CT2 set B - Nothing much

Data Mining And Analytics (SRM Institute of Science and Technology)

SRM Institute of Science and Technology
College of Engineering and Technology
School of Computing
SRM Nagar, Kattankulathur – 603203, Chengalpattu District, Tamilnadu
Academic Year: 2023-24 (ODD)

Mode of Exam
OFFLINE
SET B

Test: CLA-T2 | **Date: 26-09-2023**
Course Code & Title: 18CSC355T Data Mining and Analytics | **Duration:** 2 Hour
Year & Sem: III Year / V Sem | **Max. Marks:** 50

Course Articulation Matrix: ***(to be placed)***

S.No.	Course Outcome	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
1	**CO1**	L	H		H	L				L	L		H
2	**CO2**	M	H		H	L				M	L		H
3	**CO3**	M	H		H	L				M	L		H
4	**CO4**	M	H		H	L				M	L		H
5	**CO5**	H	H		H	L				M	L		H

Part - A
(10 x 1 = 10 Marks)
Instructions: Answer all

Q. No	Question	Marks	BL	CO	PO	PI Code
1	Which algorithm is used for frequent itemset mining? a) Decision tree algorithm b) K-nearest neighbors algorithm c) **Apriori algorithm** d) Naive Bayes algorithm	**1**	**L1**	**2**	**1**	**1.7.1**
2	Frequency of occurrence of an item set is called as ________ a) Support b) Confidence c) **Support Count** d) Rules	**1**	**L1**	**2**	**1**	**1.7.1**
3	In a supermarket there were 100 transactions. In that 20 transactions have bread, out of 20 transactions butter occurs in 8 transactions. So what is the confidence percentage for butter? a) 20 Percentage **b) 40 Percentage** c) 45 Percentage d) 8 Percentage	**1**	**L2**	**2**	**2**	**2.6.3**
4	When do you consider an association rule interesting ? a) If it only satisfies min_support b) If it only satisfies min_confidence c) **If it satisfies both min _support and min _confidence** d) There are other measures to check so	**1**	**L1**	**2**	**2**	**2.6.3**
5	How do you calculate Confidence (A -> B)? **a) Support(A ∩ B) / Support (A)** b) Support(A ∩ B) / Support (B) c) Support(A ∪ B) / Support (A) d) Support(A ∪ B) / Support (B)	**1**	**L2**	**2**	**1**	**1.7.1**
6	You are given data about seismic activity in the United States, and you want to predict the magnitude of the upcoming earthquake. This can be considered as an example of which of	**1**	**L2**	**3**	**1**	**1.7.1**

	the following methods? a) **Supervised learning** b) Unsupervised learning c) Serration d) Dimensionality reduction					
7	In some cases, telecommunication companies desire to segment their clients into distinct groups in order to send suitable and related subscription offer. This can be considered as an example of which of the following methods? a) Supervised learning **b)Unsupervised learning** c) Serration d). Data extraction	**1**	**L2**	**3**	**5**	**5.4.1**
8	Suppose your classification model predicted true for a class which actual value was false. Then this is a- **a) False positive** b) False negative c) True positive d) True negative	**1**	**L2**	**3**	**1**	**1.7.1**
9	The true positive value is 10 and the false positive value is 15. Calculate the value of precision a) 0.6 **b). 0.4** c) 0.5 d). None	**1**	**L3**	**3**	**1**	**1.7.1**
10	Which one of the following is the main reason for pruning a Decision Tree? a).to save computing time during testing b).to save space for storing the decision tree c).to make the training set error smaller **d).to avoid overfitting the training set**	**1**	**L2**	**1**	**1**	**1.7.1**

Mode of Exam
OFFLINE
SET B

SRM Institute of Science and Technology
College of Engineering and Technology
School of Computing

SRM Nagar, Kattankulathur – 603203, Chengalpattu District, Tamilnadu

Academic Year: 2021-22 (ODD)

Test: CLA-T2 **Date: 26-09-2023**
Course Code & Title: 18CSC355T Data Mining and Analytics **Duration:** 2 Hour
Year & Sem: III Year / V Sem **Max. Marks:** 50

Course Articulation Matrix: ***(to be placed)***

S.No.	Course Outcome	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
1	**CO1**	L	H		H	L				L	L		H
2	**CO2**	M	H		H	L				M	L		H
3	**CO3**	M	H		H	L				M	L		H
4	**CO4**	M	H		H	L				M	L		H
5	**CO5**	H	H		H	L				M	L		H

