Question 1

a) Data modelling often applies machine learning techniques. Machine learning can be divided into two types: supervised learning and unsupervised learning. Classification and clustering are the two most basic machine learning techniques. Which type does classification belong to?

b) Explain the main difference between supervised learning and unsupervised learning.

c) Classification involves two phases: training phase and testing phase. Training phase uses training datasets to establish data models. Testing phase applies testing datasets to examine the established data models. Briefly explain what are training errors and testing errors (a.k.a generalisation error).

d) The data models established in machine learning are unlikely to be perfect, i.e., they can be underfitted or overfitted. Briefly explain what is model overfitting.

Question 2

Consider the following document collection C which has 4 documents: $C = \{D_1, D_2, D_3, D_4\}$.

Document	Words
D_1	Communication and Data Security, Communication Technology
D_2	Introduction to Algorithms in Data Science, Discrete Structure
D_3	Data Structures and Algorithm Analysis on Massive Data Sets
D_4	Foundations of Data Base System, Computer System

Let Dict be a dictionary which consists of 5 terms (words): Dict = {t_1 = Communication, t_2 = Data, t_3 = Structure, t_4 = System, t_5 = Calculus}.

a) Denote by $\mathsf{tf}(t, D)$ the term frequency (TF) of the term t in the document D. Please fill out the blank cells in the following table, i.e., give the values $\mathsf{tf}(t_i, D_j)$ for $i = 2, 3, 5$ and $1 \leq j \leq 4$. The value $\mathsf{tf}(t_i, D_j)$ should be put into the cell specified by t_i and D_j. (**Note**, you can copy the form to your answer sheet and the fill it out.)

	D_1	D_2	D_3	D_4
t_2				
t_3				
t_5				

b) Which word would you like to set as a keyword for document D_3? Why?

c) In normal English documents, the term “the” occurs very frequently. However, “the” is certainly not suitable for being a keyword for any single document. Can you briefly explain why is that?

d) Recall the inverse document frequency (IDF) is defined as $\text{idf}(t, C) = \log_2(\frac{|C|}{|C_t|})$. Here $|C|$ denotes the number of documents in the collection C, $|C_t|$ denotes the number of documents from C that contains term t, and 2 is the logarithm base. Please compute $\text{idf}(t_2, C)$ and $\text{idf}(t_3, C)$. (You might want to use those values: $\log_2 4 = 2$, $\log_2(2) = 1$, $\log_2(\frac{4}{3}) = 0.415$, $\log_2(1) = 0$.)

e) Term frequency – inverse document frequency (TF-IDF) takes both term frequency (TF) and inverse document frequency (IDF) into consideration in identifying keywords. For the document collection C, the term t's TF-IDF value on document D_i is defined as $\text{tf-idf}(t, D_i, C) = \text{tf}(t, D_i) \times \text{idf}(t, C)$, i.e., the product of t's TF value on D_i and t's IDF value. Please compute $\text{tf-idf}(t_2, D_3, C)$ and $\text{tf-idf}(t_3, D_3, C)$.

f) Using TF-IDF prevents us from identifying “trivially frequent” terms like “the” as keywords of a text document from a given document collection. Please use the formulas of TF-IDF, i.e., $\text{idf}(t, C) = \log_2(\frac{|C|}{|C_t|})$ and $\text{tf-idf}(t, D, C) = \text{tf}(t, D) \times \text{idf}(t, C)$ to briefly explain the reason.

Question 3

A company has collected quarterly sales for the past two years which are shown in the following table. The company wants to forecast the next year's seasonal sales.

