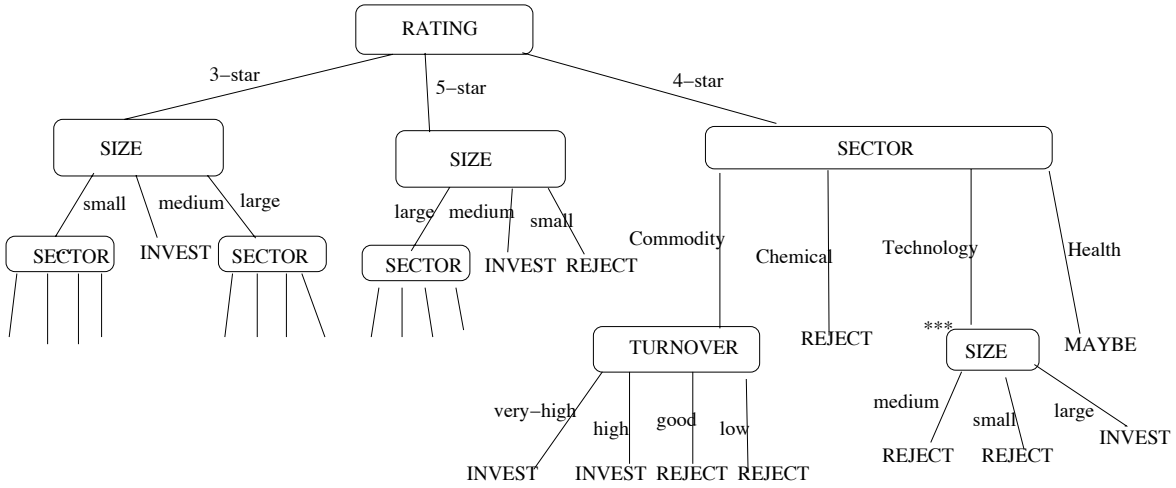
CISC-483-683: Assignment 5

Due Friday, October 6, 2023

Will not be accepted after 11:30am Wednesday October 11

65 points

1. Suppose that you have the following decision tree, where **RECOMMEND?** is the class whose value we are trying to predict.



Suppose also that the training data used to construct this decision tree had the following instances — **NOTE: The leftmost column of the table gives the number of training instances with the specified attribute values.**

# OF INSTANCES	SIZE	SECTOR	TURNOVER	RATING	DECISION?
2	large	Commodity	low	4-star	INVEST
15	large	Commodity	low	4-star	REJECT
45	large	Commodity	good	4-star	REJECT
35	medium	Commodity	good	4-star	INVEST
80	small	Commodity	high	4-star	INVEST
20	medium	Commodity	high	4-star	REJECT
90	large	Commodity	very-high	4-star	INVEST
10	medium	Commodity	very-high	4-star	REJECT
10	medium	Health	small	4-star	INVEST
20	small	Health	medium	4-star	MAYBE
3	large	Health	large	4-star	REJECT
2	small	Chemical	small	4-star	INVEST
1	medium	Chemical	medium	4-star	MAYBE
3	large	Chemical	large	4-star	REJECT
50	large	Technology	good	4-star	INVEST
10	large	Technology	good	4-star	REJECT
40	medium	Technology	high	4-star	INVEST
40	medium	Technology	good	4-star	REJECT
1	small	Technology	high	4-star	INVEST
10	small	Technology	good	4-star	REJECT
etc.					

(a) (20 points) Suppose that you are given the following pruning dataset, which includes all of the pruning data where the RATING attribute has the value *4-star*:

SIZE	SECTOR	TURNOVER	RATING	DECISION?
large	Commodity	very-high	5-star	INVEST
medium	Commodity	very-high	5-star	INVEST
medium	Chemical	good	5-star	INVEST
small	Technology	very-high	5-star	REJECT
large	Commodity	good	4-star	INVEST
medium	Commodity	good	4-star	REJECT
large	Commodity	very-high	4-star	INVEST
medium	Commodity	very-high	4-star	INVEST
large	Commodity	very-high	4-star	INVEST
small	Commodity	very-high	4-star	REJECT
large	Commodity	high	4-star	INVEST
medium	Commodity	high	4-star	INVEST
small	Commodity	high	4-star	REJECT
large	Technology	low	4-star	REJECT
large	Technology	high	4-star	INVEST
large	Technology	very-high	4-star	INVEST
medium	Technology	high	4-star	INVEST
medium	Technology	very-high	4-star	REJECT
small	Technology	low	4-star	REJECT
medium	Technology	high	4-star	INVEST
medium	Technology	good	4-star	INVEST
large	Chemical	very-high	4-star	INVEST
large	Health	very-high	4-star	INVEST
small	Health	very-high	4-star	MAYBE
medium	Health	very-high	4-star	MAYBE
small	Health	low	4-star	MAYBE
medium	Commodity	good	3-star	REJECT
large	Technology	very-high	3-star	INVEST
small	Technology	very-high	3-star	INVEST
etc.				

