Reg. No.															

SRM Institute of Science and Technology

College of Engineering and Technology

School of Computing

SETA

DEPARTMENT OF COMPUTING TECHNOLOGIES

SRM Nagar, Kattankulathur – 603203, Chengalpattu District, Tamilnadu

Academic Year: 2023-2024(ODD)

Test: CLAT-2 **Date:** 07.10.2023
Course Code & Title: 18CSE355T - Data Mining and Analytics **Duration:** 100 Minutes
Year & Sem: III &IV Year & 05th & 07th Semester **Max. Marks:** 50 Marks

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

S. No.	Course Outcome	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO1 0	PO1 1	PO1 2
1	CO1	H											
2	CO2	H		M									
3	CO3	H	M	H									
4	CO4	H	M	H									
5	CO5	H	H	H	L								

Part – A
(10 x 1 = 10 Marks)
Answer all questions. The duration for answering the part A is 20 minutes (MCQ Answer sheet will be collected after 20 minutes)

Q. No	Question	Marks	BL	CO	PO	PI Code
1	Which algorithm requires fewer scans of data? **a) FP Growth** b) Naïve Bayes c) Apriori d) Decision Tree	**1**	**1**	**2**	**1**	**1.7.1**
2	What does FP growth algorithm do? a) It mines all frequent patterns through pruning rules with lesser support b) It mines all frequent patterns through pruning rules with higher support **c) It mines all frequent patterns by constructing a FP tree** d) It mines all frequent patterns by constructing an item sets	**1**	**1**	**2**	**1**	**1.7.1**
3	You are a Data Scientist in an e-commerce company. You are analyzing all the transactions that happened over the past 1 week in your site. You observe that of the five hundred transactions that happened, two hundred of them had a mobile phone in them. What is the support for mobile phones in the last 1 week?	**1**	**3**	**2**	**1**	**1.7.1**

	a) 0.3 **b) 0.4** C) 0.5 d) 0.6					
4	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**	**2**	**2**	**1**	**1.7.1**
5	What techniques can be used to improve the efficiency of apriori algorithm? **a)Hash-based techniques** b)Transaction Increases c)Sampling d)Cleaning	**1**	**1**	**2**	**1**	**1.7.1**
6	The classification or mapping of a class using a predefined class or group is called: a) Data Sub Structure b) Data Set **c) Data Discrimination** d) Data Characterization	**1**	**1**	**3**	**2**	**2.5.2**
7	______________models continuous valued functions. **a) Prediction** **b)** Back Propagation **c)** Classification **d)** Data trends	**1**	**1**	**3**	**2**	**2.5.2**
8	_______________is a statistical methodology that is most often used for numeric prediction **a) Regression analysis** b) Classification c) Class labels analysis d) decision tree classifiers	**1**	**1**	**3**	**2**	**2.5.2**
9	________________can be used to identify whether any two given attributes are statistically related. a) Relevance Analysis b) Regression Analysis c) Attribute subset selection **d) Correlation analysis**	**1**	**1**	**3**	**2**	**2.5.2**
10	Zero Probability value can be avoided using _____________. a) Decision Trees b) If then Classification **c) Laplacian smoothing** d) Naïve Bayesian Classification	**1**	**1**	**3**	**2**	**2.5.2**

Part – B
(4 x 5 = 20 Marks)
Answer any 4 Questions

11

Consider the horizontal data format of the transaction database, D of a company. Show the transformed vertical data format. Mining can be performed on this data set by intersecting the TID sets of every pair of frequent single items. The minimum support count is 2. Because every single item is frequent in D.

TID	**LIST OF ITEM**
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Table: I -Transactional Database ‘D’ for a company.

Table 6.3: The vertical data format of the transaction data set D of Table 6.1.

itemset	*TID_set*
I1	{T100, T400, T500, T700, T800, T900}
I2	{T100, T200, T300, T400, T600, T800, T900}
I3	{T300, T500, T600, T700, T800, T900}
I4	{T200, T400}
I5	{T100, T800}

This data format is known as the horizontal data format. Alternatively, data can also be presented in item-TID set format (that is, {item : T ID set}), where item is an item name, and TID set is the set of transaction identifiers containing the item. This format is known as the vertical data format.

Mining frequent itemsets using vertical data format. Consider the horizontal data format of the transaction database, D, of Table 6.1 in Example 6.3. This can be transformed into the vertical data format shown in Table 6.3 by scanning the data set once. Mining can be performed on this data set by intersecting the TID sets of every pair of frequent single items. The minimum support count is 2. Because every single items frequent in Table 6.3, there are 10 intersections performed in total, which lead to 8 nonempty 2-itemsets as shown in Table 6.4. Notice that because the item sets {I1, I4} and {I3, I5} each contain only one transaction, they do not belong to the set of frequent 2-itemsets. Based on the Apriori property, a given 3-itemset is a candidate 3-itemset only if every one of its 2-itemset subsets is frequent. The candidate generation process here will generate only two 3-itemsets: {I1, I2, I3} and {I1, I2, I5}. By intersecting the TID sets of any two corresponding 2-itemsets of these candidate

5	2	2	8	8.4.1

Table 6.4: The 2-itemsets in vertical data format.

itemset	*TID_set*
{I1, I2}	{T100, T400, T800, T900}
{I1, I3}	{T500, T700, T800, T900}
{I1, I4}	{T400}
{I1, I5}	{T100, T800}
{I2, I3}	{T300, T600, T800, T900}
{I2, I4}	{T200, T400}
{I2, I5}	{T100, T800}
{I3, I5}	{T800}

Table 6.5: The 3-itemsets in vertical data format.

itemset	*TID_set*
{I1, I2, I3}	{T800, T900}
{I1, I2, I5}	{T100, T800}

12 What is Frequent Pattern Mining? Give example.

Explanation 3 Marks + Example 2 Marks

Frequent Pattern Mining (AKA Association Rule Mining) is an analytical process that finds frequent patterns, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other data repositories. Given a set of transactions, this process aims to find the rules that enable us to predict the occurrence of a specific item based on the occurrence of other items in the transaction.

Let's look at an example of Frequent Pattern Mining. First, we will want to understand the terminology used in this type of analysis. While there are numerous metrics and factors used in this technique, for this example, we will only consider two factors namely, Support and Confidence.

Support: The support of a rule x -> y (where x and y are each items/events etc.) is defined as the proportion of transactions in the data set which contain the item set x as well as y. So, Support (x -> y)= no. of transactions which contain the item set x & y / total no. of transactions.

Confidence: The confidence of a rule x -> y is defined as: Support (x -> y) / support (x). So, it is the ratio of the number of transactions that include all items in the consequent (y in this case), as well as the antecedent (x in this case) to the number of transactions that include all items in the antecedent (x in this case).

In the table below, Support (milk->bread) = 0.4 means milk and bread are purchased together occur in 40% of all transactions. Confidence (milk->bread) = 0.5 means that if there are 100 transactions containing milk then there will be 50 that will also contain bread.

5	**2**	**2**	**1**	**1.7.1**

TID	Milk	Bread	Butter	Beer
1	1	0	1	1
2	1	1	1	0
3	0	1	1	0
4	1	0	0	1
5	1	1	1	1

13 Bring out advantage of association rule mining in data mining.

In data mining, association rules are useful for analyzing and predicting customer behavior. They play an important part in customer analytics, market basket analysis, product clustering, catalog design and store layout. Programmers use association rules to build programs capable of machine learning.

