Exercise 2 (23 points) – individual work

- The answers can be typed or handwritten (handwriting must be clear and readable), in this exercise sheet or your own sheet (put your name & ID at the top of the sheet). All answers must be saved to only 1 PDF file.
- Some questions also require the submission of processes/workflows (file.rmp or file.ipynb).
- In case of re-submission (after first grading) or submission after solution is given, your points will be weighted by 0.5.

---

1. (Total 10 points)

1.1 (4 points) From the table, calculate GINI when splitting data by each attribute (answer in 4 decimal places). And based on GINI, which attribute should be chosen as a tree node first?

Record	Vehicle	Gender	Time	Class
1	Car	Male	pm	A
2	Car	Female	am	B
3	Bus	Male	pm	A
4	Bus	Female	pm	B
5	Bus	Male	pm	A
6	Car	Female	pm	B
7	Bus	Male	am	B
8	Bus	Male	am	A
9	Car	Male	pm	A
10	Bus	Male	am	B

Split by Vehicle	GINI =
Split by Gender	GINI =
Split by Time	GINI =
Which one to choose first?	

1.2 (4 points) Consider another dataset that has 12 records (6 in class yes, 6 in class no), and a nominal attribute "brand" with 3 categories {iPhone, Samsung, Oppo}. When splitting the data by brand, using both binary split and multi-way split, entropies after split are as shown in the table. Calculate gain ratio for each split option.

Note 1: Find entropy before split from the full data set

Note 2: Use bin sizes to calculate split penalty for each option

Option	Split Method	Bin size	Total entropy after split	Gain Ratio
1	Bin1: brand = iPhone, Samsung Bin2: brand = Oppo	4 yes, 4 no 2 yes, 2 no	1	
2	Bin1: brand = iPhone, Oppo Bin2: brand = Samsung	5 yes, 4 no 1 yes, 2 no	0.9729	
3	Bin1: brand = iPhone Bin2: brand = Samsung, Oppo	3 yes, 2 no 3 yes, 4 no	0.9793	
4	Bin1: brand = iPhone Bin2: brand = Samsung Bin3: brand = Oppo	3 yes, 2 no 1 yes, 2 no 2 yes, 2 no	0.9675	

1.3 (2 points) From question 1.2, which split option should be chosen if we use gain ratio as the criterion? And which option should be chosen if we use information gain as the criterion instead?

2. (Total 5 points) Consider the following rule set generated from training data in question 1.1

Rule 1	if gender = female	then class = B
Rule 2	if gender = male && vehicle = bus	then class = A
Rule 3	if vehicle = car && time = pm	then class = A
Rule 4	if vehicle = bus && time = am	then class = B

There are 8 combinations of attribute values e.g. (car, male, am), (car, male, pm), (car, female, am), (car, female, pm), … To check exhaustive and mutual exclusive properties, you need to check whether every combination can be predicted by only 1 rule.

2.1 (1 point) Find combination (there is 1) that indicate the lack of exhaustive property.

2.2 (2 points) Find combinations (there are 2) that indicate the lack of mutual exclusive property, and rules that are triggered by each combination.

2.3 (2 point) If we solve the conflicts in question 2.2 by following rules with the highest accuracy, what would be the prediction for each combination in 2.2.

Name _________________________ ID ____________

3. (Total 8 points) Retrieve **Past Campaign Data**.

3.1 (2 points) Build a work flow for classification task.

- Step 1. Split data into 2 partitions: 70% of the data for training and 30% for testing.
- Step 2. Use either decision tree or rule induction to classify response. Adjust parameters to get high accuracy (>= 85% on test data). If the accuracy is already high, you may adjust parameters to get better or clearer patterns (e.g. bigger tree nodes with higher coverage or smaller nodes with more conditions).
- Step 3. Submit your workflow. Make sure that it outputs at least performance table and decision tree or rule set. Name the workflow **question3.rmp.**

3.2 (3 points) Capture (a portion of) tree or rule set and highlight node/rule that predict yes. The node/rule that you choose should have high coverage, high accuracy, and contain interesting conditions. In case of rule set, don't pick default rule without any precondition.

Questions/Instructions	Answers
(a) Show the captured tree/rule & describe pattern that predicts yes in words (read all conditions in the rule, or all conditions that lead to the leaf node).	
(b) Report coverage of this pattern, i.e. number or percentage of records it covers.	
(c) Report accuracy of this pattern	

3.3 (3 points) Capture (a portion of) tree or rule set and highlight node/rule that predict no. Follow the same guidelines as in question 3.2.

Questions/Instructions	Answers
(a) Show the captured tree/rule & describe pattern that predicts no in words.	
(b) Report coverage of this pattern.	
(c) Report accuracy of this pattern	