PART B (5x 4 = 20 Marks)

Q. No	Question	Marks	BL	CO	PO	PI Code
11	What are the limitations of the Apriori algorithm and how may it be made more efficient? Limitations of the Apriori algorithm The main limitation is the costly wasting of time to hold many candidates sets with frequent itemsets, low minimum support, or large itemsets. A large amount of data needs to be stored in memory for processing, so large transaction items require much more resources. Improving the Efficiency of Apriori • Transaction Reduction(reducing the number of transactions scanned in future iterations • Partitioning(partitioning the data to find candidate itemsets): • Sampling(mining on a subset of the given data): • Dynamic itemset counting (adding candidate itemsets at different points during a scan): • Hash-based technique (hashing itemsets into corresponding buckets):	5	**L2**	2	2	2.8.2
12	Explain about multilevel association rules and their purpose. Also, list and brief the many forms of multilevel association rules.	**5**	**L1**	**2**	**1**	**1.2.2**

<table>
<tr><td></td><td>

Multilevel Association Rules

- When transactions data is taken for link analysis. It is present at the low level of abstraction that is detail form.
- It is very difficult to form association rules at the low level of abstraction as data scarcity is there. Also resultant rules can not efficiently used.
- Using concept hierarchies, transaction data can be represented at various levels of abstraction.
- In Multilevel Association Rules, association rules are generated at multiple levels of abstraction
- Instead of going at lower level of abstraction, association rules are generated from higher level of abstraction which represents common sense knowledge and be use used efficiently.

Need of Multiple-Level Association Rules?

- Sometimes at low data level, data does not show any significant pattern. But there are useful information hiding behind.
- Aim is to find the hidden information in or between levels of abstraction

Three ways

1) **Uniform Support (Using uniform minimum support for all levels)**
2) **Reduced support (Using reduced minimum support at lower levels)**
3) **Group-based support (Using item or group based minimum support)**

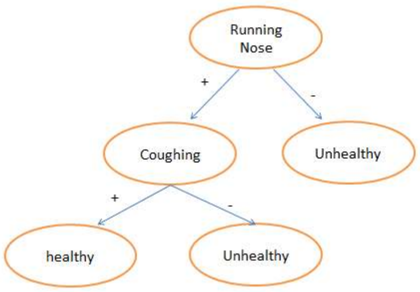
</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>13</td><td>

Brief about the Metrics for Evaluating classifier Performance

<table>
<tr><th>Measure</th><th>Formula</th></tr>
<tr><td>accuracy, recognition rate</td><td>$\frac{TP+TN}{P+N}$</td></tr>
<tr><td>error rate, misclassification rate</td><td>$\frac{FP+FN}{P+N}$</td></tr>
<tr><td>sensitivity, true positive rate, recall</td><td>$\frac{TP}{P}$</td></tr>
<tr><td>specificity, true negative rate</td><td>$\frac{TN}{N}$</td></tr>
<tr><td>precision</td><td>$\frac{TP}{TP+FP}$</td></tr>
<tr><td>F, F_1, F-score, harmonic mean of precision and recall</td><td>$\frac{2 \times precision \times recall}{precision + recall}$</td></tr>
<tr><td>F_β, where β is a non-negative real number</td><td>$\frac{(1+\beta^2) \times precision \times recall}{\beta^2 \times precision + recall}$</td></tr>
</table>

</td><td>5</td><td>L1</td><td>3</td><td>2</td><td>2.8.2</td></tr>
<tr><td>14</td><td>

Explain about decision tree and write all the possible rules from the following Decision tree.



A decision tree is a flowchart-like tree structure where each internal node denotes the feature, branches denote the rules and the leaf nodes denote