1. An estimation is used in the algorithm to prune those candidate itemsets that have no hope to be large.
2. It is suitable for low cardinality sparse transaction database
3. This algorithm has least memory consumption.
4. Easy implementation.
5. It uses Apriori property for pruning therefore, itemsets left for further support checking remain less.
6. It uses compressed representation of original database.
7. Repeated database scan is eliminated

5	**2**	**2**	**1**	**1.7.1**

14 Write algorithm for decision tree induction

Generating a decision tree form training tuples of data partition D

Algorithm : Generate_decision_tree

Input:

Data partition, D, which is a set of training tuples and their associated class labels. attribute_list, the set of candidate attributes. Attribute selection method, a procedure to determine the splitting criterion that best partitions that the data tuples into individual classes. This criterion includes a splitting_attribute and either a splitting point or splitting subset.

Output:

A Decision Tree

Method

create a node N;

if tuples in D are all of the same class, C then
   return N as leaf node labeled with class C;

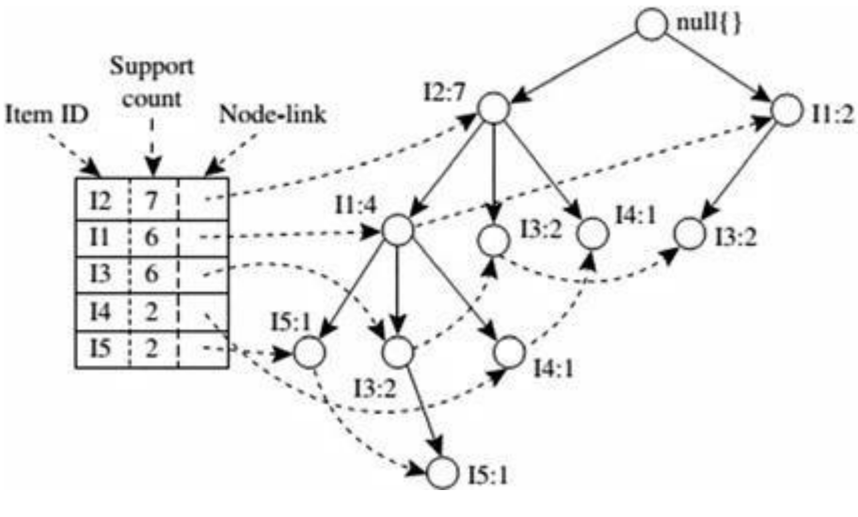
5	**2**	**3**	**2**	**2.6.4**

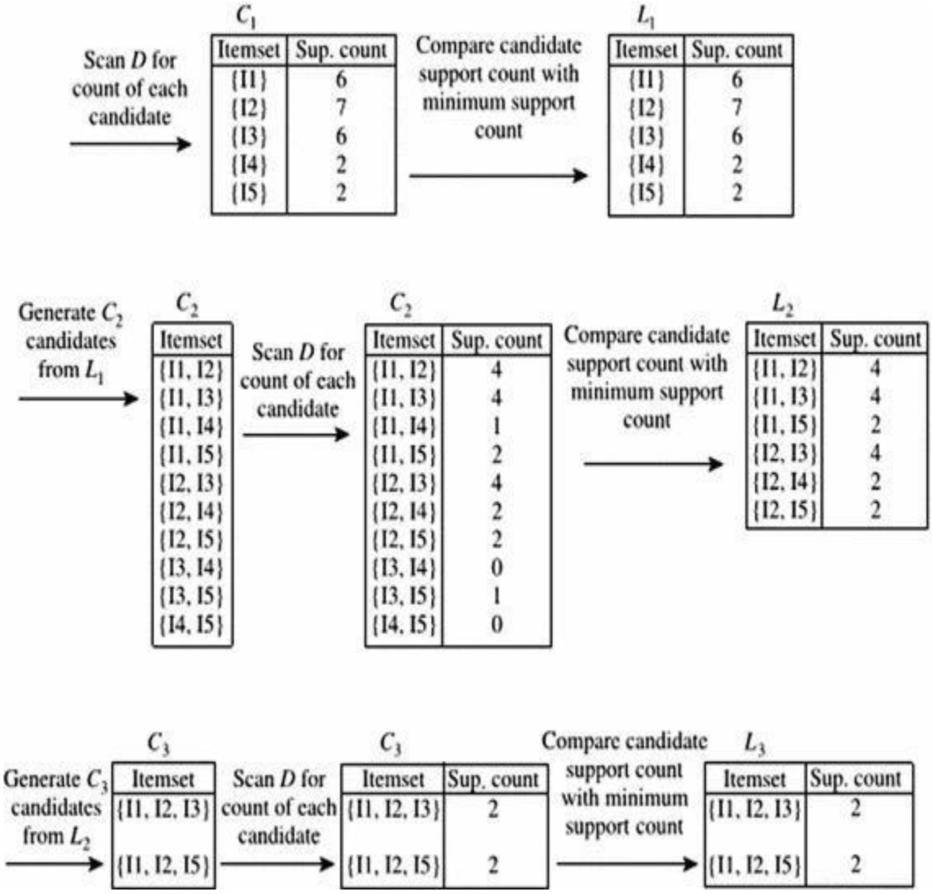
	if attribute_list is empty then
	return N as leaf node with labeled
	with majority class in D;|| majority voting
	
	apply attribute_selection_method(D, attribute_list)
	to find the best splitting_criterion;
	label node N with splitting_criterion;
	
	if splitting_attribute is discrete-valued and
	multiway splits allowed then // no restricted to binary trees
	
	attribute_list = splitting attribute; // remove splitting attribute
	for each outcome j of splitting criterion
	
	// partition the tuples and grow subtrees for each partition
	let Dj be the set of data tuples in D satisfying outcome j; // a partition
	
	if Dj is empty then
	attach a leaf labeled with the majority
	class in D to node N;
	else
	attach the node returned by Generate
	decision tree(Dj, attribute list) to node N;
	end for
	return N;					

15	Is clustering unsupervised or supervised classification? Give the reason for your answer**.** Clustering is an example of an **unsupervised learning algorithm (1 Mark)** **Justification (4 Marks)** Unlike supervised methods, clustering is an **unsupervised method** that works on datasets in which there is no outcome (target) variable nor is anything known about the relationship between the observations, that is, unlabeled data You might also hear this referred to as cluster analysis because of the way this method works. Using a clustering algorithm means you're going to give the algorithm a lot of input data with no labels and let it find any groupings in the data it can.	**5**	**2**	**3**	**2**	**2.6.4**
	Part – B **(2 x 10 = 20 Marks)**					
16	Compare FP growth and Apriori algorithm with suitable example? **Comparing Apriori and FP-Growth Algorithm (6 Marks) + Example (4 Marks)**	**10**	**3**	**2**	**1**	**1.7.1**

One of the most important features of any frequent itemset mining algorithm is that it should take lower timing and memory. Taking this into consideration, we have a lot of algorithms related to FIM algorithms. These two Apriori and FP-Growth algorithms are the most basic FIM algorithms. Other algorithms in this field are improvements of these algorithms. There are some basic differences between these algorithms let’s take a look at

Apriori	**FP Growth**
Apriori generates the frequent patterns by making the itemsets using pairing such as single item set, double itemset, triple itemset.	FP Growth generates an FP-Tree for making frequent patterns.
Apriori uses candidate generation where frequent subsets are extended one item at a time.	FP-growth generates conditional FP-Tree for every item in the data.
Since apriori scans the database in each of its steps it becomes time-consuming for data where the number of items is larger.	FP-tree requires only one scan of the database in its beginning steps so it consumes less time.
A converted version of the database is saved in the memory	Set of conditional FP-tree for every item is saved in the memory
It uses breadth-first search	It uses a depth-first search.





