Name: ___________________________ **26 points possible**

CS 5402 – Intro to Data Mining
Fall 2021
HW #3

Submit as a single pdf file via Canvas by 11:59 p.m. on Oct. 8, 2021

1. Consider the following dataset:

married	education	income	creditLine	cardCategory
no	college	low	10k	Blue
yes	college	low	5k	Gold
no	college	low	10k	Blue
yes	highSchool	middle	7k	Silver
yes	graduate	middle	7k	Silver
no	highSchool	high	5k	Red
no	college	middle	10k	Gold

a. Compute the **coverage** of each item set listed below. **(1 pt.)**

Item Set	**Coverage**
education = highSchool, ***cardCategory*** = Red	___
married = no, ***income*** = low, ***creditLine*** = 7k	___

b. Write down every **association rule** that could be generated from the 2-item set listed below, regardless of whether or not there are actually any instances of that rule in our given dataset. Hint: You should be able to generate 3 rules. **(1.5 pts.)**

married = no, ***cardCategory*** = Blue

c. Compute the **accuracy** of each rule listed below. Express accuracy as a **fraction** (e.g., 2/3, 2/2, etc.), **NOT** as a decimal number (e.g., 0.67, 1.0, etc.). **(1.5 pts.)**

Rule	**Accuracy**
If *married* = *yes* then *income* = middle	___
If *married* = no and *education* = college then *creditLine* = *10k* and *cardCategory* = Blue	___
If _ then *cardCategory* = Red and *married* = yes	___

Name: ______________________________ 26 points possible

2. The dataset shown below is posted on Canvas (along with this assignment) as **creditBinary.csv**. Run the **Prism** algorithm on it in **Weka** specifying ***cardCategory*** as the decision attribute. List the classification rules that are produced (you can just include a screenshot of your Weka output). Then work out the Prism algorithm **by hand** starting with a rule for ***cardCategory = Blue*** to **show** what classification rules you would get; who knows, they might be different than what Weka produces! **SHOW ALL OF YOUR WORK!!! (6.5 pts.)**

   If there is a **tie between 2 attributes**, choose the attribute that comes first in the table as listed from left to right (e.g., ***education*** comes before ***creditCardDebt***). This will make it easier on the grader (i.e., multiple possible solutions won't have to be considered!).

married	**education**	**income**	**creditCardDebt**	**cardCategory**
yes	highSchool	ge50k	low	Blue
yes	highSchool	ge50k	high	Blue
no	highSchool	ge50k	low	Blue
no	college	lt50k	low	Gold
no	college	lt50k	high	Gold
yes	college	lt50k	low	Gold
yes	highSchool	lt50k	high	Gold
no	college	ge50k	high	Gold
no	highSchool	lt50k	low	Gold
yes	college	ge50k	high	Blue

Name: ____________________________ **26 points possible**

3. Consider the dataset shown below where the decision attribute is ***paidCash***. Assume that attribute weights $\mathbf{w_{milk}}$, $\mathbf{w_{beer}}$, $\mathbf{w_{diapers}}$, and $\mathbf{w_{chips}}$ (corresponding to attributes *boughtMilk, boughtBeer, boughtDiapers,* and *boughtChips*, respectively) are all initialized to 2. If **ϴ** is 2, **α** is 2, and **β** is 0.5, what will the **attribute weights** (i.e., $\mathbf{w_{milk}}$, $\mathbf{w_{beer}}$, $\mathbf{w_{diapers}}$, and $\mathbf{w_{chips}}$) be after **one** iteration of the **Winnow** algorithm? **YOU MUST SHOW YOUR WORK** in computing these values; otherwise, you will receive **NO CREDIT! (2 pts.)**

	boughtMilk	**boughtBeer**	**boughtDiapers**	**boughtChips**	**paidCash**
x1	0	1	0	1	0
x2	1	1	0	0	1
x3	0	0	0	1	1
x4	0	1	0	0	0

Final values: $\mathbf{w_{milk}}$ = ____ $\mathbf{w_{beer}}$ = ____ $\mathbf{w_{diapers}}$ = ____ $\mathbf{w_{chips}}$ = ____

Name: ____________________________ **26 points possible**

4. Consider the dataset given below where the decision attribute is the one labeled ***z***. Build a **kd-tree** where **k = 2**. **No partial credit will be given unless you SHOW YOUR WORK! (8.5 pts.)**

   When computing medians, if you have a real number, **round** .1 to .4 **down** to the next integer, and **round** .5 to .9 **up** to the next integer (e.g., round 2.5 to 3, round 2.3 to 2, etc.).

   When processing the non-decision attributes, process them in alphabetical order (i.e., x before y).

x	**y**	**z**
1	5	green
2	8	blue
2	10	red
3	20	blue
4	20	green
5	30	red
6	40	blue
7	50	green
8	60	red

Name: ____________________________ 26 points possible

5. Consider the dataset given below where the decision attribute is the one labeled ***class***. Show how **k-means clustering** using **k = 3** would cluster the instances on attributes ***a*** and ***b*** assuming that the initial cluster centers you start with are **(2, 4)**, **(5, 6)**, and **(8, 1)**. **SHOW ALL OF YOUR WORK!**

   Use **Manhattan distance** for your calculations. When computing centers, if you have a real number, **round** .1 to .4 **<u>down</u>** to the next integer, and **round** .5 to .9 **<u>up</u>** to the next integer (e.g., round 2.5 to 3, round 2.3 to 2, etc.).

   Do **<u>NOT</u>** draw a graph showing the final clusters; simply specify what the clusters will be in terms of **what each cluster's center is and what instances from the dataset will be in each cluster**. **(5 pts.)**

a	**b**	**c**	**class**
2	4	11	true
5	6	5	false
8	1	7	false
7	3	4	true
4	10	8	true
3	0	3	true
9	8	1	false