Họ và tên: Nguyễn Minh Tâm

MSSV: 1952968

Question 1

a) Classification belongs to supervised learning.

b) The main difference between supervised learning and unsupervised learning: labeled data. Supervised learning uses labeled input and output data, while an unsupervised learning algorithm does not.
In supervised learning, the algorithm learns from the training dataset by iteratively making predictions on the data and adjusting for the correct answer. It requires upfront human intervention to label the data appropriately.
Unsupervised learning models work on their own to discover the inherent structure of unlabeled data. They still require some human intervention for validating output variables.

c) Training error: the error that you will get when you run the trained model back to the training data that was used before.
Testing error: the error you will get when you run the trained model on a completely different set of data that it has never been exposed to.

d) Model overfitting is a concept in data science, which occurs when a statistical model fits exactly against its training data. When a model trains for too long on sample data or when the model is too complex, it can start to learn the noise or irrelevant information within the dataset. When the model memorizes the noise and fits too closely to the training set, the model becomes overfitted and is unable to generalize well to new data. As a result, it will not be able to perform the classification or prediction tasks that it wass intended for.

Question 2

a)

	D1	D2	D3	D4
t2	1	1	2	1
t3	0	1	1	0
t5	0	0	0	0

b) In D3, there are 2 words appearing in dictionary: Data and Structure. To determine which word is more importance, I will calculate the idf of each word.
First calculating the document frequency of each word (df)

	t2 (Data)	t3 (Structure)
df	4	2

Then I calculate the inverse document frequency (idf = $\log\frac{|C|}{|Ct|}$)

	t2 (Data)	t3 (Structure)
idf	0	0.69

As you can see, t3 (Structure) has higher idf than t2 -> Choose t3 to be the keyword.

c) When a term occurs very frequently in many documents, it means that it has less discriminatory power. Considering the idf equation:

$$\text{idf} = \log\frac{|C|}{|Ct|}$$

From this equation, if a term occurs very frequently, value of Ct will be big

➔ idf will be close to 0.

➔ ‘the’ term is just a redundant like a stop-word.

d) idf(t2, C) = 0
idf(t3, C) = 1

e) tf-idf(t2, D3, C) = tf(t2, D3) * idf(t2, C) = 2 * 0 = 0
tf-idf(t3, D3, C) = tf(t3, D3) * idf(t3, C) = 1 * 1 = 1

f) Term occurs frequently -> idf(t, C) = 0 -> tf-idf(t, D, C) = 0.
If tf-idf(t, D, C) = 0, the term is not important. In other words, the term is a redundant.

Question 3

a) A1 = 1600 + 1800 + 2000 + 2600 = 8000
A2 = 1680 + 2160 + 2400 + 3360 = 9600
F(2019) = (A1 + A2) / 2 = (8000 + 9600) / 2 = 8800
The predicted total sales for 2019 is 8800$.

b) Total sales for 2017 = 8000
Total sales for 2018 = 9600
Predicted total sales for 2019 = 8800.
In general, we would expect the total sales for 2019 gets increased. However, if we use MA method to forecast, we cannot catch up the trend because MA method takes the average of total sales in the past, so the result cannot be bigger than the past total sales.

c) The average seasonal sales for 2017 = (1600 + 1800 + 2000 + 2600) / 4 = 2000
The average seasonal sales for 2018 = (1680 + 2160 + 2400 + 3360) / 4 = 2400

d)

Quarter	2017	Seasonal index	2018	Seasonal index	Average index	2019
Spring	1600	0.8	1680	0.7	0.75	2025

Summer	1800	0.9	2160	0.9	0.9	2430
Fall	2000	1	2400	1	1	2700
Winter	2600	1.3	3360	1.4	1.35	3645
Average	2000		2400			2700

Question 4

a)

Vertex	Centrality	Ranking
A	3	3
B	4	2
C	3	3
D	4	2
E	5	1
F	5	1
G	2	4

b)

1. Cliques with 3 vertices: {A, B, C}, {B, C, D}, {D, E, F}
2. {A, B, C} & {B, C, D} share two vertices (B & C) -> {A, B, D, C}
3. {A, B, D, C} & {D, E, F}

c) Looking to this graph, I figure out there are 2 main communities: {1, 2, 4, 3} & {5, 6, 7, 8}.
So we should select 1 node of each community to make sure that all people will buy the product.
We would like to choose 2 nodes that can directly make their neighbors to buy the product. In this case, we choose node 3 & 6. By choosing these 2 nodes, node 5 will buy our product right away. This leads to node 7 and then node 8. Now we have finished the below community.
With the above community, clearly after node 8 buys our product, node 4 will buy the product too because 2/3 of its neighbors have bought our products (node 3 & 8). Finally node 1 & 2 will buy our product.