Scan D for count of each candidate → C_1

Itemset	Sup. count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

Compare candidate support count with minimum support count → L_1

Itemset	Sup. count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

Generate C_2 candidates from L_1 → C_2

Itemset
{I1, I2}
{I1, I3}
{I1, I4}
{I1, I5}
{I2, I3}
{I2, I4}
{I2, I5}
{I3, I4}
{I3, I5}
{I4, I5}

Scan D for count of each candidate → C_2

Itemset	Sup. count
{I1, I2}	4
{I1, I3}	4
{I1, I4}	1
{I1, I5}	2
{I2, I3}	4
{I2, I4}	2
{I2, I5}	2
{I3, I4}	0
{I3, I5}	1
{I4, I5}	0

Compare candidate support count with minimum support count → L_2

Itemset	Sup. count
{I1, I2}	4
{I1, I3}	4
{I1, I5}	2
{I2, I3}	4
{I2, I4}	2
{I2, I5}	2

Generate C_3 candidates from L_2 → C_3

Itemset
{I1, I2, I3}
{I1, I2, I5}

Scan D for count of each candidate → C_3

Itemset	Sup. count
{I1, I2, I3}	2
{I1, I2, I5}	2

Compare candidate support count with minimum support count → L_3

Itemset	Sup. count
{I1, I2, I3}	2
{I1, I2, I5}	2

[OR]

17	A database has five transactions. Let min support = 60% and minconfidence = 80%.	**10**	**3**	**2**	**8**	**8.4.1**

TID	Items bought
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I, E}

(a) Find all frequent item sets using Apriori and FP-growth, respectively. Compare the efficiency of the two mining processes. **(6 Marks)**

(b) List all the strong association rules (with support s and confidence) **(2Marks)**

c) Matching the following metarule, where X is a variable representing customers, and item denotes variables representing items (e.g., “A,” “B,”):

∀x ∈ transaction, buys(X, item1) ∧ buys(X,item2) ⇒ buys(X,item3) [s,c] **(2 Marks)**

List out items and their support counts

Item | Support Count

A | 1

C | 2

D |1

E |4

I |1

K |5

M |3

N |2

O |3

U |1

Y |3

Generate L, but with the largest support count to smallest

L = {{K: 5}, {E: 4}, {M: 3}, {O:3}, {Y: 3}}

Item |Support Count

K |5

E |4

M |3

O |3

Y |3

So, rewriting the table with the new column:

TID	items_bought	Ordered Itemset
T100	{M, O, N, K, E, Y}	K,E,M,O,Y
T200	{D, O, N, K, E, Y}	K,E,O,Y
T300	{M, A, K, E}	K,E,M
T400	{M, U, C, K, Y}	K,M,Y
T500	{C, O, O, K, I, E}	K,E,O

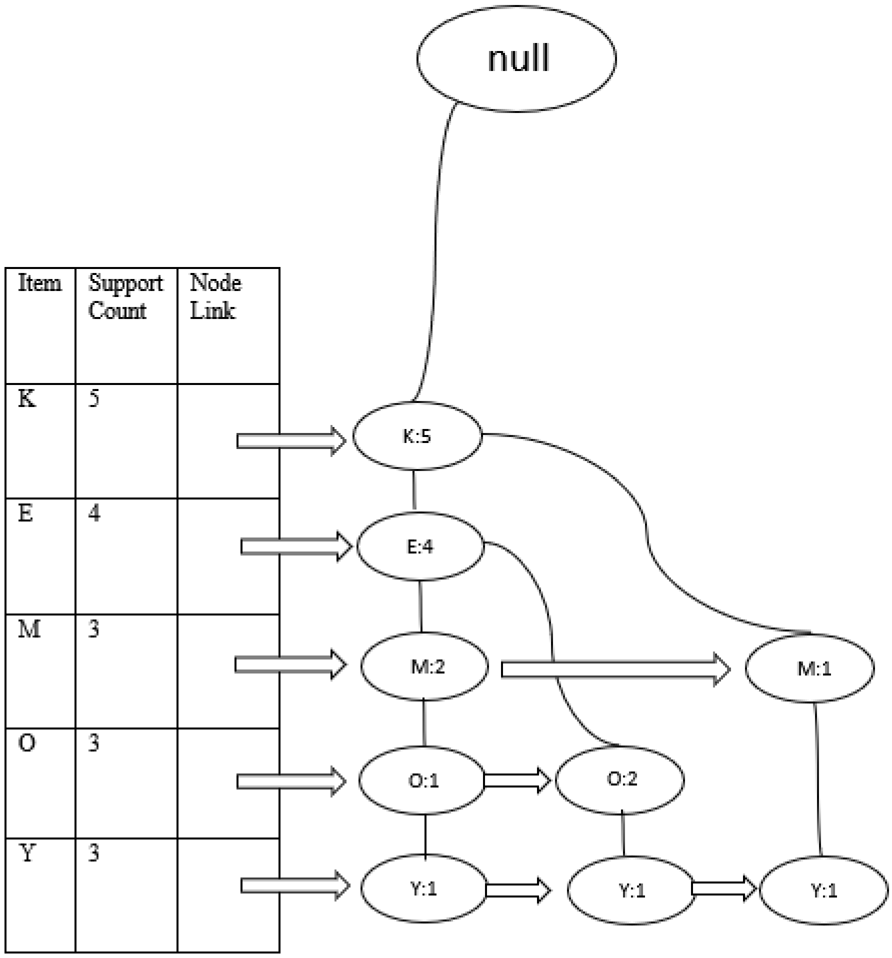
Creating the tree

Steps

1. Write out table structure

2. Add empty node (null)

3. Go through each transaction's ordered item-set and draw nodes and link them together. If the paths match, then add one to the count on the note itself.



Find Conditional Pattern Base by analyzing all paths to an item.

Find Conditional FP Tree by finding common terms among record in Conditional Patter Base and summing its support count.

