



Roll No.

**PRESIDENCY UNIVERSITY  
BENGALURU**

**SCHOOL OF ENGINEERING**

**TEST 1**

**Sem & AY:** Odd Sem. 2019-20

**Date:** 27.09.2019

**Course Code:** CSE 307

**Time:** 11AM to 12PM

**Course Name:** DATA MINING

**Max Marks:** 40

**Program & Sem:** B.Tech (CSE) & V DE

**Weightage:** 20%

**Instructions:**

- (i) all questions carry equal marks
- (ii) answer all questions sequentially

**Part A [Memory Recall Questions]**

**Answer all the Questions. Each Question carries two marks. (4x2M=8M)**

1. What is Data Mining? (C.O.NO.1) [Knowledge]
2. Briefly explain on category of Data Mining Tasks. (C.O.NO.1) [Knowledge]
3. Infer the condition of entropy and entropy gain for an effective split up.  
(C.O.NO.2) [Comprehension]
4. What is Stratified Sampling? (C.O.NO.2) [Knowledge]

**Part B [Thought Provoking Questions]**

**Answer both the Questions. Each Question carries two marks. (2x6M=12M)**

5. Describe in detail the challenges on Data Mining (C.O.NO.1) [Knowledge]
6. For the following vectors, X and Y, Calculate the Similarity and Dissimilarity measures.
  - i)  $x = (1, 1, 1, 1)$ ,  $y = (2, 2, 2, 2)$  cosine, correlation, Euclidean
  - ii)  $x = (0, 1, 0, 1)$ ,  $y = (1, 0, 1, 0)$  cosine, correlation, Euclidean, Jaccard

**Part C [Problem Solving Questions]**

**Answer both the Questions. Each Question carries ten marks. (2x10M=20M)**

7. Explain in detail with example and relevant diagram different types of Data sets.

(C.O.NO.1) [Knowledge]

8. From the given data shown in Table 1 set by using entropy based discretization, Identify the best among the three boundary values of hours 4.5, 6.5, 10 with justification.

(C.O.NO.2) [Application]

User ID	1	2	3	4	5
Hours Studied	4	5	8	12	15
Test Attended	No	Yes	No	Yes	Yes

Table. 1



## SCHOOL OF ENGINEERING

Semester: V

Course Code: CSE307

Course Name: Data Mining

Date: 27-09-2019

Time: 11AM to 12PM

Max Marks: 40

Weightage: 20%

### Extract of question distribution [outcome wise & level wise]

Q.NO	C.O.NO	Unit/Module Number/Unit /Module Title	Memory recall type [8 Marks] Bloom's Levels			Thought provoking type [12 Marks] Bloom's Levels			Problem Solving type [ 20 Marks ]			Total Marks
			K & C			K&C			K&A			
1	CO1	UNIT 1	K	2	-	-	-	-	-	-	-	2
2	CO1	UNIT 1	K	2	-	-	-	-	-	-	-	2
3	CO2	UNIT 2	C	2	-	-	-	-	-	-	-	2
4	CO2	UNIT 2	K	2	-	-	-	-	-	-	-	2
5	CO1	UNIT 1	-	-	-	K	6	-	-	-	-	6
6	CO2	UNIT 2	-	-	-	C	6	-	-	-	-	6
7	CO1	UNIT 1	-	-	-	-	-	-	K	10	-	10
8	CO2	UNIT 2	-	-	-	-	-	-	A	10	-	10
	Total Marks		-	8	-	-	12	-	-	20	-	40

K =Knowledge Level C = Comprehension Level, A = Application Level



Note: While setting all types of questions the general guideline is that about 60%

Of the questions must be such that even a below average students must be able to attempt, About 20% of the questions must be such that only above average students must be able to attempt and finally 20% of the questions must be such that only the bright students must be able to attempt.

[I hereby certify that All the questions are set as per the above guide lines. Dr. Thivakaran  
T K]

Reviewers' Comments



## Annexure- II: Format of Answer Scheme



### SCHOOL OF ENGINEERING

#### SOLUTION

Semester: V

Course Code: CSE307

Course Name: Data Mining

Date: 27-09-19

Time: 11.00AM to 12.00PM

Max Marks: 40

Weightage: 20%

#### Part A

(4 x 2 = 8 Marks)

Q No	Solution	Scheme of Marking	Max. Time required for each Question
1.	i) Discovering Useful Information ii) Discovering Patterns iii) Predict Information.	02 Marks	05 minutes
2.	i) Predictive Tasks ii) Descriptive Tasks	02 Marks	05 minutes
3.	i) Entropy value to be Minimum ii) Entropy gain to be Maximum	02 Marks	05 minutes
4.	i) pre specified groups of objects. ii) Equal numbers of objects are drawn from each group even though the groups are of different sizes. another iii) The number of objects drawn from each group is proportional to the size of that group.	02 Marks	05 minutes

#### Part B

(2 x 6 = 12 Marks)





	Solution	Scheme of Marking	Max. Time required for each Question
5.	i) Scalability ii) Dimensionality iii) Heterogeneous / Compound data iv) Data Ownership and distribution	Any three out of four points with brief descriptions Each carries 2 marks Total 6 Marks	10 minutes
6.	i) $\cos(x, y) = 1$ , $\text{corr}(x, y) = 0/0$ (undefined), $\text{Euclidean}(x, y) = 2$  ii) $\cos(x, y) = 0$ , $\text{corr}(x, y) = -1$ , $\text{Euclidean}(x, y) = 2$ , $\text{Jaccard}(x, y) = 0$	Each carries Three Marks Total 6 Marks	10 minutes

**Part C**

(2 x 10 = 20 Marks)

Q No	Solution	Scheme of Marking	Max. Time required for each Question
7.	i) Record data ii) Transaction data iii) Data Matrix iv) Graph based data v) Sequential data	Each carries of 2 marks Total 10 Marks	15 minutes
8.	i) For 4.5 Net Entropy = 0.648 Gain = 0.322 ii) For 6.5 Net Entropy = 0.944 Gain = 0.027. iii) For 10 Net Entropy = 0.55 Gain = 0.421	Each carries of 3 Marks Formula carries of 1 marks Total 10 Marks	10 minutes





Roll No.

