Exercise I (Classification) (20 pts)

We will use the dataset below to learn a decision tree which predicts if people pass machine learning (Yes or No), based on their previous GPA (High, Medium, or Low) and whether or not they studied.

GPA	Studied	Passed
L	F	No
L	T	Yes
M	F	No
M	T	Yes
H	F	Yes
H	T	Yes

1. [4 points] What is the total entropy?
2. [4 points] What is the entropy of *GPA*?
3. [4 points] What is the entropy of *Studied*?
4. [8 points] Draw the full decision tree that would be learned for this dataset. Show all necessary calculations.

Exercice II (Confusion Matrix) (10 pts)

In a medical application domain, suppose we build a classifier for patient screening (True means patient has cancer). Suppose that the confusion matrix is from testing the classifier on some test data.

		Predicted	
		True	False
Actual	True	*TP*	*FN*
	False	*FP*	*TN*

Which of the following situations would you like your classifier to have? Explain

a. $FP >> FN$
b. $FN >> FP$
c. $FN = FP \times TP$
d. $TN >> FP$
e. $FN \times TP >> FP \times TN$
f. All of the above

Exercise III (Association Rules) (25 pts)

LEBANESE UNIVERSITY

Faculty of Sciences – Section I

الجامعة اللبنانيّة

كلية العلوم - الفرع الأول

2nd Session Exam 2023: INFO 4447 -E -- Duration: 1h:30m

Consider the following dataset as shown below.

Customer ID	Transaction ID	Items bought
1	1001	{i1, i4, i5}
1	1024	{i1, i2, i3, i5}
2	1012	{i1, i2, i4, i5}
2	1031	{i1, i3, i4, i5}
3	1015	{i2, i3, i5}
3	1022	{i2, i4, i5}
4	1029	{i3, i4}
4	1040	{i1, i2, i3}
5	1033	{i1, i4, i5}
5	1038	{i1, i2, i5}

1. Compute the support for itemsets {i5}, {i2, i4}, and {i2, i4, i5} by treating each transaction ID as a market basket.
2. Use the results in part (a) to compute the confidence for the association rules {i2, i4}→ {i5} and {i5}→ {i2, i4}.
3. Is confidence a symmetric measure? Justify your answer.

Exercise IV (K means Clustering)

(25 pts)

1. Explain the definition of a centroid in k-means.
2. The following is a set of one-dimensional points: {1,1,2,3,5,8,13,21,33,54}. Perform all iterations of k-means on these points using the two initial centroids 0 and 11.

Exercise V (Agglomerative Clustering)

(20 pts)

Use single and complete link agglomerative clustering to group the data described by the following distance matrix. Show the dendrograms.

	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0