Apriori

ITERATION 1:

STEP 1: (C1)

Item	Count
A	1
C	2
D	1
E	4
I	1
K	5

M	3
N	2
0	3
U	1
Y	3
STEP 2: (L2)	
Item	Count
E	4
K	5
M	3
O	3
Y	3
ITERATION 2:	
STEP 3: (C2)	
Item	Count
E,K	4
E,M	2
E,O	3
E,Y	2

ASSOCIATION RULE:

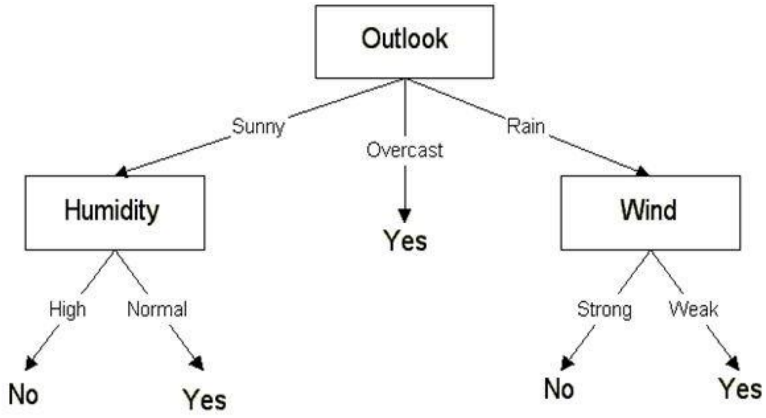
1. [E,K] →→ 0 = 3/4 = 75%
2. [K,O] →→ E = 3/3 = 100%
3. [E,O] →→ K = 3/3 = 100%
4. E →→ [K,O] = 3/4 = 75%
5. K →→ [E,O] = 3/5 = 60%
6. O →→ [E,K] = 3/3 = 100%

18 Construct at least five decision tree from the dataset. Write 5 different rules derived from the constructed tree. **10** **3** **3** **8** **8.4.1**

Outlook	Temperature	Humidity	Wind	Played football (Yes / No)
sunny	Hot	High	Weak	No
sunny	Hot	High	Strong	No
overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No

overcast	Cool	Normal	Strong	Yes
sunny	Mild	High	Weak	No
sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
sunny	Mild	Normal	Strong	Yes
overcast	Mild	High	Strong	Yes
overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

1st tree: (Each tree 2 Marks)



[OR]

19

You are a data scientist which data mining task do you prefer under the following conditions.

A) You are given with a dataset with 3 attributes. 1. Keyword, 2.Length of the document and 3. Spam or not. The attribute “keyword” has the values “accepted” and “Not accepted”. Length of the document has the values “Less than 30” and “More than 30”.

B) A data table with 2 attributes Transaction Id and Items purchased.

i) Justify the mining task chosen.

ii) The algorithm you prefer to do the task.

iii) The information which can be derived.

For A: 5 Marks

i) Justify the mining task chosen:

Here Mining task is classification.

It is a data analysis task, i.e. the process of finding a model that describes and distinguishes data classes and concepts. Classification is the problem of identifying to which of a set of categories (subpopulations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known.

ii) The algorithm you prefer to do the task: (any one)

1. Decision Trees
2. Bayesian Classifiers
3. Neural Networks
4. K-Nearest Neighbour
5. Support Vector Machines

10	3	3	8	8.4.1

6. Linear Regression
7. Logistic Regression

iii) The information which can be derived:

What information gathered from the above data should be explained and justification.

For B: 5 Marks

i) Justify the mining task chosen:

Here Mining task is association rule.

Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently a item set occurs in a transaction. A typical example is a Market Based Analysis.

Market Based Analysis is one of the key techniques used by large relations to show associations between items. It allows retailers to identify relationships between the items that people buy together frequently.

ii) The algorithm you prefer to do the task:

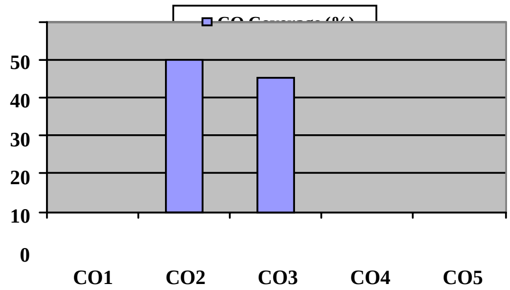
Apriori
Eclat
FP-Growth etc.

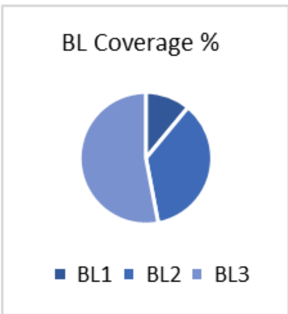
iii) The information which can be derived:

What information gathered from the above data should be explained and justification.

***Performance 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





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: 2023-24 (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 - 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**

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

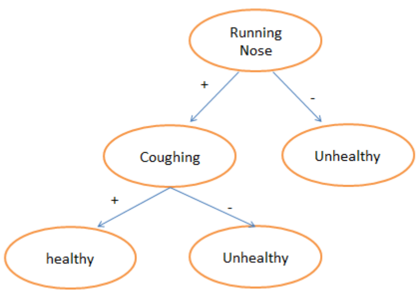
Three ways

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)**

13	Brief about the Metrics for Evaluating classifier Performance	5	L1	3	2	2.8.2

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

14	Explain about decision tree and write all the possible rules from the following Decision tree.	**5**	**L2**	**3**	**2**	2.6.4



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

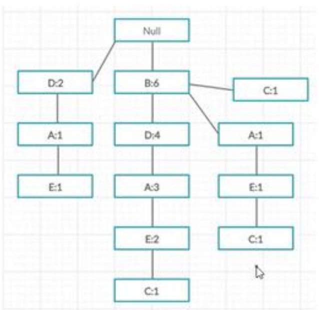
	the result of the algorithm. It is a versatile supervised machine-learning algorithm, which is used for both classification and regression problems Root Node: It is the topmost node in the tree, which represents the complete dataset. It is the starting point of the decision-making process. Decision/Internal 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

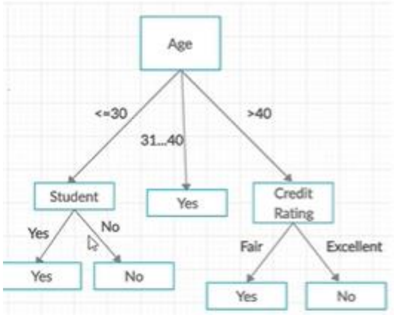
Information gain for second left node

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>**

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

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

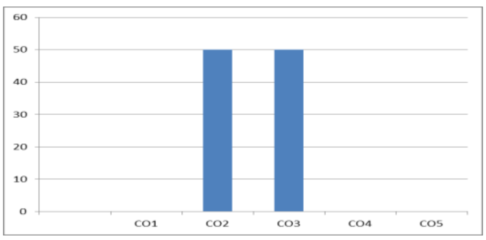
P(No)>P(Yes)

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

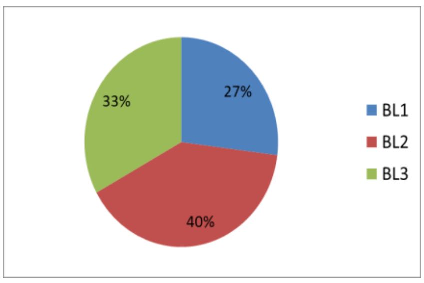
***Program Indicators are available separately for Computer Science and Engineering in AICTE examination reforms policy.**

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

CO Coverage in %



BL Coverage in %



Reg. No.															