**PRESIDENCY UNIVERSITY  
BENGALURU**

**SCHOOL OF ENGINEERING**

**TEST – 2**

**Sem & AY:** Odd Sem. 2019-20

**Date:** 16.11.2019

**Course Code:** CSE 307

**Time:** 11:00 AM to 12:00 PM

**Course Name:** DATA MINING

**Max Marks:** 40

**Program & Sem:** B.Tech (CSE) & V

**Weightage:** 20%

---

**Instructions:**

- I. Answer all questions sequentially
- 

**Part A [Memory Recall Questions]**

**Answer all the Questions. Each Question carries two marks. (4Qx2M=8M)**

1. Differentiate model overfitting and underfitting with respect to decision tree based algorithms. (C.O.NO.4) [Comprehension]
2. Name the two most significant characteristics of rule based classifiers. (C.O.NO.4) [Knowledge]
3. Find the Gini Index of a set, which has 2 examples of class C1 and 4 examples of class C2. (C.O.NO.4) [Comprehension]
4. Define the two rule evaluation metrics namely, support and confidence of an association rule  $X \rightarrow Y$ . (C.O.NO.3) [Knowledge]

**Part B [Thought Provoking Questions]**

**Answer both the Questions. Each Question carries six marks. (2Qx6M=12M)**

5. Given frequent 3-itemset  $L_3 = \{\{1,2,3\}, \{1,2,4\}, \{1,3,4\}, \{1,3,5\}, \{2,3,4\}\}$ , generate the candidate 4-itemsets  $C_4$  using Apriori algorithm. Justify the answer. (C.O.NO.3) [Application]
6. Draw the FP Tree for the following transaction database using minimum support count as 2. (C.O.NO.3) [Application]

TID	Items
1	A, C, D
2	B, C, E
3	A, B, C, E
4	B, E

**Part C [Problem Solving Questions]**

**Answer both the Questions. Each Question carries ten marks. (2Qx10M=20M)**

7. Following is the training data set for a decision tree classifier to predict the factors affecting sunburn. (C.O.NO.4) [Application]
- What is the entropy of this data set?
  - Using multi-way split on the attributes and entropy as the impurity measure, find the root node of the tree.
  - Draw the decision tree after first iteration.

Name	Hair	Height	Lotion	Sunburned
Sarah	Blonde	Average	No	Yes
Dana	Blonde	Tall	Yes	No
Alex	Brown	Short	Yes	No
Annie	Blonde	Short	No	Yes
Emily	Red	Average	No	Yes
Pete	Brown	Tall	No	No
John	Brown	Average	No	No
Katie	Blonde	Short	Yes	No

8. a) Using a neat diagram define confusion matrix with its four terms. (C.O.NO.4) [Knowledge]
- b) For the following confusion matrix of a certain binary classifier, calculate the classification accuracy and precision. (C.O.NO.4) [Application]

		PREDICTED CLASS	
		Class = Yes	Class = No
ACTUAL CLASS	Class = Yes	100	5
	Class = No	10	50



## SCHOOL OF ENGINEERING

Semester: V

Course Code: CSE 307

Course Name: Data Mining

Date: 16.11.2019

Time: 11.00am – 12.00noon

Max Marks: 40

Weightage: 20%

### Extract of question distribution [outcome wise & level wise]

Q.NO	C.O. NO	Unit/Module Number/Unit /Module Title	Memory recall type [Marks allotted] Bloom's Levels			Thought provoking type [Marks allotted] Bloom's Levels			Problem Solving type [Marks allotted]			Total Marks
			K			C			A			
1	4	4 – Classification					2					2
2	4	4. Classification	2									2
3	4	4. Classification					2					2
4	3	3. Frequent Patterns	2									2
5	3	3 Frequent Patterns							6			6
6	3	3 Frequent Patterns							6			6
7	4	4 Classification							10			10
8	4	4 Classification							10			10
	Tot al		4				4		32			40



	Marks														
--	-------	--	--	--	--	--	--	--	--	--	--	--	--	--	--

K = Knowledge Level C = Comprehension Level, A = Application Level

Note: While setting all types of questions the general guideline is that about 60%

Of the questions must be such that even a below average students must be able to attempt, About 20% of the questions must be such that only above average students must be able to attempt and finally 20% of the questions must be such that only the bright students must be able to attempt.