Index	Time	Sales ($)	Index	Time	Sales ($)
1	Spring 2017	1600	5	Spring 2018	1680
2	Summer 2017	1800	6	Summer 2018	2160
3	Fall 2017	2000	7	Fall 2018	2400
4	Winter 2017	2600	8	Winter 2018	3360

a) Let A_1 and A_2 be the actual total sales (i.e., the sum of all four seasonal sales) in 2017 and 2018, respectively. Assume current time is t, the n-moving average (MV) technique makes forecast for the time $t+1$ by taking the average of previous n actual values where $n \leq t$. (The formula can be written as $F_{t+1} = \frac{\sum_{i=t-n+1}^{t} A_i}{n}$ where A_i is the i-th actual data). Predict the total sales for 2019 using MV with 2 actual values A_1 and A_2.

b) In general, we would expect the total sales gets increased in both 2018 and 2019 if the economy situation has been keeping going well since 2017. Use this and your

answer to sub-question **a)** to explain the limitation of moving average (MV) method in forecasting.

c) Calculate the average seasonal sales for both 2017 and 2018.

d) Here are the steps of forecasting with seasonality. Please follow the steps to fill out the blank cells in the following form.

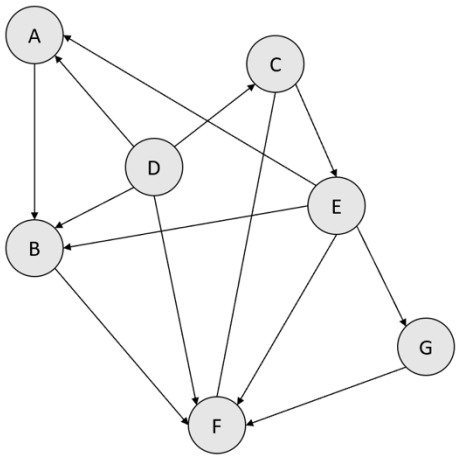
1. Calculate the average seasonal sales for each year;
2. Calculate each seasonal index (by dividing the actual seasonal sales by the average seasonal sales);
3. Compute the average indexes;
4. Predict the average seasonal sales for the next year (i.e., 2019);
5. Multiple next year's average seasonal sales by each average seasonal index.

(**Note**, you have already done the step 1) by answering sub-question c) and please put those values into the table. Also, step 4 is already done for you (the red value). You can copy the form to your answer sheet then fill blank cells).

Quarter	2017	Seasonal	Sales	Seasonal	Average	2019
Spring	1600		1680			
Summer	1800		2160			
Fall	2000		2400			
Winter	2600		3360			
Average						2700

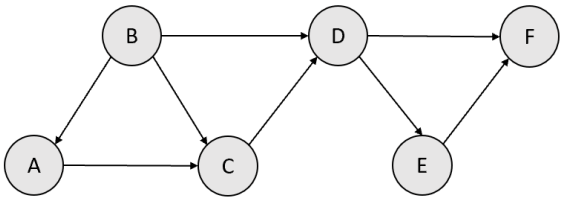
Question 4

a) Degree centrality is the most basic centrality metric. Find the degree centrality of each vertex in the graph and rank the vertices using their degree centrality. Fill out the following table. (**Note**, you can copy the form to your answer sheet then fill blank cells.)



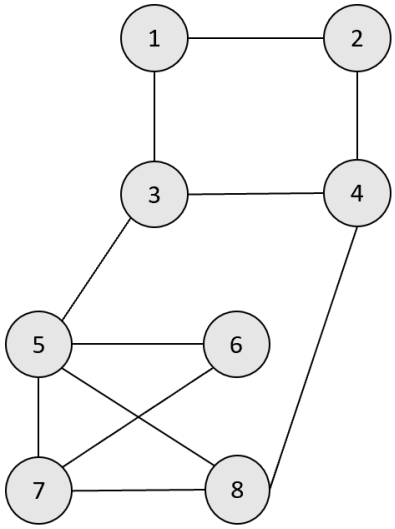
Vertex	Centrality	Ranking

b) Clique Percolation Method (CPM) is used to discover communities for network data. Below is the graph representation of a given network dataset. Your tasks include: 1) write down all cliques with three vertices; 2) connecting any two cliques if they share with two vertices; 3) output the communities (as sets of vertices) you discovered.



c) You are trying to sell your product to a group of people of which you have obtained the connection network, as depicted below. Suppose you can only convince at most two people from the network to buy your product. People in this network will buy a product if

≥50% of their neighbours have bought one. Please select two people from the network so that eventually all people will buy your product and explain why that is the case.



END OF EXAM