IS 665 Data Analytics

Fall 2023

Final Project

You may submit more than one file if necessary, and you may submit files of different formats (Word, Excel, Python, pdf, picture files of screenshot, etc.)

Please submit using your first name last name as the file name. For example, I would name my answer lin_lin.docx. In the event where you have more than one files, please put the question number after your name, e.g., lin_lin_Q6.docx.

Good luck! I hope that you have learned something useful through this class, and that some of the skills you will need for this final exam can help you in your work in the future.

Part 1. Statistics Basics (20 pts.)

1. (5 pts.) Here is some data about Covid-19 patients and their symptoms.

Table 1. Baseline Characteristics of Patients Infected With 2019-nCoV

	No. (%)		
	Total (N = 138)	ICU (n = 36)	Non-ICU (n = 102)
Signs and symptoms			
Fever	136 (98.6)	36 (100)	100 (98.0)
Fatigue	96 (69.6)	29 (80.6)	67 (65.7)
Dry cough	82 (59.4)	21 (58.3)	61 (59.8)
Anorexia	55 (39.9)	24 (66.7)	31 (30.4)
Myalgia	48 (34.8)	12 (33.3)	36 (35.3)
Dyspnea	43 (31.2)	23 (63.9)	20 (19.6)
Expectoration	37 (26.8)	8 (22.2)	29 (28.4)
Pharyngalgia	24 (17.4)	12 (33.3)	12 (11.8)
Diarrhea	14 (10.1)	6 (16.7)	8 (7.8)
Nausea	14 (10.1)	4 (11.1)	10 (9.8)
Dizziness	13 (9.4)	8 (22.2)	5 (4.9)
Headache	9 (6.5)	3 (8.3)	6 (5.9)
Vomiting	5 (3.6)	3 (8.3)	2 (2.0)
Abdominal pain	3 (2.2)	3 (8.3)	0 (0)

a. Of the 138 patients studied, how many were admitted to ICU? How many weren't? How many of the ICU patients have fever? How many of the non-ICU patients have fever? Using these data, construct a contingency table like the following:

	ICU	Non-ICU	**Total**
Fever			
No Fever			
Total			

b. Given that a certain patient has a fever, what is the likelihood that he will be in ICU? Show your computation.

c. Given that a patient is in ICU, how likely would he/she has a fever?

2. (5 pts.) The table below displays some of the results of last summer's Fishing Festival, showing how many contestants caught n fish for various values of n

n	0	1	2	3	...	13	14	15
number of contestants who caught n fish	9	5	7	23	...	5	2	1

In the newspaper story covering the event, it was reported that
(a) the winner caught 15 fish;
(b) those who caught 3 or more fish averaged 6 fish each;
(c) those who caught 12 or fewer fish averaged 5 fish each.

What was the total number of fish caught during the festival? Please show your process.

3. (10 pts) Suppose that you are working for a phone company that wants to predict churn. Churn is when customers cancel their subscriptions. Let "*actual*" be a binary random variable meaning that the customer actually churns, and let "*flag*" be a binary random variable meaning that a classifier **predicts** that the customer will churn. Note that P(*actual* = 1), the churn base rate, is the probability that a customer would actually churn. This can be represented by the fraction of customers with "*actual*" value of 1. Similarly, P(*flag* = 1) is the fraction of customers that are predicted to churn. We know that the churn base rate is 5%. We define:

   a = P(*actual* = 1|*flag* = 1)
   b = P(*actual* = 0|*flag* = 0).

   You have hired Dr. Lin as a data mining consultant. He claims that he can train a classifier that will achieve a = 0.8 and b = 0.9.

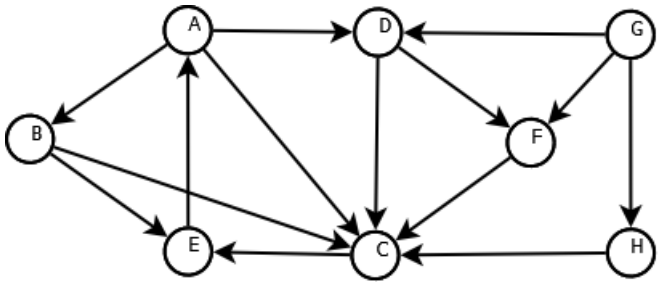
   (a) Drawn a confusion matrix. Show how to define P(*actual* = 1) and P(*flag* = 1) in terms of the entries in the confusion matrix (TP, TN, FP, FN).

   (b) Show how to define a and b in terms of the entries in the confusion matrix (TP, TN, FP, FN).