## Annexure- II: Format of Answer Scheme



### SCHOOL OF ENGINEERING

#### SOLUTION

Semester: V

Course Code: CSE 307

Course Name: DATA MINING

Date: 16.11.2019

Time: 11.00am – 12.00 noon

Max Marks: 40

Weightage: 20%

#### Part A

(4Q x 2M = 8Marks)

Q No	Solution	Scheme of Marking	Max. Time required for each Question
1.	<p><b>Underfitting</b> - when model is too simple, both training and test errors are large. The model has yet to learn the true structure of the data. Hence it performs poorly on training and test data.</p> <p><b>Overfitting</b> : As the size of the tree increases, the tree has fewer training and test errors. When the tree becomes too large, the test error increases although training error decreases.</p>	Definition 1 + 1 = 2 marks	5 mins
2.	Rules generated have to be mutually exclusive and exhaustive.	1 + 1 = 2 marks	5 mins
3.	$P(C1) = 2/6$ $P(C2) = 4/6$ $Gini = 1 - (2/6)^2 - (4/6)^2 = 0.444$	Partial – 1 Complete - 2	5 mins
4.	Support – fraction of transactions that have both X and Y.	1 + 1 = 2	5 mins

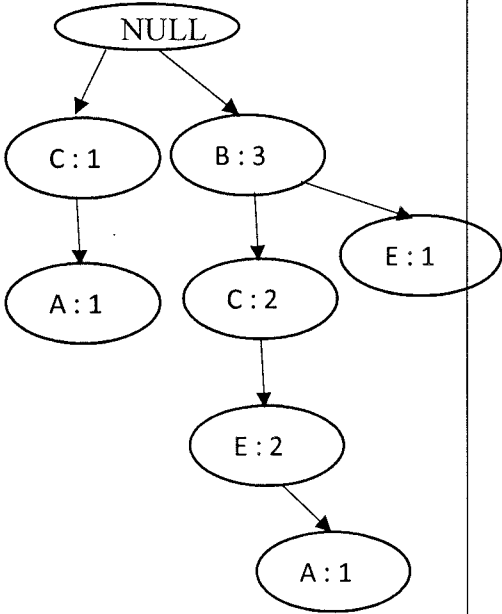




	Confidence – measures how often items in Y appear in transactions having X		
--	--	--	--

**Part B**

(2Q x 6M = 12Marks)

Q No	Solution	Scheme of Marking	Max. Time required for each Question																				
5.	$C_4 = \{\{1,2,3,4\}\}$ . The $\{1,3,4,5\}$ cannot be a frequent itemset, since one of its subsets $\{1,4,5\}$ is not in $F_3$	Results of Join step ---- 2 marks Results of Prune step ----- 2 marks Justification----- 2 marks	5 mins																				
6.	<p><math>L_1</math> Re-arranged TDB</p> <table border="1" data-bbox="272 564 493 792"> <thead> <tr> <th>Item</th> <th>Support count</th> </tr> </thead> <tbody> <tr> <td>A</td> <td>2</td> </tr> <tr> <td>B</td> <td>3</td> </tr> <tr> <td>C</td> <td>3</td> </tr> <tr> <td>E</td> <td>3</td> </tr> </tbody> </table> <table border="1" data-bbox="525 795 802 1055"> <thead> <tr> <th>TID</th> <th>Items</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>C A</td> </tr> <tr> <td>2</td> <td>B C E</td> </tr> <tr> <td>3</td> <td>B C E A</td> </tr> <tr> <td>4</td> <td>B E</td> </tr> </tbody> </table> <p>FP Tree</p>  <pre> graph TD     NULL(NULL) --&gt; C1(C:1)     NULL --&gt; B3(B:3)     C1 --&gt; A1(A:1)     B3 --&gt; C2(C:2)     B3 --&gt; E1(E:1)     C2 --&gt; E2(E:2)     E2 --&gt; A2(A:1)     </pre>	Item	Support count	A	2	B	3	C	3	E	3	TID	Items	1	C A	2	B C E	3	B C E A	4	B E	2 + 3 + 3 = 6 marks	10 mins
Item	Support count																						
A	2																						
B	3																						
C	3																						
E	3																						
TID	Items																						
1	C A																						
2	B C E																						
3	B C E A																						
4	B E																						



Q No	Solution	Scheme of Marking	Max. Time required for each Question													
7.	<p>a) <math>Ent(S) = 0.95443</math></p> <p><b>b) For attribute 'Hair':</b>            Values(Hair) : [Blonde, Brown, Red]  <math>S = [3+, 5-]</math>  <math>S_{Blonde} = [2+, 2-]</math>      <math>E(S_{Blonde}) = 1</math>  <math>S_{Brown} = [0+, 3-]</math>      <math>E(S_{Brown}) = 0</math>  <math>S_{Red} = [1+, 0-]</math>      <math>E(S_{Red}) = 0</math>  <math>Gain(S, Hair) = 0.95443 - [(4/8)*1 + (3/8)*0 + (1/8)*0]</math>  <math>= 0.45443</math></p> <p><b>For attribute 'Height':</b>            Values(Height) : [Average, Tall, Short]  <math>S_{Average} = [2+, 1-]</math>      <math>E(S_{Average}) = 0.91829</math>  <math>S_{Tall} = [0+, 2-]</math>      <math>E(S_{Tall}) = 0</math>  <math>S_{Short} = [1+, 2-]</math>      <math>E(S_{Short}) = 0.91829</math>  <math>Gain(S, Height) = 0.95443 - [(3/8)*0.91829 + (2/8)*0 + (3/8)*0.91829]</math>  <math>= 0.26571</math></p> <p><b>For attribute 'Lotion':</b>            Values(Lotion) : [Yes, No]  <math>S_{Yes} = [0+, 3-]</math>      <math>E(S_{Yes}) = 0</math>  <math>S_{No} = [3+, 2-]</math>      <math>E(S_{No}) = 0.97095</math>  <math>Gain(S, Lotion) = 0.95443 - [(3/8)*0 + (5/8)*0.97095]</math>  <math>= 0.01571</math></p> <p><b>Hence root node is 'Hair'.</b></p> <p><b>c) Root node = Hair</b>  <b>Hair = red then leaf node is "YES"</b>  <b>Hair = brown then leaf node is "NO"</b>  <b>Hair = blonde then decision is not clear</b></p>	<p>a) <math>Ent(S) = 2</math> M            b) Gain of each attribute <math>3 \times 2 = 6</math> M            Root node = 2 marks            c) tree after first iteration = 2 marks</p>	10 mins													
8.	<table border="1" data-bbox="312 1279 903 1503"> <thead> <tr> <th colspan="2" rowspan="2"></th> <th colspan="2">PREDICTED CLASS</th> </tr> <tr> <th>Class = Yes</th> <th>Class = No</th> </tr> </thead> <tbody> <tr> <th rowspan="2">ACTUAL CLASS</th> <th>Class = Yes</th> <td>TP</td> <td>FN</td> </tr> <tr> <th>Class = No</th> <td>FP</td> <td>TN</td> </tr> </tbody> </table> <p>a)            b) Accuracy = <math>150/165 = 0.91</math>            Precision = <math>TP/Total\ Yes = 100/110 = 0.91</math></p>			PREDICTED CLASS		Class = Yes	Class = No	ACTUAL CLASS	Class = Yes	TP	FN	Class = No	FP	TN	2 + 3 + 5 = 10M	10 mins
				PREDICTED CLASS												
		Class = Yes	Class = No													
ACTUAL CLASS	Class = Yes	TP	FN													
	Class = No	FP	TN													





