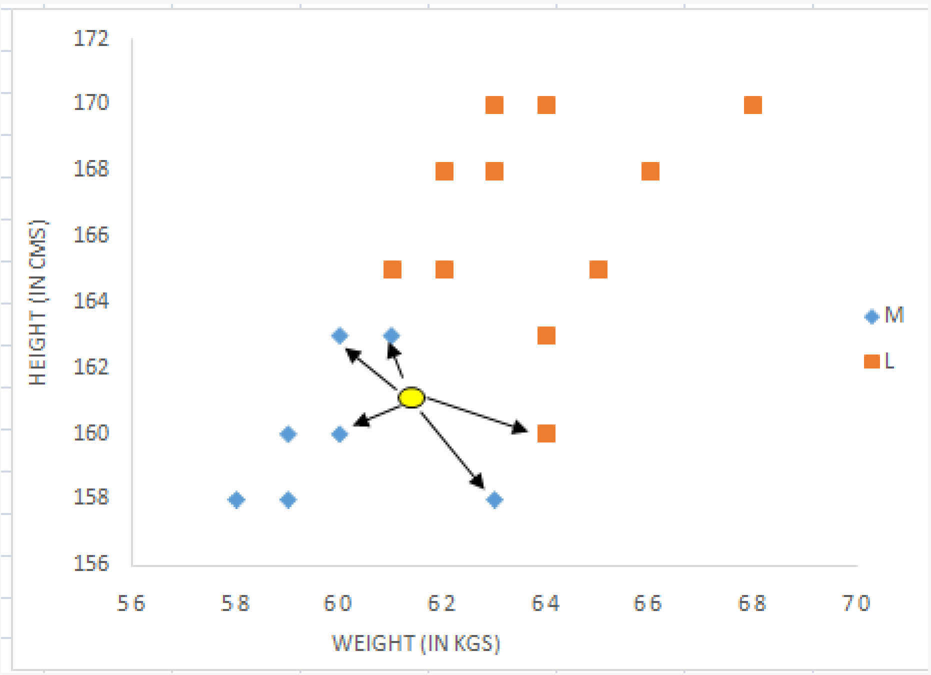


ID	Height	Age	Weight
1	5	45	77
2	5.11	26	47
3	5.6	30	55
4	5.9	34	59
5	4.8	40	72
6	5.8	36	60
7	5.3	19	40
8	5.8	28	60
9	5.5	23	45
10	5.6	32	58
11	5.5	38	?



In the above graph, the y-axis represents the height of a person (in feet) and the x-axis represents the age (in years). The points are numbered according to the ID values. The yellow point (ID 11) is our test point.

since ID11 is closer to points 5 and 1, so it must have a weight similar to these IDs, probably between 72-77 kgs (weights of ID1 and ID5 from the table



KNN: Visual Representation

.