SRM Institute of Science and Technology

College of Engineering and Technology

School of Computing

SET - C

DEPARTMENT OF COMPUTING TECHNOLOGIES

SRM Nagar, Kattankulathur – 603203, Chengalpattu District, Tamilnadu

Academic Year: 2023-2024 (ODD)

Test: CLAT-2 **Date:** 07.10.2023

Course Code & Title: 18CSE355T - Data Mining And Analytics **Duration:** 2 Periods

Year & Sem: III & IV Year & 05th & 07th Semester **Max. Marks:** 50 Marks

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

S. No.	Course Outcome	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
1	CO2	3							3				
2	CO3		3						3				

Part – A

(10 x 1 = 10 Marks)

Answer all questions. The duration for answering the part A is 15 minutes (MCQ Answer sheet will be collected after 15 minutes)

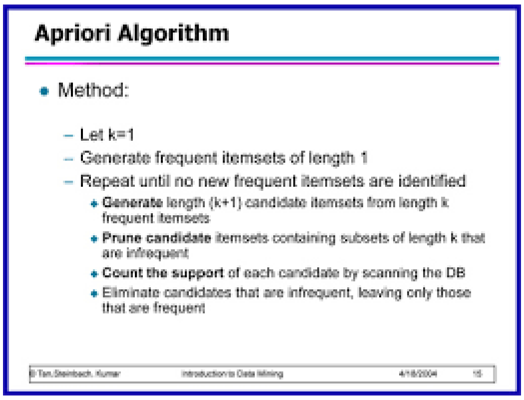
Q. No	Question	Marks	BL	CO	PO	PI Code
1	When do you consider an association rule interesting? (a) If it only satisfies min_support (b) If it only satisfies min_confidence **(c) If it satisfies both min_support and min_confidence** (d) There are other measures to check so	1	1	2	1	1.7.1
2	What is the relation between candidate and frequent itemsets? a) A candidate itemset is always a frequent itemset **b) A frequent itemset must be a candidate itemset.** c) No relation between the two d) Both are same	1	1	2	1	1.7.1
3	Decision tree is a _______ algorithm. a) supervised learning **b) unsupervised learning** c) Both d) None of these	1	3	3	1	1.7.1

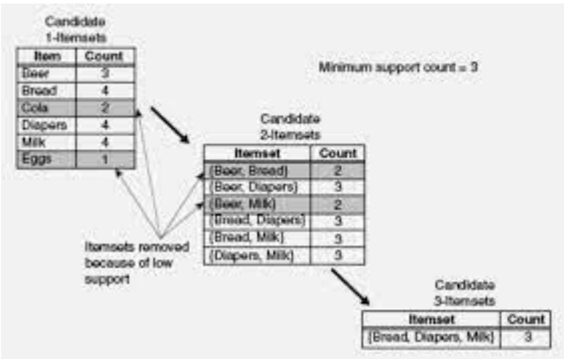
4	Spam Classification is an example for ? a) **Naive Bayes** b) Probabilistic condition c) Random Forest d) All the Above	**1**	**2**	**3**	**1**	**1.7.1**
5	Ensemble methods seek to ____________. a) Reduce variance of individual weak learners by aggregating their predictions. b) Improve performance by exploiting prediction diversity. **c) Both (a) and (b).** d) None of the above.	**1**	**1**	**3**	**1**	**1.7.1**
6	Training in parallel that occurs in bagging aims to capitalize on the ___________ of each base learner, while the sequential training in boosting capitalizes on the _________ of the learners. **a) Independence , dependence.** b) Dependence, Independence. c) Dependence , Dependence. d) Independence, Independence.	**1**	**1**	**3**	**2**	**2.5.2**
7	In Random Forest the Memory requirement for the storage process? **a) High Memory** b) Low Memory c)No Memory d) None of the Above	**1**	**1**	**3**	**2**	**2.5.2**
8	5. Which of the following is an example of a classification problem? a) Predicting the price of a house based on its features b) Predicting the weight of a person based on their height **c) Predicting whether a customer will churn or not** d) Predicting the age of a person based on their income	**1**	**1**	**2**	**2**	**2.5.2**
9	Which one of the following correctly refers to the task of the classification? a) A measure of the accuracy, of the classification of a concept that is given by a certain theory b) **The task of assigning a classification to a set of examples** c) A subdivision of a set of examples into a number of classes d) None of the above	**1**	**1**	**2**	**2**	**2.5.2**

10	Which of the following is the direct application of frequent itemset mining? a) Social Network Analysis **b) Market Basket Analysis** c) Outlier Detection d) Intrusion Detection	**1**	**1**	**2**	**2**	**2.5.2**

Part – B
(4 x 5 = 20 Marks)
Answer any 4 Questions

11	How to generate candidate itemset in Apriori algorithm? Explain in detail	**5**	**2**	**2**	**8**	**8.4.1**

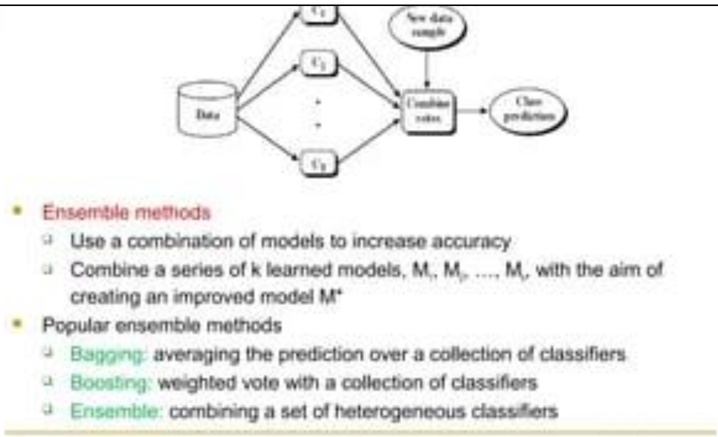


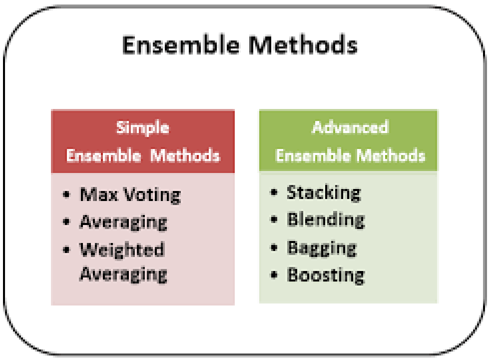


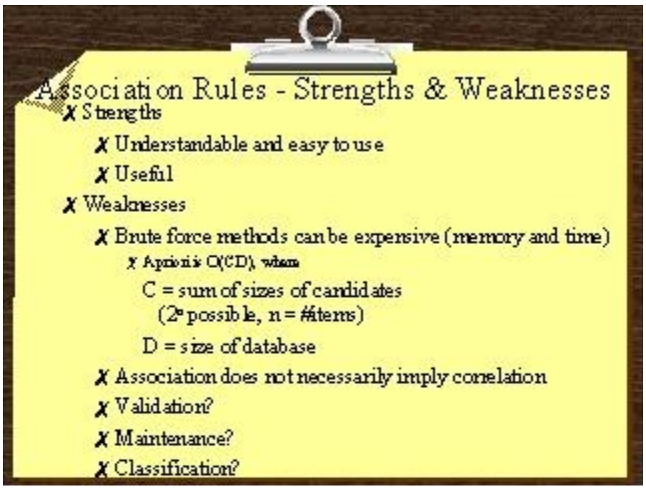
12	The following table contains training dataset of grade database. Use information gain decision tree algorithm to construct a decision tree from the given dataset.	**5**	**2**	**3**	**1**	**1.7.1**