Roll No																			
---------	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

**PRESIDENCY UNIVERSITY  
BENGALURU**  
**SCHOOL OF ENGINEERING**

**END TERM FINAL EXAMINATION**

**Semester:** Odd Sem. 2019 - 20

**Course Code:** CSE 307

**Course Name:** DATA MINING

**Program & Sem:** B.Tech (CSE) & V (DE-I)

**Date:** 20 December 2019

**Time:** 9:30 AM to 12:30 PM

**Max Marks:** 80

**Weightage:** 40%

**Instructions:**

- (i) Read the all questions carefully and answer accordingly.
- (ii) Write the answers legibly.

**Part A [Memory Recall Questions]**

**Answer all the Questions. Each Question carries 2 marks. (10Qx2M=20M)**

1. Identify the type of attributes in each of the following cases as nominal, ordinal, interval or ratio.  
a) Weight                      b) Gender                      c) Position in class                      d) Age in Years  
(C.O.No.1) [Knowledge]
2. Define Data warehousing. (C.O.No.1) [Knowledge]
3. A certain attribute called "Marks" in the data set has values 4, 8, 15, 25. It is required to transform these values using Z-score normalization where the mean is 13 and standard deviation is 7.96.what are the transformed values of "Marks". (C.O.No.2) [Knowledge]
4. Differentiate between Mutually exclusive rules and Exhaustive rules. (C.O.No.2) [Knowledge]
5. Define anti-monotone property (apriori property). (C.O.No.3) [Knowledge]
6. What is training set and test set? (C.O.No.3) [Knowledge]
7. Mention any two requirements of clustering in data mining. (C.O.No.2) [Knowledge]
8. Calculate Minkowski distance for the following points p(3,2) and q(4,6).  
(C.O.No.2) [Knowledge]
9. Define Web mining. (C.O.No.3) [Knowledge]
10. Define Text Mining. (C.O.No.3) [Knowledge]

**Part B [Thought Provoking Questions]**

**Answer all the Questions. Each Question carries 10 marks. (3Qx10M=30M)**

11. Using Naive bayes classifier from the given training set predict the class of the test set:  
Test Set =Sample x = {Green,Nonveg,Indian} (C.O.No.2) Comprehension]

Order	Box Color	Type	Origin	Pays or Not (class)
1	Green	Veg	Indian	Yes
2	Green	Veg	Indian	No
3	Green	Veg	Indian	Yes
4	Red	Veg	Indian	No

5	Red	Veg	Mexican	Yes
6	Red	Non-veg	Mexican	No
7	Red	Non-veg	Mexican	Yes
8	Red	Non-veg	Indian	No
9	Green	Non-veg	Mexican	No
10	Green	Veg	Mexican	Yes

12. a) Explain the KDD process with appropriate diagram. (C.O.No.1) [Knowledge]

b) Using Min Max normalization find the normalized value for the attribute height (cm) in the range (1, 2), given the value of attributes are: 65cm, 40cm, 85cm, 72cm, 50cm.

(C.O.No.2) [Comprehension]

13. Perform K mean clustering for the given data points. Consider K1 (92, 36) and K2 (85, 28) in the beginning. (C.O.No.2) [Comprehension]

Data Point	Length	Breadth
1	92	36
2	85	28
3	84	30
4	89	34
5	91	36
6	94	38

### Part C [Problem Solving Questions]

Answer both the Questions. Each Question carries 15 marks.

(2Qx15M=30M)

14. For the given proximity (Distance) matrix, perform agglomerative single link hierarchical clustering and represent the cluster using Dendrogram. (C.O.No.3)[Application]

	M1	M2	M3	M4	M5
M1	0				
M2	18	0			
M3	6	14	0		
M4	12	10	18	0	
M5	22	20	4	16	0

15. Form the following customer details create a decision tree using information gain.

Customer No.	Age Group	Income	Credit Range	Class Buy computer
1	Young	High	Weak	Yes
2	Senior	Low	Weak	Yes
3	Young	High	Weak	Yes
4	Young	High	Weak	Yes
5	Young	Low	Weak	No
6	Young	Low	Strong	No

(C.O.No.3)[Application]