</td><td>5</td><td>L2</td><td>3</td><td>2</td><td>2.6.4</td></tr>
</table>

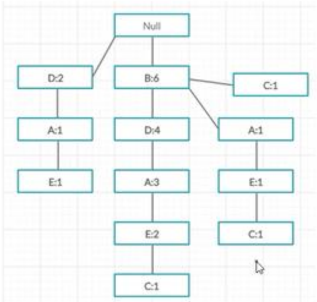
	the result of the algorithm. It is a versatile supervised machine-learning algorithm, which is used for both classification and regression problems Root Node: It is the topmost node in the tree, which represents the complete dataset. It is the starting point of the decision-making process. Decision/Internal Node: A node that symbolizes a choice regarding an input feature. Branching off of internal nodes connects them to leaf nodes or other internal nodes. Leaf/Terminal Node: A node without any child nodes that indicates a class label or a numerical value. Splitting: The process of splitting a node into two or more sub-nodes using a split criterion and a selected feature. Branch/Sub-Tree: A subsection of the decision tree starts at an internal node and ends at the leaf nodes. Parent Node: The node that divides into one or more child nodes. Child Node: The nodes that emerge when a parent node is split. R1: If(Running Nose=-) then Status=Unhealthy R2: If(Running Nose=+) AND (Coughing=+) then Status=Healthy R2: If(Running Nose=+) AND (Coughing=-) then Status=Unhealthy					
15	Is clustering unsupervised or supervised classification? Give the reason for your answer. Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be “the process of organizing objects into groups whose members are similar in some way”. A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.	**5**	**L2**	**3**	**2**	2.6.4

PART C (2 x 10 = 20 Marks)

16	i.Built the FP tree for the given transactions , Find all frequent item sets from conditional pattern base of each node using Fp growth algorithm and the minimum support is 30% (7 marks)	10	L3	2	2	2.8.2

TID	List of Items
T1	E,A,D,B
T2	D,A,C,E,B
T3	C,A,B,E
T4	B,A,D
T5	D
T6	D,B
T7	A,D, E
T8	B,C

ii .Explain the support and confidence (3 marks)



TID	List of Items
T1	B, D, A, E
T2	B, D, A, E, C
T3	B, A, E, C
T4	B, D, A
T5	D
T6	B, D
T7	D, A, E
T8	B, C

Frequent pattern are

C:3,CB:3

E:4, DE:3, AE:4, BE:3, ADE:3, ABE:3

A:5, AD:5, AB:4, ABD:3

D:6, BD:4

B:6

(OR)

17	Consider the Dataset for finding frequency pattern using Apriori Algorithm Find the frequent item sets and generate association rules on this. Assume that minimum support threshold (s = 30%) and minimum confident threshold (c = 70%)	**10**	**L3**	**3**	**4**	**4.4.2**

TID	List of Items
T1	E,A,D,B
T2	D,A,C,E,B
T3	C,A,B,E

T4	B,A,D
T5	D
T6	D,B
T7	A,D, E
T8	B,C

Step 1: Calculate Minimum support.

Minimum support count (30/100 * 8) = 2.4=3

Step 3: Generating 3-itemset frequent pattern.

L2

Itemset	Sup. Count
A, B	4
A, D	4
A, E	4
B, D	4
B, E	3
D, E	3

C3

Itemset	Sup. Count
A, B, D	3
A, B, E	3
A, D, E	3
B, D, E	2

Generate C3 from L2 and scan D for count of each candidate

L3

Itemset	Sup. Count
A, B, D	3
A, B, E	3
A, D, E	3

Compare candidate sup. count with min. Sup. count

{A, B, D}

Confidence =75%

Association Rule	Confidence	Confidence (%)
A ^ B→ D	C(A, B, D)/C(A, B)=3/4	75%
A^ D →B	C(A, B, D)/C(A, D)=3/4	75%
B ^ D→ A	C(A, B, D)/C(B, D)=3/4	75%
A → B ^ D	C(A, B, D)/C(A)=3/5	60%
B → A ^ D	C(A, B, D)/C(B)=3/6	50%
D → A ^ B	C(A, B, D)/C(Hotdogs)=3/6	50%

{A,B,E}

{A,D,E}

18 Construct Decision tree and Rules for the given dataset using Information Gain. **10** **L3** **3** **8** 8.4.1

Age	Income	Student	Credit Rating	Buys_Computer
<=30	High	No	Fair	No
<=30	High	No	Excellent	No
31..40	High	No	Fair	Yes
>40	Medium	No	Fair	Yes
>40	Low	Yes	Fair	Yes
>40	Low	Yes	Excellent	No
31..40	Low	Yes	Excellent	Yes
<=30	Medium	No	Fair	No
<=30	Low	Yes	Fair	Yes
>40	Medium	Yes	Fair	Yes
<=30	Medium	Yes	Excellent	Yes
31..40	Medium	No	Excellent	Yes
31..40	High	Yes	Fair	Yes
>40	Medium	No	Excellent	No