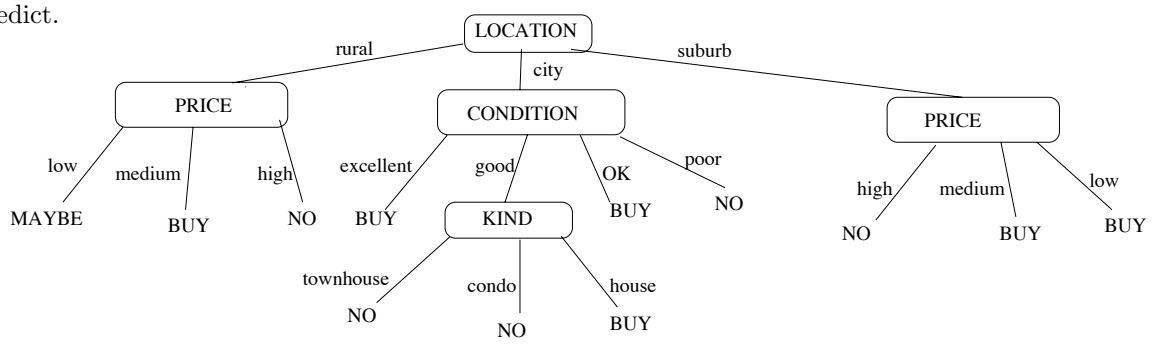
Consider the rightmost subtree of RATING (the one on the branch *4-star*). Use reduced-error pruning to determine which nodes should be pruned. Please show the details of your work and draw the final decision tree after pruning.

(b) (20 points) Suppose that you don't have a pruning dataset and must instead rely on the training set for pruning. Instead of the training set given earlier, suppose that the following training instances end up at the leaf nodes of the rightmost SIZE subtree indicated by *** in the decision tree on the preceding page.

# OF INSTANCES	SIZE	SECTOR	TURNOVER	RATING	DECISION?
17	large	Technology		4-star	INVEST
14	large	Technology		4-star	REJECT
15	medium	Technology		4-star	INVEST
22	medium	Technology		4-star	REJECT
4	small	Technology		4-star	INVEST
6	small	Technology		4-star	REJECT

Use pessimistic error pruning to determine whether to prune the node indicated by *** — use a confidence value of 58%. Please show the details of your work. If you decide not to prune, explain why. If you decide to prune, please explain why and draw the resulting decision tree.

2. (25 points) Suppose that you have the following decision tree, where **DECISION?** is the class whose value we are trying to predict.



Find-Home is a real estate comapny that helps individuals find homes. Find-Home has constructed a decision tree for recommending property. Suppose that this decision tree was constructed from the data shown below (and other instances not included in the table but which do not affect your answers to this problem.) Note that there are instances that are missing an attribute value. The class attribute is DECISION.

PRICE	LOCATION	KIND	CONDITION	DECISION
high	city	townhouse	excellent	BUY
medium	city	townhouse	OK	BUY
medium	city	townhouse	OK	MAYBE
medium	city	house	OK	MAYBE
medium	city	condo	OK	BUY
high	city	condo	OK	NO
medium	city	condo	poor	NO
high	city	townhouse	poor	NO
medium	city	townhouse	poor	NO
medium	city	condo	poor	NO
medium	city	condo	poor	NO
high	city	condo	poor	MAYBE
low	city	townhouse	good	MAYBE
medium	city	condo	good	NO
high	city	townhouse	good	NO
medium	city	house	good	MAYBE
high	city	house	good	BUY
medium	city	house	good	BUY
medium	city	townhouse	good	NO
medium	city	townhouse		NO
low	city	condo		BUY
high	city	condo		NO

In the following problems, use the method described in class for utilizing information in instances that are missing values for some predictor attributes.

(a) What is the entropy of the instances that arrive at the node that is labelled KIND on the path LOCATION=city, CONDITION=good. Please show your work.

(b) The class value on the path LOCATION=city, CONDITION=OK is BUY. Taking into account the training instances with a missing attribute value, show why BUY is the class label on this path in the decision tree.

(c) **CISC-483**: any two test cases (extra credit for doing all three test cases)
CISC-683: all three test cases Assume the following:
A. 22 training instances have LOCATION=city (they are shown in the above table)
B. 13 training instances have LOCATION=suburb
C. 15 training instances have LOCATION=rural

Suppose that you have the following three test instances:

Test-instance	PRICE	LOCATION	KIND	CONDITION
1.	medium	rural	condo	
2.	high		townhouse	OK
3.	low	city	townhouse	

For each test case, determine what class value the decision tree would assign and why — please make sure to show your work and computations.