## SCHOOL OF Engineering

### END TERM FINAL EXAMINATION

#### Extract of question distribution [outcome wise & level wise]

Q.NO	C.O.NO (% age of CO)	Unit/Module Number/Unit /Module Title	Memory recall type	Thought provoking type	Problem Solving type [Marks allotted]	Total Marks
			[Marks allotted] Bloom's Levels	[Marks allotted] Bloom's Levels		
			K	C	A	
1		Unit 1	2			
2		Unit 1	2			
3		Unit 2	2			
4		Unit 2	2			
5		Unit 3	2			
6		Unit 3	2			
7		Unit 4	2			
8		Unit 4	2			
9		Unit 5	2			
10		Unit 5	2			
11		Unit 4		10		
12		Unit 1,3	5	5		
13		Unit 2		10		
14		Unit 5			15	
15		Unit 4			15	
	Total Marks		25	25	30	

K = Knowledge Level C = Comprehension Level, A = Application Level

Note: While setting all types of questions the general guideline is that about 60% Of the questions must be such that even a below average students must be able to attempt, About 20% of the questions must be such that only above average students must be

able to attempt and finally 20% of the questions must be such that only the bright students must be able to attempt.

I hereby certify that all the questions are set as per the above guidelines.

Faculty Signature:

Reviewer Comment:

### Format of Answer Scheme



## SCHOOL OF ENGINEERING

### SOLUTION

Semester: 2019-2020

Course Code: CSE307

Course Name: Data Mining

Program & Sem: CSE, V Sem

Date: 20 Dec 2019

Time: 9:30AM-12:30PM

Max Marks: 80

Weightage: 40 %

#### Part A

(10Q x 2M = 20Marks)

Q No	Solution	Scheme of Marking	Max. Time required for each Question
1 a b c d	Ratio Nominal Ordinal Interval	0.5M For each	3Min
2	<b>Data Warehouse:</b> A Data warehouse is an integrated, subject-oriented and time variant repository of information in support of management's decision making process.	2M	3Min
3	-1.13, -0.62, 0.25, 1.50	0.5*4= 2M	4Min
4	Mutually exclusive rules <ul style="list-style-type: none"> <li>Classifier contains mutually exclusive rules if the rules are independent of each other</li> <li>Every record is covered by at most one rule</li> </ul> Exhaustive rules <ul style="list-style-type: none"> <li>Classifier has exhaustive coverage if it accounts for every possible combination of attribute values</li> </ul> Each record is covered by at least one rule	2M (0.5 for each difference)	4Min
5	Any nonempty subset of a frequent itemset must be frequent	2M	2Min
6	Training set is the information used to train an algorithm. The training data includes both input data and the corresponding expected output. Testing data, on the other hand, includes only	2M	



	input data, not the corresponding expected output. The testing data is used to assess how well your algorithm was trained, and to estimate model properties	(1+1)	3Min
7	<ul style="list-style-type: none"> <li>Scalability (Any two)</li> <li>Able to deal with noise and outliers</li> </ul>	2M (1+1)	2Min
8	$P(\text{Green} \text{yes}) = \frac{3}{5}$ $P(\text{Green} \text{no}) = \frac{2}{5}$ $P(\text{Red} \text{yes}) = \frac{2}{5}$ $P(\text{Red} \text{no}) = \frac{3}{5}$	2M (1M formula 1M Ans)	3Min
9	Web Mining is the use of the data mining techniques to automatically discover and extract information from web documents/services Discovering useful information from the World-Wide Web and its usage patterns	2M	3Min
10	Text Mining is used to extract relevant information or knowledge or pattern from different sources that are in unstructured or semi-structured form.	2M	3Min

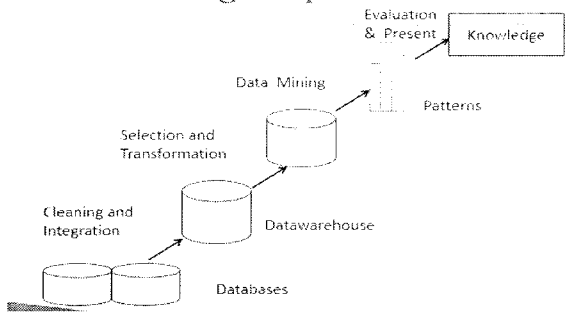
**Part B**

(3Q x 10M = 30 Marks)