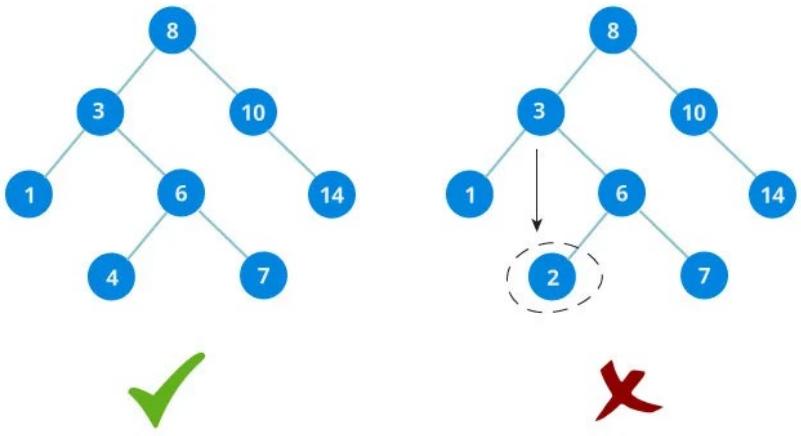
   (c) Prove Dr. Lin wrong, because no such classifier exists.

Part 2. Data Structure Basics (10 pts.)

1. (5 pts) Show the order of visit using **DFS** and **BFS** to traverse the following graph, starting at node A (this is a directed graph). When you have more than one node to choose from, always choose the letter that comes first in the alphabet.



2. (5 pts) A binary search tree is a binary tree where the left child is smaller than the parent and the right child is greater than the parent. In the diagram below, the tree to the left is a BST and the tree to the right isn't.



Using the tree on the left in the picture above (the one with a green checkmark under it), answer the following questions:

a). draw the tree after I insert the following numbers: 11, 5, 2, 13

b). Does the order of insertion matter? In other words, if I insert the numbers in a different order such as 5,11,2,13, would the resulted tree always have the same shape?

Part 3. Data Warehousing (10 pts.)

The following is the meta-data of a call center database. It records information about all calls made by the customers to call center agents regarding certain products. The agents' jobs are to take the calls, help customer figuring out what to do, and to take actions by directing customers to a certain web page where the agent and the customer could conduct online chat and run diagnostic sessions.

a. Identify the measures and constraints in the following business question: Report how often agents in different positions access web pages of each page type to perform customer support tasks for customers from each state in each month. Use attributes from the Metadata below to serve as the constraints. For example, the constraint "customer state" should be expressed as the attribute "Customer.State" (5 pts)

Measures:

Constraints:

b. Given the metadata below, design a star schema that helps you answer question (a). You need to:
   - Clearly name each dimension and fact table
   - Make sure that all keys are properly created
   - Include some attributes in each table. All attributes in your star schema must come from the metadata blow. You can choose the attributes that you think to be necessary for meaningful queries to be generated. (10 pts)

Table Name	Attributes	Description
Customer	CID Name Gender City State Country Income Level Education level Internet usage level	 Value: 1 - 10 Value: 1 - 10 Value: 1 - 10
Product	PID PName PCID	
Agent	AID AName Agender Abranch Postion Experience_Level Hiring_Date Manager_ID	
Task	TID AID CID PID Journal_Entry Begin_Time End_Time TC_ID	

Task Category	TC_ID TC_Name	
Action	AC_ID Page_URL Description	
WebPage	PageURL PageType	URL of a web page customer support page for a product, search page, product description, corporate information and order page etc.
Agent_Click	TID PageURL Entry_Time	

Part 4. Data Mining (60 pts.)

THEORY (30 points)

I. Association Rule (10 pts)

Consider the following transaction database:

TransID	Items
1	B,E
2	E,F
3	A,B,C,F
4	B,E,F
5	C,D,E,F

Suppose that minimum support is set to 40% and minimum confidence to 60%.

a) List **all frequent** (large) **TWO itemsets**, together with their support. (4 pts)

b) For all frequent itemsets of **maximal** length, list all corresponding association rules satisfying the requirements on (minimum support and) minimum confidence, together with their **confidence**. (4 pts)

c) Compute the **lift** for the association rule **A → B** and **C → F** (2 pts)

II. Clustering (10 pts)

Assume the following dataset is given:
(2,5), (3,4), (7,3), (4,4),(9,8) ,(0,4), (2,0),(5,6),(7,8),(1,1),(6,7).

K-Means is run with k=3 to cluster the dataset. Moreover, Manhattan distance is used as the distance function to compute distances between centroids and objects in the dataset.