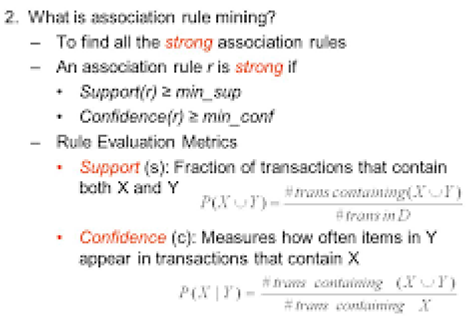
subject	standard	marks	Grade
Science	10	75-100	A
Science	10	55-74	B
Maths	9	90-100	A
Maths	9	55-89	B
Maths	10	80-100	A
Maths	10	55-79	B
Science	9	80-100	A
Science	9	55-79	B

13	How ensemble method improves the classification accuracy? Justify with suitable example.	**5**	**2**	**3**	**1**	**1.7.1**

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




</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>
Give comparison between the strong and the weak association rules




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

15

Describe the following

a. What are attributes and how they influence a learning model. Give Examples

•Attribute selection measures are also known as **splitting**

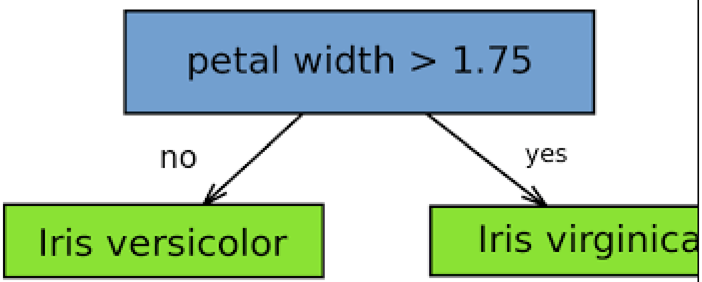
because they determine how the tuples at a given node a

be split.

- Three popular attribute selection **measures**
  - ***information gain***
  - ***gain ratio*, and**
  - ***gini index*.**

b. Define a stump.

A decision stump is a machine learning model consisting of a one-level decision tree. That is, it is a decision tree with one internal node (the root) which is immediately connected to the terminal nodes (its leaves). A decision stump makes a prediction based on the value of just a single input feature.



5	**2**	**2**	**2**	**2.6.4**

Part – B
(2 x 10 = 20 Marks)

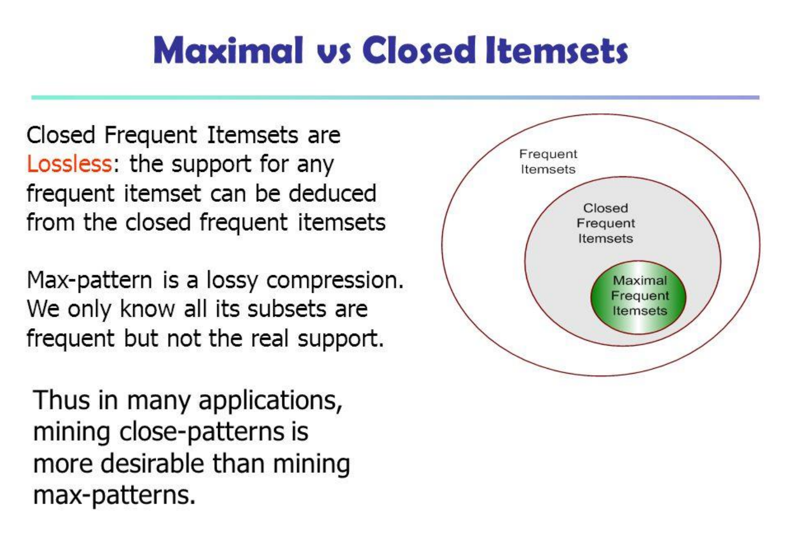
16

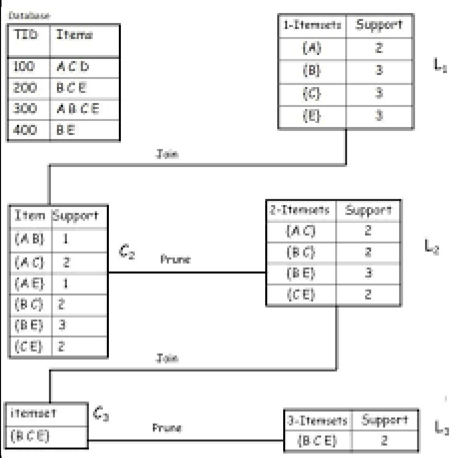
Brand	Fuel type	Mileage	Model	Buy
A1	petrol	low	SUV	Not recommended
A1	petrol	medium	Sports	recommended
A2	diesel	high	SUV	recommended
A1	diesel	medium	SUV	Not recommended
A3	diesel	low	Sports	Not recommended
A2	CNG	medium	SUV	recommended
A4	petrol	low	SUV	recommended
A3	petrol	medium	Sports	Not recommended
A1	diesel	high	SUV	recommended
A2	diesel	medium	SUV	Not recommended
A3	diesel	low	Sports	Not recommended
A4	CNG	medium	SUV	Not recommended
A1	diesel	low	Sports	Not recommended
A1	diesel	medium	Sports	recommended
A2	petrol	medium	Sports	recommended

Consider the following dataset that contains four attributes and one class, use Naïve base classifier to determine the class of Buy for A1 diesel low SUV.

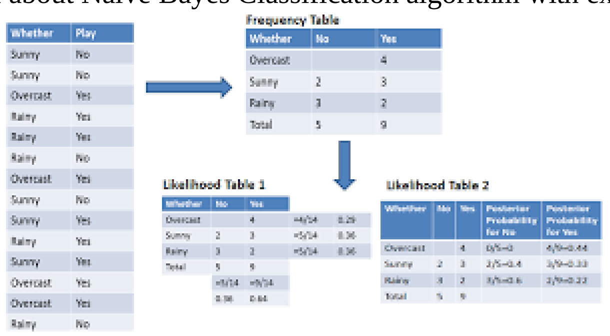
10	L3	2	4	2.7.1

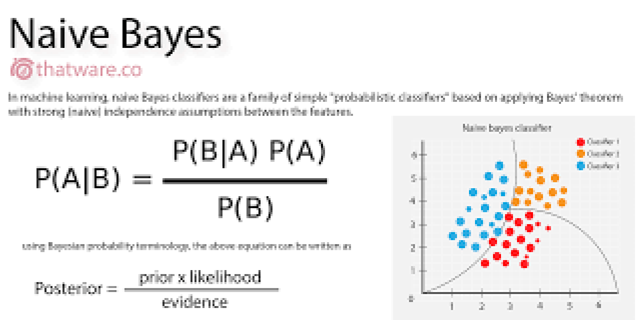
[OR]						
17	What is Closed frequent itemset and maximal frequent item set? Explain in detail about how frequent item sets are identified with any example using Apriori Algorithm.	10	L3	2	4	2.7.1

