CSC3170 Assignment
This is an individual assignment and should be submitted by 5 pm 20 May 2022 via Blackboard

Question 1

A Contingency Table for X → Y is defined as follows (where X' signifies Not X, and likewise for Y):

	Y	Y'	
X	f_{11}	f_{10}	f_{1+}
X'	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	|T|

f_{11}: support of X and Y

f_{10}: support of X and Y'

f_{01}: support of X' and Y

f_{00}: support of X' and Y'

Suppose we are given the following contingency table

	Coffee	Coffee'	
Tea	15	5	20
Tea'	75	5	80
	90	10	100

(a) Determine if confidence is a useful measure for the Rule

Tea → Coffee

Now, consider another measure called Lift, defined as follows.

$$Lift(X,Y) = \frac{P(Y \mid X)}{P(Y)}$$

(b) Comment on this measure for the cases of Lift=1, Lift>1, and Lift<1.

(c) Calculate the Lift of the contingency table of Tea and Coffee, and comment on its usefulness in relation to confidence.

Question 2

An international shipping company is planning to optimize its delivery service by forming clusters of its consignments. A sample set of consignments is given in the following table.

Consignment	Volume	Weight
1	8	4
2	5	4
3	2	4
4	2	6
5	2	8
6	8	6

Use the *k*-means algorithm to cluster the data, and assume that $k = 3$, with Consignments 1, 3, and 5 used for the initial cluster centroids.

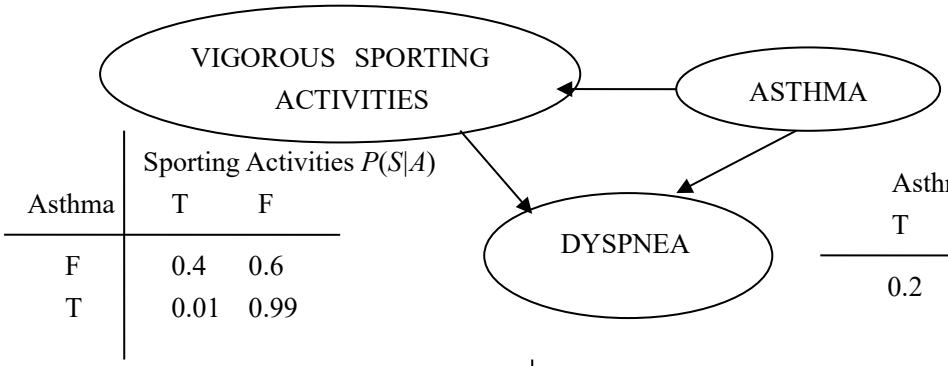
Provide **three** solutions to form three clusters of the consignments.

Questions 3 and 4 are related to Bayesian Networks

Bayesian networks are often used for data mining. Bayes' Rule helps to express conditional probabilities—the likelihood of one event given that another has occurred; i.e. the conditional probability of event A given event B is given by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

A Bayesian network is a way to put Bayes' Rule to work by laying out graphically which events influence the likelihood of occurrence of other events. The following figure shows a Bayesian network for the diagnosis of breathing difficulty (dyspnea).



Asthma	Sporting Activities $P(S\|A)$ T	F
F	0.4	0.6
T	0.01	0.99

Asthma $P(A)$ T	F
0.2	0.8

Sporting Activities	Asthma	Breathing Difficulty $P(B\|S,A)$ T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

Suppose that there are two events which could lead to dyspnea:

- the person with the condition *asthma*, and experiences an asthma attack
- after vigorous sporting activities/physical exertion

Also, suppose that an asthmatic person may also undertake vigorous sporting activity. Then the situation can be modelled with Bayesian network as indicated in the above figure. All three variables have two possible values T (for true) and F (for false). The above also gives the relevant conditional probability tables (CPT), where the events to the left of the vertical line of a table signify the events conditioned upon.

Question 3

(a) Abbreviating the names of the variables as

B = Breathing Difficulty
S = Vigorous Sporting Activities
A = Asthma,

determine for the above situation the following probabilities:

- *P (B|S, A)*
- *P (S|A)*
- *P (A)*

(b) Express *P (B, S, A)* in terms of *P (B|S,A), P (S|A)* and *P (A)*.

Question 4

(a) For the above situation, determine the numerical values of the probabilities of $P(B, S, A)$, for each truth value combination of B, S, A.

(b) Using (a) or otherwise, calculate the probability that a person has an asthma attack, given that the person experiences breathing difficulty.

Questions 5 and 6 are related to the following supermarket basket analysis situation.

Consider a database of customer transactions where a number of items are purchased. In general, the number of possible association rules in such a database is very large, giving rise to a huge amount of processing in support and confidence evaluations. In more complex associations, rules can be of the form

$$A_1 \,\&\, A_2 \,\&\, \ldots \,\&\, A_n \rightarrow B, \qquad (*)$$

where the antecedent can be a conjunction of several items, but the consequent is a single item. Consider a database of customer transactions where a number of items are purchased as shown in the table below, with the first column indicating the Transaction ID, and the second column giving the items purchased.

Transaction#	**Items List**
T100	Apple, Beer, Eggs
T200	Apple, Cake, Diaper
T300	Apple, Cake
T400	Beer, Cake
T500	Apple, Beer, Cake
T600	Apple, Beer, Cake, Diaper, Eggs

Question 5

(a) How many possible rules of the form (*) are there for the above database?

(b) In general, for a database where the item list has a total of N items, determine the total number of possible rules of the form (*).

Question 6

(a) For the above database, determine all rules having a minimum support of 50% using the *Apriori Algorithm*, giving the support of each rule.

(b) Suppose further that we are only interested in rules having a confidence of at least 70%. Determine the set of rules having minimum support of 50%, and minimum confidence of 70% for this database.