K-Mean's initial centroids C1, C2, and C3 are as follows:

C1: (2,0)

C2: (5,3)

C3: (8,7)

Now K-means is run for a single iteration.

a. Assign each data point to one of the three centroid based on k-means algorithm. (5 pts)

Cluster 1 – C1:

Cluster 2 – C2:

Cluster 3 – C3:

b. Based on the above result, re-calculate the new centroid for each cluster (5 pts)

New Centroid for cluster 1 :

New Centroid for cluster 2 :

New Centroid for cluster 3 :

III. Classification (10 pts)

1. **Naïve Bayes (5 pts).** Below is a data set about stolen cars.

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

We have a RED DOMESTIC SUV car. Predict whether it is stolen using Naïve Bayes. Show your process.

2. Decision Trees (5 pts)

Consider the training dataset given below. In this dataset, X1, X2, and X3 are the input attributes and Y is the class variable. “+ve” and “-ve” are the two outcome classes, similar to “Yes” and “no”.

Example#	X_1	X_2	X_3	Y
E1	0	0	0	+ve
E2	0	0	1	-ve
E3	0	1	0	-ve
E4	0	1	1	+ve
E5	1	0	0	-ve

Which attribute has the highest information gain, i.e., reduction in Gini Index? Justify your answer. Use Gini Index ($1 – p_1^2 – p_2^2$) as measurement criteria

APPLICATIONS (30 points)

Case 1 (15 points):

In a research paper, 36 years of accident investigation reports from 1983 to 2018 from news websites, emergency management department and tourism websites around the world, seventy-five typical emergency cases of scenic areas were screened. The criterion of case collection is directly or indirectly related to the security of scenic areas. The scenic areas involved are located in the United States, the United Kingdom, China, Thailand, Japan, Malaysia, Singapore and various other countries. They include 9 types of accidents: traffic accidents, amusement facilities accidents, cable car accidents, natural disasters, accidents involving collapses and trampling, fire accidents, accidents involving animals attacking humans, drowning accidents and social safety accidents. The data set can be downloaded here:

https://data.mendeley.com/public-files/datasets/6pc4x652c8/files/ae76bfef-eb65-433c-a88a-3cade989e9df/file_downloaded

Of special interest to us is the question: What type of strategies are associated with certain type of mechanisms?

The following paragraphs give you background information about what "mechanisms" and "strategies" mean in this specific problem domain. THEY ONLY SERVE AS A BACKGROUND REFERENCE IN CASE YOU NEED TO UNDERSTAND THE PROBLEM BETTER.

----Reference for Association Rule Question ------

"Mechanisms" refer to the internal logic of things. Research on the mechanism of unexpected events is conducive to find out the source of events and the motive force of development, so as to take the corresponding strategies in emergency management. The evolution mechanism of emergencies includes spread mechanism, coupling mechanism and derivation mechanism.

Spread refers to the occurrence of similar incidents caused by the original event. It should be noted that new events are the same type as the original ones, such as a fire in a building that causes a fire in a nearby building. The spread can be an expansion of the spatial extent or the impact time.

According to the causes of the spread, emergencies can be divided into two types: occupancy types and transitive types. Occupancy type refers to original things which occupy the time or space of other similar things.

> **Capacity expension strategy.** This refers to increasing the amount of resources required by original things and similar things in common. When the amount of resources increases, the things that can be accommodated per unit time increases, which will reduce or no longer occupy the time of the former things occupying the latter things.

Adjustment strategy. It is an adjustment to optimize the chronological or spatial location of the original event and spread event to maximize buffer time or make better use of resources. With the using of buffer time and alternative resources to meet the needs of things, the extent of spread can be reduced. (3)

Abandonment strategy. This involves removing several things from the space-time sequence of the original or spread event in order to reduce the number of things that consume resources. The essence of abandonment strategy is to increase buffer time.

Transitive spread refers to the effect of the original things transmitted to the associate things through the material, energy or information. The condition is that things can be connected through media. Transitive spread will be prevented by the isolation strategy, resistance strategy and dredging strategy.

Isolation strategy. This entails isolating the original thing from the latter thing, to ensure that the function of the original thing cannot affect the following things.

Resistance strategy. This refers to that during process of spreading, weaken the impact of the event by reducing the energy transmitted to the latter things or increasing the resistance ability of the latter things.

Dredging strategy. This involves reducing the energy of the original things transmitted to other things by allowing this energy transmission.

Coupling stage is a phenomenon in which two or more factors interact and influence each other, eventually leading to emergencies or causing emergencies to become more severe. When the coupling factor increases the energy of the disaster source and reaches a critical value that can trigger an event, it will lead to an emergency. For an emergency that has already occurred, when the coupling factor increases, the incident will be more serious. According to the characteristics of coupling objects, it can be divided into three types: event-event coupling, event-factor coupling and factor-factor coupling. The difference between the factors and the events mentioned here is that the factors themselves are not harmful, and the events are harmful. According to the role of the coupling object, it can be divided into four types: occurrence coupling, acceleration coupling, aggravation coupling and transformation coupling. The isolation strategy, insurance strategy (including the proofreading strategy, sequential strategy, conformity strategy, buffer strategy), hedging strategy, domination strategy, reform strategy and transition strategy will be adopted to cope with the coupling of events.