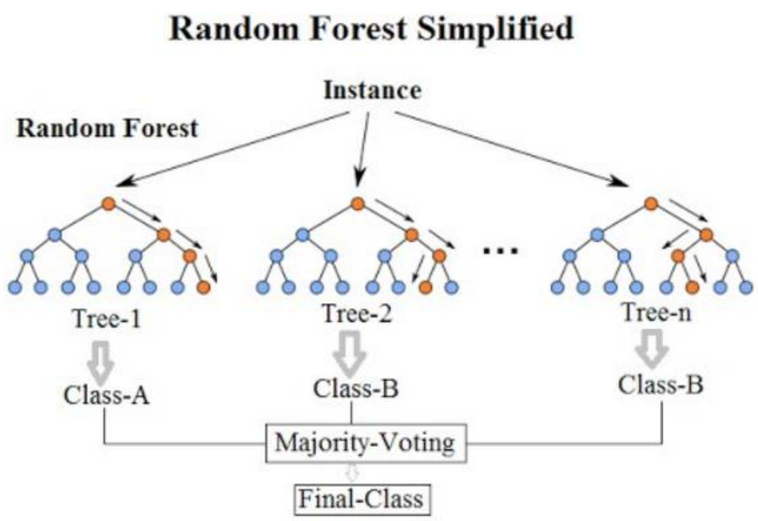
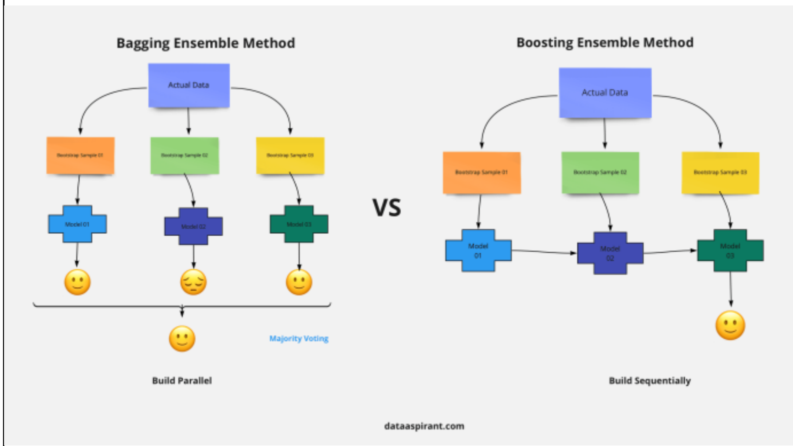




18	Explain about Naive Bayes Classification algorithm with example?	10	L2	3	2	2.5.2

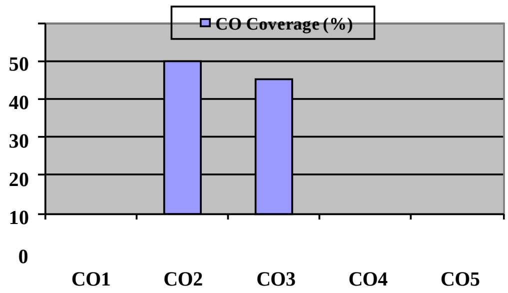


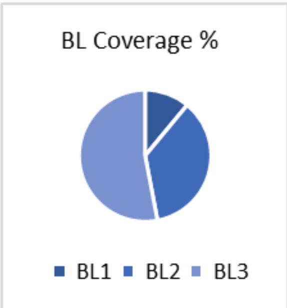


[OR]						
19	Write in detail about Random Forests and write how they fit well with the ensembling, bagging and bootstrapping concept.  	**10**	**L2**	**3**	**4**	**2.6.4**

***Performance 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





Approved by the Audit Professor/Course Coordinator

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:**
Course Code & Title: 18CSC355T Data Mining and Analytics | **Duration:** 1 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

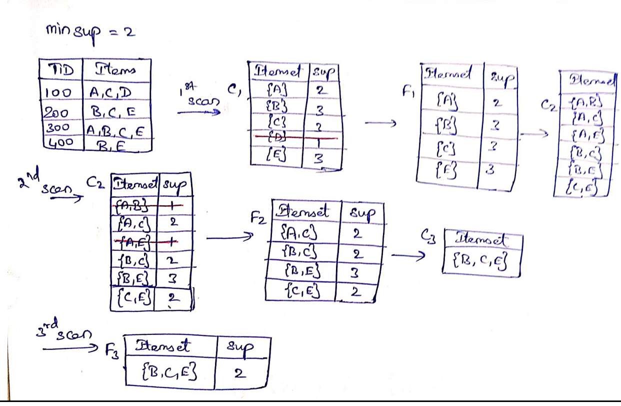
Q. No	Question	Marks	BL	CO	PO	PI Code
1	In some cases, telecommunication companies desire to segment their clients into distinct groups in order to send suitable and related subscription offer. This can be considered as an example of which of the following methods? a. Supervised learning b. **Unsupervised learning** c. Serration d. Data extraction	1	L1	3	1	1.7.1
2	What is the relation between a candidate and frequent itemsets? (a) A candidate itemset is always a frequent itemset **(b) A frequent itemset must be a candidate itemset** (c) No relation between these two (d) Strong relation with transactions	1	L2	2	1	1.7.1
3	Which algorithm requires fewer scans of data? **(a) FP Growth** (b) Naïve Bayes (c) Apriori (d) Decision Tree	1	L2	2	2	1.7.1
4	For the question given below consider the data Transactions : 1. I1, I2, I3, I4, I5, I6 2. I7, I2, I3, I4, I5, I6 3. I1, I8, I4, I5 4. I1, I9, I10, I4, I6 5. I10, I2, I4, I11, I5 With support as 0.6 find all frequent itemsets? **(a)<I1>, <I2>, <I4>, <I5>, <I6>, <I1, I4>, <I2, I4>, <I2, I5>, <I4, I5>, <I4, I6>, <I2, I4, I5>** (b)<I2>, <I4>, <I5>, <I2, I4>, <I2, I5>, <I4, I5>, <I2, I4, I5> (c)<I11>, <I4>, <I5>, <I6>, <I1, I4>, <I5, I4>, <I11, I5>, <I4,	1	L2	2	2	2.6.3

	I6>, <I2, I4, I5> (d)<I1>, <I4>, <I5>, <I6>					
5	**Which of the following is a classification algorithm?** a. K-means **b. Decision tree** c. Apriori d. DBSCAN	**1**	**L1**	**3**	**1**	**1.7.1**
6	Frequency of occurrence of an itemset is called as ______ (a) Support (b) Confidence **(c) Support Count** (d) Rules	**1**	**L1**	**2**	**1**	**1.7.1**
7	In the example of predicting number of babies based on storks' population size, number of babies is **(a) Outcome** (b) Feature (c) Attribute (d) observation	**1**	**L2**	**3**	**5**	**1.7.1**
8	**Which of the following is not a type of decision tree node?** a. Root node b. Leaf node c. Decision node **d. Branch node**	**1**	**L2**	**3**	**1**	**1.7.1**
9	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**
10	The classification or mapping of a class using a predefined class or group is called: a. Data Sub Structure b. Data Set **c. Data Discrimination** d. Data Characterisation	**1**	**L2**	**3**	**1**	**1.7.1**

Part – B
(5x 5 = 25 Marks)

1	**Consider the data set D. Given the minimum support 2, apply Apriori algorithm on this dataset**	**5**	**L3**	**2**	**2**	**2.5.2**