Q No	Solution	Scheme of Marking	Max. Time required for each Question																
11	<p>Pay or Not</p> <table border="1"> <tr> <td>P(yes)</td> <td>5/10</td> </tr> <tr> <td>P(No)</td> <td>5/10</td> </tr> </table> <p>Box color</p> <table border="1"> <tr> <td>P(Green yes) = <math>\frac{3}{5}</math></td> <td>P(Green No) = <math>\frac{2}{5}</math></td> </tr> <tr> <td>P(Red yes) = <math>\frac{2}{5}</math></td> <td>P(Red No) = <math>\frac{3}{5}</math></td> </tr> </table> <p>Type</p> <table border="1"> <tr> <td>P(Nonveg yes) = <math>\frac{1}{5}</math></td> <td>P(Nonveg No) = <math>\frac{3}{5}</math></td> </tr> <tr> <td>P(veg yes) = <math>\frac{4}{5}</math></td> <td>P(veg No) = <math>\frac{2}{5}</math></td> </tr> </table> <p>Origin</p> <table border="1"> <tr> <td>P(Indian yes) = <math>\frac{2}{5}</math></td> <td>P(Indian No) = <math>\frac{3}{5}</math></td> </tr> <tr> <td>P(Mexican yes) = <math>\frac{3}{5}</math></td> <td>P(Mexican No) = <math>\frac{2}{5}</math></td> </tr> </table> <p>for sample X :- {Green, Nonveg, Indian}</p> $P(X \text{yes}) = P(\text{yes}) = P(\text{Green} \text{yes}) \times P(\text{Nonveg} \text{yes}) \times P(\text{Indian} \text{yes}) \times P(\text{yes})$ $= \frac{3}{5} \times \frac{1}{5} \times \frac{2}{5}$ $= 0.024$ $P(X \text{No}) = P(\text{No}) = P(\text{Green} \text{No}) \times P(\text{Nonveg} \text{No}) \times P(\text{Indian} \text{No}) \times P(\text{No})$ $= \frac{2}{5} \times \frac{3}{5} \times \frac{3}{5}$ $= 0.072$ <p>Here the probability of No is higher so for the test sample class will be No</p> <p>{Green, Nonveg, Indian} = "No"</p>	P(yes)	5/10	P(No)	5/10	P(Green yes) = $\frac{3}{5}$	P(Green No) = $\frac{2}{5}$	P(Red yes) = $\frac{2}{5}$	P(Red No) = $\frac{3}{5}$	P(Nonveg yes) = $\frac{1}{5}$	P(Nonveg No) = $\frac{3}{5}$	P(veg yes) = $\frac{4}{5}$	P(veg No) = $\frac{2}{5}$	P(Indian yes) = $\frac{2}{5}$	P(Indian No) = $\frac{3}{5}$	P(Mexican yes) = $\frac{3}{5}$	P(Mexican No) = $\frac{2}{5}$	10 M	26Mins
P(yes)	5/10																		
P(No)	5/10																		
P(Green yes) = $\frac{3}{5}$	P(Green No) = $\frac{2}{5}$																		
P(Red yes) = $\frac{2}{5}$	P(Red No) = $\frac{3}{5}$																		
P(Nonveg yes) = $\frac{1}{5}$	P(Nonveg No) = $\frac{3}{5}$																		
P(veg yes) = $\frac{4}{5}$	P(veg No) = $\frac{2}{5}$																		
P(Indian yes) = $\frac{2}{5}$	P(Indian No) = $\frac{3}{5}$																		
P(Mexican yes) = $\frac{3}{5}$	P(Mexican No) = $\frac{2}{5}$																		

# Data Mining Steps in KDD

12  
a)



5M

13Min

- Data Cleaning : Remove noise and inconsistent data
- Data Integration : Multiple data source combined
- Data Selection : Relevant data for analysis is retrieved
- Data Transformation : consolidation, aggregations, summary
- Data Mining : intelligent method to extract interesting patterns
- Pattern Evaluation: Truly interesting patterns.
- Knowledge Presentation : Visualization (charts),

12  
a)  
b)

Height (cm)  
65  
40  
85  
72  
50

→

$min_n = 40$   
 $max_n = 85$   
 $newmin_n = 1$   
 $newmax_n = 2$

a)  $= 65 = \frac{65-40}{45} (1) + 1 = 1.55$

b)  $40 = \frac{40-40}{45} (1) + 1 = 1$

c)  $85 = \frac{85-40}{45} (1) + 1 = 2$

d)  $72 = \frac{72-40}{45} (1) + 1 = 1.71$

e)  $50 = \frac{50-40}{45} (1) + 1 = 1.22$

5M

13Mins

13

$K_1$  (92, 36)       $K_2$  (85, 28)

for point 3

$\rightarrow K_1 = \sqrt{(92-84)^2 + (36-30)^2} = 10$   
 $\rightarrow K_2 = \sqrt{(85-84)^2 + (28-30)^2} = 2.23$

new centroid for  $K_2$

$K_1$  (92, 36)       $K_2$  (84.5, 29)

$K_2 = \left( \frac{85+84}{2}, \frac{28+30}{2} \right)$   
 $K = (84.5, 29)$

for point 4 (89, 34)

$K_1 = \sqrt{(92-89)^2 + (36-34)^2} = 2.82$   
 $K_2 = \sqrt{(84.5-89)^2 + (29-34)^2} = 7.07$

It goes in  $(K_1)$

10M

26Mins

Point 5 (91, 36)

$$\text{distance to } K_1 = \sqrt{(91-90.5)^2 + (36-35)^2}$$

$$= 1.1$$

$$\text{distance to } K_2 = \sqrt{(91-84.5)^2 + (36-29)^2}$$

$$= 9.55$$

It goes to  $K_1$

new centroid for  $K_1$

$$\left( \frac{92+89+91}{3}, \frac{36+34+36}{3} \right)$$

$$= 90.66, 35.33$$

centroid for  $K_2$

$$= (84.5, 29)$$

Point 6 (94, 38)

$$\text{distance } K_1 = \sqrt{(94-90.66)^2 + (38-35.33)^2}$$

$$= 4.27$$

$$\text{distance } K_2 = \sqrt{(94-84.5)^2 + (38-29)^2}$$

$$= 12.45$$

$$K_1 = \{ 1, 4, 5, 6 \}$$

$$K_2 = \{ 2, 3 \}$$

### Part C

(2Q x 15M = 30Marks)