Isolation strategy. This involves isolating the coupling factors from their objects to avoid the factors affecting one another. It includes spatial isolation, time isolation, media isolation etcetera. Unlike the isolation strategy in the spread mechanism, which isolates two specific things, the isolation strategy here aims at avoiding factors.

Insurance strategy. The insurance strategy includes the proofreading strategy, sequential strategy, conformity strategy and buffer strategy. 1) Proofreading strategy: This is a means to avoid errors by adding a verification procedure in each step of an event, so as to eliminate the factors caused by various errors. 2) Sequential strategy: This refers to the joint or sequential execution of more than two actions to complete the work. 3) Compatibility strategy: This uses different measures of shapes, mathematical models, and quantities to prevent human error. 4) Buffer strategy: It means to reduce the possibility of error by adding time buffers or a forced buffer.

Hedging strategy. Hedging strategy refers to applying a neutralization reaction to counteract the effect of coupling factors

Domination strategy. This uses a certain inhibitory factor to inhibit the role of factor and its target.

Reform strategy. The reform strategy refers to changing the state of characteristics of the factor itself, including the structure, energy, speed and other attributes, so that the factor cannot exert a coupling effect.

Transition strategy. This involves changing the environment of factors, so that the factors cannot take effect.

Emergency derivation refers to the measures taken to deal with an emergency leading to other harmful events. It emphasizes the subjective of human coping and the negative effects of the influence.

According to the effect of response measures on derivative events, it can be divided into two types: excessive derivation and harmful derivation of measures. The excessive derivative of measures is that the excessive response of the original events lead to new hazard events. The domination strategy or neutralization strategy should be implemented for excessive derivation of measures.

Domination strategy. The domination strategy in derivation stage involves controlling the speed, degree or quantity of response measures to prevent excessive measures from being taken. The domination strategy in the coupling mechanism is to control the interaction of coupling factors, but the control strategy in the derivation mechanism is to prevent excessive response measures
.
Neutralization strategy. This refers to the use of another measure to neutralize or buffer the destructive power of excessive coping strategies.

The elimination strategy or pre-evaluation strategy should be adopted to restrict harmful derivatives of measures.

Elimination strategy. The elimination strategy refers to the adoption of other measures to eliminate the harmful effects of response measures. It can take the following forms: the strategy acts on the response measures directly to reduce their harmfulness; the strategy adopts a protective role on the disaster bearing body of derivative events; and the strategy adopted acts on the original emergencies to prevent them from evolving into new ones.

Pre-evaluation strategy. If eliminating the derivative effect of harmful strategies is impossible, only the pre-evaluation strategy can be adopted to gage the hazards caused by adopting or not adopting the counter measures, to then decide whether to implement the measures or not.

----End of Reference for Association Rule Question ------

Please download the data and conduct Association Rule Mining on it. You may use any tool you feel comfortable with. Please report:

1. **The support and confidence threshold you used**
2. **List the top 5 rules you discovered, i.e., the rules with the highest confidence that is interesting. An interesting rule is one that shows a connection between <u>STRATEGY and MECHANISM</u>. Report the confidence and lift of your rules.**
3. **Interpret your findings – what types of strategy should be recommended for what type of mechanisms?**

Case 2 (15 points):

Forbes recently announced billionaires of 2021. The data set is here: https://www.kaggle.com/alexanderbader/forbes-billionaires-of-2021-20?select=forbes_billionaires.csv

Please take some time to read and understand the data.

This is an open-ended question. Please

a. **Formulate a problem** – it can be supervised, i.e., predicting something, or unsupervised, i.e., analyzing or discovering some patterns in the data. Be clear about what you are looking for, phrase your **research question** clearly.

b. Given the problem statement, provide a solution. You need to:
    a. List the attributes you need for your task
    b. Name the data mining algorithm you will use
    c. If you are running association rules, state what type of association you are looking for, and state the support and confidence you are planning to use. Justify your parameters.
    d. If you plan to use clustering, explain what type of cluster you are trying to form, and the type of clustering algorithms you plan to use.
    e. If you are running prediction, define the label. Explain which algorithm you will be using, and any parameters you might consider tuning.

c. RUN the mining using any tool of your choice, show either your code (Python) or screenshot of your mining process (RapidMiner). Discuss your findings.