Information gain for root node

Attributes	Gain
Age	**0.25**
Income	0.03
Student	0.15
Credit rating	0.05

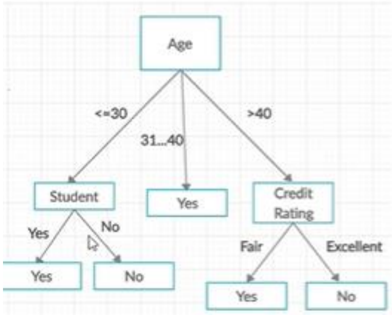
Information gain for second left node

Attributes	Gain
Income	0.57
student	**0.97**
Credit rating	0.02

Information gain for second Right node

Attributes	Gain
Income	0.02
student	0.02
Credit rating	**0.97**

Decision Tree



Decision Tree Rules

R1: If(Age=31...40) then Buys_Computer=Yes

R2: If(Age<=30) AND (Student=Yes) then Buys_Computer=Yes

R3: If(Age<=30) AND (Student=No) then Buys_Computer=No

R4: If(Age=>40) AND (Credit Rating=Fair) then Buys_Computer=Yes

R5: If(Age=>40) AND (Credit Rating=Excellent) then Buys Computer=No

(OR)

19	i. Create the naive bayes model from the given dataset and the target class is stolen ii. find the target class (stolen) for the unseen data x= (red ,Sports and Domestic) iii. find the target class (stolen) for the unseen data x= (yellow ,SUV and imported)	**10**	**L3**	**3**	**8**	8.4.1

Car no.	Color	Type	Origin	Stolen
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Prior Probability

P(Yes)=**5/10** P(No)=**5/10**

Likelihood

Classification→	Yes (5)	No(5)
Color		
Red (5)	3/5	2/5
Yellow (5)	2/5	3/5

Classification→	Yes(5)	No(5)
Type		
Sports(6)	4/5	2/5
SUV (4)	1/5	3/5

Classification→	Yes(5)	No(5)
Origin		
Domestic(5)	2/5	3/5
Imported(5)	3/5	2/5

Testing Model **Unseen Sample (X)=<Red , Sports, Domestic>**

$$P(Yes \mid X) = P(X \mid Yes) \times P(YES)$$
$$= P(Red \mid Yes) \times P(Sports \mid Yes) \times P(Domestic \mid Yes) \times P(Yes)$$
$$= \frac{3}{5} \times \frac{4}{5} \times \frac{2}{5} \times \frac{5}{10} = 0.0960$$

$$P(No \mid X) = P(X \mid No) \times P(No)$$
$$= P(Red \mid No) \times P(Sports \mid No) \times P(Domestic \mid No) \times P(No)$$
$$= \frac{2}{5} \times \frac{2}{5} \times \frac{3}{5} \times \frac{5}{10} = 0.0480$$

P(Yes)>P(No)

Unseen sample(X) is classified as yes. That is prediction for car stolen is yes.

Testing Model **Unseen Sample (X)=<Yellow , SUV, Imported>**

$$P(Yes \mid X) = P(X \mid Yes) \times P(YES)$$
$$= P(Yellow \mid Yes) \times P(SUV \mid Yes) \times P(Imported \mid Yes) \times P(Yes)$$
$$= \frac{2}{5} \times \frac{1}{5} \times \frac{3}{5} \times \frac{5}{10} = 0.0240$$

$$P(No \mid X) = P(X \mid No) \times P(No)$$
$$= P(Yellow \mid No) \times P(SUV \mid No) \times P(Imported \mid No) \times P(No)$$
$$= \frac{3}{5} \times \frac{3}{5} \times \frac{2}{5} \times \frac{5}{10} = 0.0720$$

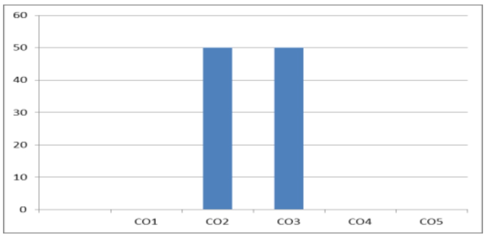
P(No)>P(Yes)

Unseen sample(X) is classified as no. That is prediction of car stolen is no.

***Program Indicators are available separately for Computer Science and Engineering in AICTE examination reforms policy.**

Course Outcome (CO) and Bloom's level (BL) Coverage in Questions

CO Coverage in %



BL Coverage in %