Q No	Solution	Scheme of Marking	Max. Time required for each Question																																																													
14	<table border="1"> <thead> <tr> <th></th> <th>M1</th> <th>M2</th> <th>M3</th> <th>M4</th> <th>M5</th> </tr> </thead> <tbody> <tr> <th>M1</th> <td>0</td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <th>M2</th> <td>18</td> <td>0</td> <td></td> <td></td> <td></td> </tr> <tr> <th>M3</th> <td>6</td> <td>14</td> <td></td> <td></td> <td></td> </tr> <tr> <th>M4</th> <td>12</td> <td>10</td> <td>18</td> <td>0</td> <td></td> </tr> <tr> <th>M5</th> <td>22</td> <td>20</td> <td>4</td> <td>16</td> <td>0</td> </tr> </tbody> </table> <p>Min = 4 (Thus [M3, M5] comes in same cluster)</p> <table border="1"> <thead> <tr> <th></th> <th>M1</th> <th>M2</th> <th>[M3, M5]</th> <th>M4</th> </tr> </thead> <tbody> <tr> <th>M1</th> <td>0</td> <td></td> <td></td> <td></td> </tr> <tr> <th>M2</th> <td>18</td> <td>0</td> <td></td> <td></td> </tr> <tr> <th>[M3, M5]</th> <td>6</td> <td>14</td> <td>0</td> <td></td> </tr> <tr> <th>M4</th> <td>12</td> <td>10</td> <td>16</td> <td>0</td> </tr> </tbody> </table> <p><math>d(M1, [M3, M5])</math>  <math>= \min(d(M1, M3), d(M1, M5))</math>  <math>= \min(6, 22)</math>  <math>= 6</math></p> <p><math>d(M2, [M3, M5])</math>  <math>= \min(d(M2, M3), d(M2, M5))</math>  <math>= \min(14, 20)</math>  <math>= 14</math></p> <p><math>d(M4, [M3, M5])</math>  <math>= \min(d(M4, M3), d(M4, M5))</math>  <math>= \min(18, 16)</math>  <math>= 16</math></p> <p>∴ the above matrix minimum is 6          So [M3, M5, M1] comes in cluster</p>		M1	M2	M3	M4	M5	M1	0					M2	18	0				M3	6	14				M4	12	10	18	0		M5	22	20	4	16	0		M1	M2	[M3, M5]	M4	M1	0				M2	18	0			[M3, M5]	6	14	0		M4	12	10	16	0	15M	35Min
	M1	M2	M3	M4	M5																																																											
M1	0																																																															
M2	18	0																																																														
M3	6	14																																																														
M4	12	10	18	0																																																												
M5	22	20	4	16	0																																																											
	M1	M2	[M3, M5]	M4																																																												
M1	0																																																															
M2	18	0																																																														
[M3, M5]	6	14	0																																																													
M4	12	10	16	0																																																												

Thus proximity Matrix will be

	[M1, M3, M5]	M2	M4
[M1, M3, M5]	0		
M2	14	0	
M4	12	10	0

$$= d(M2(M1, M3, M5))$$

$$= \min [d(M2, M1), d(M2, M3), d(M2, M5)]$$

$$= \min (18, 14, 20)$$

$$= 14$$

$$= d(M4(M1, M3, M5))$$

$$= \min [d(M4, M1), d(M4, M3), d(M4, M5)]$$

$$= \min (12, 18, 16)$$

$$= 12$$

Again after choosing minimum i.e. 10 in the above matrix, [M2, M4] will become a cluster.

	[M1, M3, M5]	[M2, M4]
[M1, M3, M5]	0	
[M2, M4]	12	0

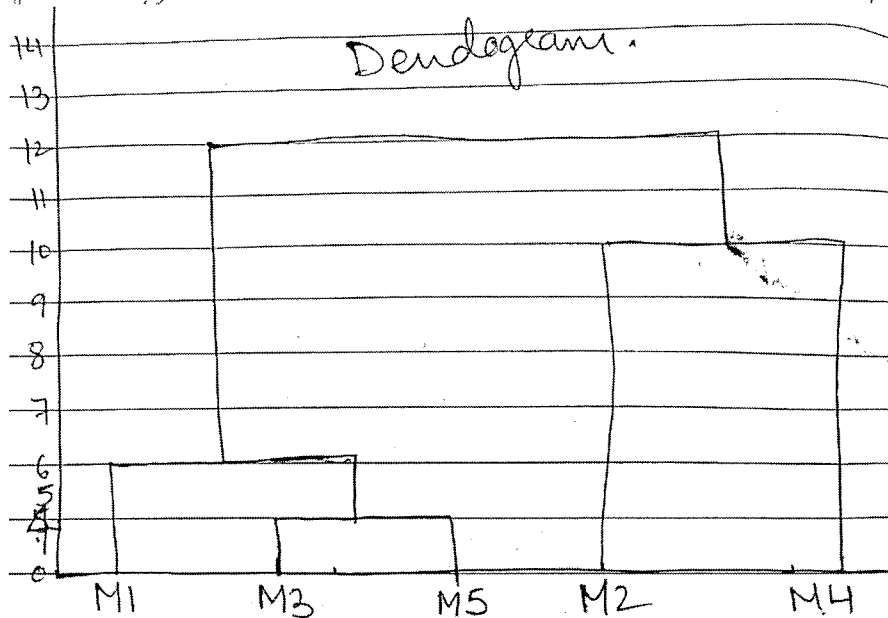
$$d([M1, M3, M5], [M2, M4])$$

$$= \min [d(M1, M2), d(M3, M2), d(M5, M2), d(M1, M4), d(M3, M4), d(M5, M4)]$$

$$= \min (18, 14, 20, 12, 18, 16)$$

$$\min = 12$$

Finally  $\min = 12$  so last cluster [M1, M3, M5, M2, M4].



