Assignment 6. (**For all questions, use Euclidian distance**)

Q1[2]. What are random forests? Explain how they work. What is the advantage of using a random forest over a decision tree?

'Random forests' in machine learning refer to an ensemble of decision trees that are created by randomly selecting a subset of training data and creating a decision tree out of them.

'Random forests' are used in the process of classification. Once the forests are created, bootstrap samples (also known as bagging) are drawn with replacement from the original set of training data. Each sample is used to train a separate decision tree model. As these decision trees grow, a random subset of the features is considered at each split instead of evaluating all features in order to introduce randomness and improve overall accuracy. These decision trees are allowed to grow to a maximum specified depth in order to capture complex patterns and avoid overfitting the data. From there, the random forest models arrive at a single prediction for a new instance by comparing all of the predictions from the individual trees and selecting the highest category percentage.

The advantage of using a random forest over a decision tree, as stated above, is to reduce overfitting the model by introducing random noise/reducing variance as well as creating a higher overall accuracy by feature selection with low-bias.

Q2[6]. Consider the following data.

ID (not used for clustering)	age	salary (in thousands)
1	23	22
2	33	50
3	40	80
4	11	5
5	70	30

Show two iterations of applying k-means clustering to the data. Use k = 3. Pick records 1, 3, and 5 as initial centroids.

Data Points - (23, 22); (33, 50); (40, 80); (11, 5); (70, 30)

Initial Centroids - (23, 22); (40, 80); (70, 30)

Round #1

C1:

ID 2: sqrt((23 - 33)^2 + (22 - 50)^2) = 29.7

ID 4: sqrt((23 - 11)^2 + (22 - 5)^2) = 20.8

C2:

ID 2: sqrt((40 - 33)^2 + (80 - 50)^2) = 30.8

ID 4: sqrt((40 - 11)^2 + (80 - 5)^2) = 80.4

C3:

ID 2: sqrt((70 - 33)^2 + (30 - 50)^2) = 42.1

ID 4: sqrt((70 - 11)^2 + (30 - 5)^2) = 64.1

Clusterings:

C1: (23, 22); (33, 50); (11, 5)

C2: (40, 80);

C3: (70, 30)

Round #2

Changed Centroids - (22.33, 25.67); (40, 80); (70, 30)

C1:

ID 1: sqrt((22.33 - 23)^2 + (25.67 - 22)^2) = 3.73

ID 2: sqrt((22.33 - 33)^2 + (25.67 - 50)^2) = 26.6

ID 4: sqrt((22.33 - 11)^2 + (25.67 - 5)^2) = 23.6

C2:

ID 1: sqrt((40 - 23)^2 + (80 - 22)^2) = 60.5

ID 2: sqrt((40 - 33)^2 + (80 - 50)^2) = 30.8

ID 4: sqrt((40 - 11)^2 + (80 - 5)^2) = 80.4

C3:

ID 1: sqrt((70 - 23)^2 + (30 - 22)^2) = 47.7

ID 2: sqrt((70 - 33)^2 + (30 - 50)^2) = 42.1

ID 4: sqrt((70 - 11)^2 + (30 - 5)^2) = 64.1

Clusterings:

C1: (23, 22); (33, 50); (11, 5)

C2: (40, 80);

C3: (70, 30)

Q3[2] What are the main weaknesses of k-means clustering?

One main weakness of k-means clustering is initially determining the k value. Since there is no way to know the optimal k beforehand, it has to be determined through a rough estimate or randomly. However, if the wrong k is chosen, there may be varying clusterings and classification results produced.

Another main weakness of k-means clustering is the placement of centroids. Since the initial placement of the centroids is also random, there may also be varying clusterings and classification results produced. This can especially occur when outliers are present within the data, expanding the cluster size and empty space within them due to the calculated distance between points.

Q4[6]. Apply agglomerative (bottom-up) hierarchical clustering to the data from Q2. Show the steps and the final result. Use the **single-link method** to compute the distance between two clusters.

Data Points - P1 (23, 22); P2 (33, 50); P3 (40, 80); P4(11, 5); P5(70, 30)

Euclidean Distance Between Each Point:

d(P1, P2)	(23, 22); (33, 50) → sqrt((23 - 33)^2 + (22 - 50)^2) = 29.73
d(P1, P3)	(23, 22); (40, 80) → sqrt((23 - 40)^2 + (22 - 80)^2) = 60.44
d(P1, P4)	(23, 22); (11, 5) → sqrt((23 - 11)^2 + (22 - 5)^2) = 20.80
d(P1, P5)	(23, 22); (70, 30) → sqrt((23 - 70)^2 + (22 - 30)^2) = 47.67
d(P2, P3)	(33, 50); (40, 80) → sqrt((33 - 40)^2 + (50 - 80)^2) = 30.80
d(P2, P4)	(33, 50); (11, 5) → sqrt((33 - 11)^2 + (50 - 5)^2) = 50.01
d(P2, P5)	(33, 50); (70, 30) → sqrt((33 - 70)^2 + (50 - 30)^2) = 42.06
d(P3, P4)	(40, 80); (11, 5); → sqrt((40 - 11)^2 + (80 - 5)^2) = 80.44
d(P3, P5)	(40, 80); (70, 30) → sqrt((40 - 70)^2 + (80 - 30)^2) = 58.30
d(P4, P5)	(11, 5); (70, 30) → sqrt((11 - 70)^2 + (5 - 30)^2) = 64.08

