Institute of Information Technology
Jahangirnagar University
Professional Masters in IT

1st Semester Special Final Examination

Semester: Spring 2019

Duration: 3 Hours

Full Marks: 60

Course Code: PMIT – 6307

Course Title: Data mining & Knowledge Discovery

Do not write anything on the question paper.

There are 7 (Seven) questions. Answer any 5 (Five) of them.

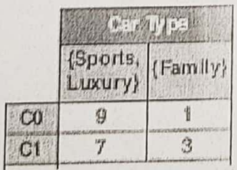
Figures in the right margin indicate marks.

1. a) Define Data Mining. List the steps of the Knowledge Discovery in Databases (KDD). 1+2
   b) What is meant by outlier? 2
   c) Discuss (shortly) whether or not each of the following activities is a data mining task. 3
      i) Computing the total sales of a company.
      ii) Predicting the outcomes of tossing a (fair) pair of dice.
      iii) Monitoring the heart rate of a patient for abnormalities.
   d) What is simple random sampling? Is simple random sampling (without replacement) a good approach to sampling? Why or why not? 2+2

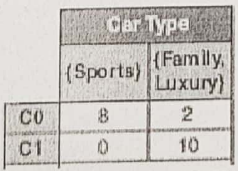
2. a) Mention the different types of attributes with necessary example. 3
   b) What is meant by aggregation? What are the purposes of aggregation? Give 2 examples in which aggregation is useful. 2+2+1
   c) What is meant by regression? 2
   d) What summary statistics can be used on nominal attributes? 2

3. a) Define training set and testing set. 2
   b) Discuss disadvantages of testing a model over its training set. 2
   c) Explain the procedure of multi-way and binary splitting for ordinal and continuous attribute. 3
   d) Explain the procedure to compute Entropy of discrete attribute. 3
   e) What are the stopping criteria for tree induction? 2

4. a) Discuss issues that are important to consider when employing a Decision Tree based classification algorithm. 2
   b) Mention three parameters to measures the node impurity. Compare to each other. 3
   c) Consider the decision tree shown in the following figure. Calculate the GINI (children) or weighted average GINI index for each grouping and compare the results. 5



	Car Type	
	{Sports, Luxury}	{Family}
C0	9	1
C1	7	3

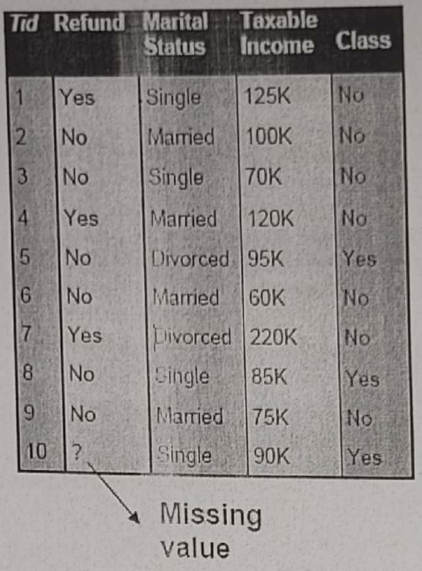


	Car Type	
	{Sports}	{Family, Luxury}
C0	8	2
C1	0	10

   d) What are the main advantages and disadvantages of Decision Tree classification algorithms? 2

5. a) What is visualization? Explain the procedure of histogram technique. 2+2
   b) Explain the procedure of multi-way and binary splitting for nominal attribute. 3

c) Consider the following figure. Calculate the entropy gain while splitting on refund attribute. 5

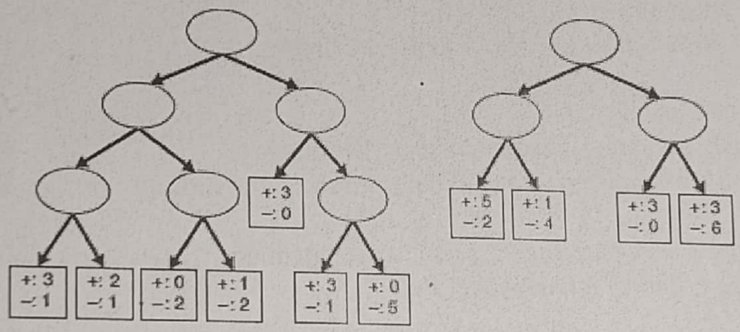


Tid	Refund	Marital Status	Taxable Income	Class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	?	Single	90K	Yes

Missing value

6. a) Explain the procedure to calculate classification error. 2

b) Consider the binary decision trees (shown in the figure) with training error 4 and penalty term is equal to 0.5. Calculate the generalization error. 5



Decision Tree, T_L Decision Tree, T_R

c) Write down the algorithm to compute the GINI index for the case of continuous attribute. 3

d) Consider the following figure. 2

Node N_2	Count
Class=0	1
Class=1	5

Calculate the classification error.

7. a) What is meant by a proximity matrix? 2

b) Describe the principles and ideas regarding Agglomorative Hierarchical Clustering. 2

c) Explain the working of the K-means algorithm and its advantages and disadvantages. 3

d) Explain the problem of selecting good initial points for the K-means algorithm. 3

e) How do you evaluate K-means clustering? 2