## Q.2 Solution

Cost. No	Age Group	Income	Credit Range	Buy's-comp.
1	Young	High	Weak	Yes
2	Senior	Low	Weak	Yes
3	Young	High	Weak	Yes
4	Young	High	Weak	Yes
5	Young	Low	Weak	No
6	Young	Low	Strong	No

$$\text{Entropy (S)} = \sum_{i=1}^n P_i \log_2 P_i$$

$$\text{Entropy (4Y, 2N)} = -\frac{4}{6} \log_2 \left(\frac{4}{6}\right) - \left(\frac{2}{6}\right) \log_2 \left(\frac{2}{6}\right)$$

$$\text{Entropy} = 0.9182$$

• For Attribute Age Group

Age G	P <sub>i</sub>	n <sub>i</sub>	I
Young	3	2	0.966
Senior	1	0	0

$$I = -\frac{3}{5} \log_2 \left(\frac{3}{5}\right) - \frac{2}{5} \log_2 \left(\frac{2}{5}\right)$$

$$\text{Information Gain} = 0.966$$

$$E = \left(\frac{3+2}{4+2}\right) \times 0.966 + \left(\frac{1+0}{4+2}\right) \times 0$$

$$E = 0.833 \times 0.966$$

$$E = 0.804$$

$$\text{Gain} = 0.9182 - 0.804$$

$$\text{Gain (S, Age Group)} = 0.1142$$

• For Attribute Income

Income	P <sub>i</sub>	n <sub>i</sub>	I
High	3	0	0
Low	1	2	0.915

$$\text{Information Gain (1,2)} = -\frac{1}{3} \log_2 \left(\frac{1}{3}\right) -$$

$$\left(\frac{2}{3}\right) \log_2 \left(\frac{2}{3}\right)$$

$$I = 0.915$$

$$E = \left(\frac{1+2}{4+2}\right) \times 0.915 + 0$$

$$E = 0.457$$

$$\text{Gain} = 0.9182 - 0.457$$

$$\text{Gain (S, Income)} = 0.461$$

• For attribute Credit Range

CR	P <sub>i</sub>	n <sub>i</sub>	I
Weak	4	1	
Strong	0	1	0

$$I = -\frac{1}{5} \log_2 \left(\frac{1}{5}\right) - \left(\frac{4}{5}\right) \log_2 \left(\frac{4}{5}\right)$$

$$I = 0.720$$

$$E = \left( \frac{11+1}{142} \right) \times 0.720 + 0$$

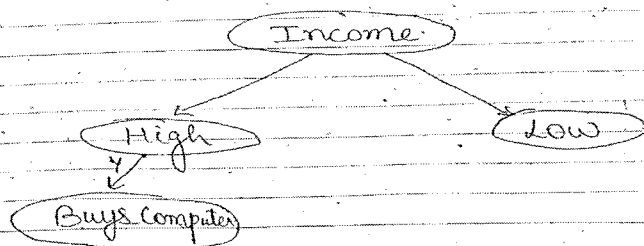
$$= 0.6$$

$$\text{Gain} = 0.918 - 0.6$$

$$\text{Gain}(S, \text{CreditR}) = 0.318$$

- $\text{Gain}(S, \text{AgeGroup}) = 0.1142$
- $\text{Gain}(S, \text{Income}) = 0.461$
- $\text{Gain}(S, \text{CreditRange}) = 0.318$

So Maximum Gain is from Attribute Income, so it is considered as root Node.



As all high are yes. so, it will Buy Computer

When Income is low, we are not finding one appropriate decision so we will repeat the process, with respect to low income attribute

Cust. No	Age Group	Income	Credit Range	Buys Comp.
2	senior	low	Weak	yes
5	young	low	Weak	NO
6	young	low	Strong	NO

For Entropy (S) Here, yes = 1  
NO = 2.

$$E(S) = -\frac{1}{3} \log_2 \left( \frac{1}{3} \right) - \frac{2}{3} \log_2 \left( \frac{2}{3} \right)$$

$$E(S) = 0.915$$

For attribute Age Group

AG	P <sub>i</sub>	n <sub>i</sub>	I
Senior	1	0	0
young	0	2	0

$$I = 0$$

$$E = 0$$

$$\text{Gain} = 0.915 - 0$$

$$\text{Gain}(S, \text{Age group}) = 0.915$$

For attribute Credit Range

Credit R <sub>o</sub>	P <sub>i</sub>	n <sub>i</sub>	I
Weak	1	1	1
Strong	0	1	0

$$I = -\frac{1}{2} \log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right) \log_2\left(\frac{1}{2}\right)$$

$$I = 1$$

$$\text{Gain } E = \left(\frac{1+1}{1+2}\right) \times 1$$

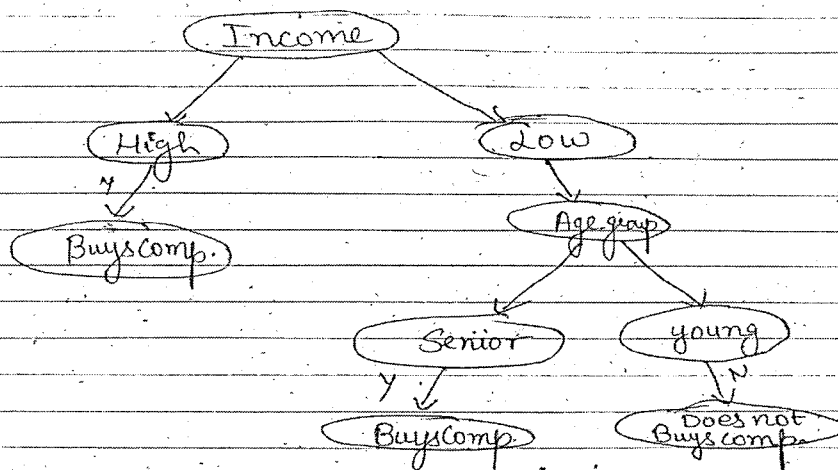
$$\text{Gain } E = 0.66$$

$$\text{Gain} = 0.915 - 0.66$$

$$\text{Gain}(S, CR) = 0.255$$

Gain Age group is Maximum.

So final Decision tree will be



1. The first question has 2 parts