	P1	P2	P3	P4	P5

P1	0				
P2	29.73	0			
P3	60.44	30.80	0		
P4	20.80	50.01	80.44	0	
P5	47.67	42.06	58.30	64.08	0

Smallest Difference {P5, P3} - 12.77

	P1	P2	P3	P4	P5
P1	0				
P2	29.73	0			
{P3, P5}	60.44	30.80	0		
P4	20.80	50.01	80.44	0	

Smallest Difference {P4, P2} - 12.77

	P1	P2	P3	P4	P5
P1	0				
{P2, P4}	29.73	0			
{P3, P5}	60.44	30.80	0		

P1

|

| |

{P2, P4} {P3, P5}

Q5[3]. Consider the following two data points:

Age	Salary	Number of friends	Own a house
Young	High	A lot	Yes
young	medium	Not many	yes

How can you compute the distance between the two data points? Show all methods covered in class. Show your work and the actual distance as a number.

Age	Salary	Number of friends	Own a house
1	3	3	Yes
1	2	1	yes

A. Euclidean Distance : sqrt((x1 - x2)^2 + (y1 - y2)^2 + (z1 - z2)^2)

→ sqrt((1 - 1)^2 + (3 - 2)^2 + (3 - 1)^2)) = 2.24

B. Manhattan Distance : |x1 - x2 | + |y1 - y2| + |z1 - z2|

→ |1 - 1 | + |3 - 2| + |3 - 1| = 3.0

C. Minkowski Distance : sqrt((|x1 - x2|)^2 + (|y1 - y2|)^2 + (|z1 - z2|)^2)

→ sqrt((|1 - 1|)^2 + (|3 - 2|)^2 + (|3 - 1|)^2) = 2.24

Q6[4] Consider a classifier that classified documents as follows:
C1 :{1,3,5} C2:{2,4,6}, C3{7,8}. Suppose the correct classification, as determined by humans, is: C1:{1,3,6} C2:{2,4,5},C3{7,8}. What is the **entropy** and **purity** of the algorithm on the example? Show your work.

Note:

- Entropy = – Σ (pi * log2 pi)
- Proportion : (TP + FN) / (TP + FN + FP + TN)
- Purity: (TP + TN) / ((TP + FN + FP + TN)

Cluster #1 (C1)

TP (True Positive): 2 (correct numbers included being 1 and 3)
TN (True Negative): 4 (correct numbers discluded being 2, 4, 7, 8)
FP (False Positive): 1 (wrong number being 5)
FN (False Negative): 1 (number needed being 6)
Proportion: (2 + 1) / (2 + 1 + 1 + 4) = 0.375
Entropy: - (0.375 * log2(0.375)) - (0.625 * log2(0.625)) = 0.95
Purity: (2 + 4) / (2 + 1 + 1 + 4) = 0.75

Cluster #2 (C2)
TP (True Positive): 2 (correct numbers included being 2 and 4)
TN (True Negative): 4 (correct numbers discluded being 1, 3, 7, 8)
FP (False Positive): 1 (wrong number being 6)
FN (False Negative): 1 (number needed being 5)
Proportion: (2 + 1) / (2 + 1 + 1 + 4) = 0.375
Entropy: - (0.375 * log2(0.375)) - (0.625 * log2(0.625)) = 0.95
Purity: (2 + 4) / (2 + 1 + 1 + 4) = 0.75

Cluster #3 (C3)
TP (True Positive): 2 (correct numbers included being 7 and 8)
TN (True Negative): 6 (correct numbers discluded being 1, 2, 3, 4, 5, 6)
FP (False Positive): 0 (no wrong numbers)
FN (False Negative): 0 (no number needed)
Proportion: (2 + 0) / (2 + 0 + 0 + 6) = 0.25
Entropy: - (0.25 * log2(0.25)) - (0.75 * log2(0.75)) = 0.81
Purity: (2 + 6) / (2 + 1 + 1 + 4) = 1.0

Overall Algorithm Entropy
Entropy:
= (C1 Proportion * C1 Entropy) + (C2 Proportion * C2 Entropy) + (C3 Proportion * C3 Entropy)
= (0.375 * 0.95) + (0.375 * 0.95) + (0.25 * 0.81)
= 0.915

Overall Algorithm Purity
Purity:
= (C1 Proportion * C1 Purity) + (C2 Proportion * C2 Purity) + (C3 Proportion * C3 Purity)
= (0.375 * 0.75) + (0.375 * 0.75) + (0.25 * 1.0)
= 0.813