Transaction ID	**Items**
100	**A,C,D**
200	**B,C,E**
300	**A,B,C,E**
400	**B,E**

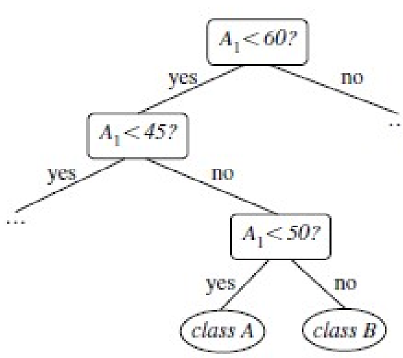
<table>
<tr>
<td></td>
<td>
</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>

How does tree pruning approach works?
There are two common approaches to tree pruning:
-prepruning
-Postpruning
Prepruning approach

- A tree is “pruned” by halting its construction early e.g., by deciding not to further split or partition the subset of training tuples at a given node
- Upon halting, the node becomes a leaf.
- The leaf may hold the most frequent class among the subset tuples or the probability distribution of those tuples.
- When constructing a tree, measures such as statistical significance, information gain, Gini index, and so on, can be used to assess the goodness of a split.
- If partitioning the tuples at a node would result in a split that falls below a prespecified threshold, then further partitioning of the given subset is halted.
- There are difficulties, however, in choosing an appropriate threshold.
- High thresholds could result in oversimplified trees, whereas low thresholds could result in very little simplification.

Postpruning Approach

- The second and more common approach is which removes subtrees from a “fully grown” tree.
- A subtree at a given node is pruned by removing its branches and replacing it with a leaf.
- The leaf is labeled with the most frequent class among the subtree being replaced.



(a)
</td>
<td>5</td>
<td>L3</td>
<td>3</td>
<td>2</td>
<td>1.7.1</td>
</tr>
<tr>
<td>3</td>
<td>

Explain the various metrics for evaluating the classifier performance.
Metrics for Evaluating classifier Performance

- It presents measures for assessing how good or how “accurate” your classifier is at predicting the class label of tuples
</td>
<td>5</td>
<td>L1</td>
<td>2</td>
<td>1</td>
<td>2.6.3</td>
</tr>
</table>

- Consider the case of where the class tuples are more or less evenly distributed, as well as the case where classes are unbalanced
    - e.g., where an important class of interest is rare such as in medical tests
- They include accuracy (also known as recognition rate), sensitivity (or recall), specificity, precision, *F1, and F.*
- *Note that although accuracy is a specific measure, the word "accuracy" is* also used as a general term to refer to a classifier's predictive abilities.
- Using training data to derive a classifier and then estimate the accuracy of the resulting learned model can result in misleading overoptimistic estimates due to overspecialization of the learning algorithm to the data.

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

Meaning of the various measures

- **True positives** (*TP*): These refer to the positive tuples that were correctly labeled by the classifier. Let *TP* be the number of true positives.
- **True negatives** (*TN*): These are the negative tuples that were correctly labeled by the classifier. Let *TN* be the number of true negatives.
- **False positives** (*FP*): These are the negative tuples that were incorrectly labeled as positive (e.g., tuples of class *buys_computer = no* for which the classifier predicted *buys_computer = yes*). Let *FP* be the number of false positives.
- **False negatives** (*FN*): These are the positive tuples that were mislabeled as negative (e.g., tuples of class *buys_computer = yes* for which the classifier predicted *buys_computer = no*). Let *FN* be the number of false negatives.

4	**Suppose the data contain the frequent itemset l = {I1, I2, I5}. What are the association rules that can be generated from l? The nonempty subsets of l are {I1, I2}, {I1, I5}, {I2, I5}, {I1}, {I2} and {I5}. Show the resulting association rules and its confidence.**	**5**	**L2**	**2**	**2**	**2.5.2**

Consider a database, D, consisting of 9 transactions. Suppose min. support count required is 2 (i.e. min_sup = 2/9 = 22 %).
Let the minimum confidence required is 70%.
We have to first find out the frequent itemset using Apriori algorithm.
Then, Association rules will be generated using min. support & min. confidence.

Step 1: Generating 1-Itemset Frequent Pattern

Itemset	Count
{I1}	6
{I2}	7

{I3}	6
{I4}	2
{I5}	2

The above table is L1.
In the first iteration of the algorithm, each item is a member of the set of candidate.
The set of frequent 1-itemsets, L1, consists of the candidate 1-itemsets satisfying minimum support.

Step 2: Generating 2-Itemset Frequent Pattern

To discover the set of frequent 2-itemsets, L2, the algorithm uses L1 Join L1 to generate a candidate set of 2-itemsets, C2. Next, the transactions in D are scanned and the support count for each candidate itemset in C2 is accumulated (as shown in the middle table).
The set of frequent 2-itemsets, L2, is then determined, consisting of those candidate 2-itemsets in C2 having minimum support.
L1 = {I1,I2,I3,I4,I5}.
Since L2 = L1 join L1 then {I1,I2,I3,I4,I5} join {I1,I2,I3,I4,I5}.
It becomes -> C2= [{I1,I2} {I1,I3}, {I1,I4}, {I1,I5}, {I2,I3}, {I2,I4}, {I2,I5}, {I3,I4} {I3,I5}, {I4,I4}].
Now we need to check the frequent itemsets with min support count.
Then we get -> (C2*C2) L2= [{I1,I2} {I1,I3}, {I1,I5}, {I2,I3}, {I2,I4}, {I2,I5}].
Similarly, We do it for L3.

Step 3: Generating 3-Itemset Frequent Pattern

L2= [{I1,I2} {I1,I3}, {I1,I5}, {I2,I3}, {I2,I4}, {I2,I5}].
L3 = L2 JOIN L2 i.e.
C3 = [{I1,I2,I3},{I1,I2,I5}].
Now, the Join step is complete and the Prune step will be used to reduce the size of C3. Prune step helps to avoid heavy computation due to large Ck.
Procedure Step 1: Find Items starting with I2 in B
It gives {I1, I2, I3}, {I1, I2, I4}, {I1, I2, I5}.
Step 2: Find Items starting with I3 in B
It gives NIL, Similarly I4, I5.
Step 3: Find out infrequent items sets using min support count and remove them.
Based on the Apriori property that all subsets of a frequent itemset must also be frequent, we can determine that four latter candidates cannot possibly be frequent. How?
For example, lets take {I1, I2, I3}. The 2-item subsets of it are {I1, I2}, {I1, I3} & {I2, I3}. Since all 2-item subsets of {I1, I2, I3} are members of L2, We will keep {I1, I2, I3} in C3.
Lets take another example of {I2, I3, I5} which shows how the pruning is performed. The 2-item subsets are {I2, I3}, {I2, I5} & {I3,I5}.
BUT, {I3, I5} is not a member of L2 and hence it is not frequent violating Apriori Property. Thus We will have to remove {I2, I3, I5} from C3.
Therefore, C3 = {{I1, I2, I3}, {I1, I2, I5}} after checking for all members of the result of Join operation for Pruning.

Now, the transactions in D are scanned in order to determine L3, consisting of those candidates 3-itemsets in C3 having minimum support.

Step 4: Generating 4-Itemset Frequent Pattern

The algorithm uses L3 Join L3 to generate a candidate set of 4-itemsets, C4. Although the join results in {{I1, I2, I3, I5}}, this itemset is pruned since its subset {{I2, I3, I5}} is not frequent.
Thus, C4 = φ(Null), and algorithm terminates, having found all

	of the frequent items. This completes our Apriori Algorithm. Step 5: Generating Association Rules From Frequent Itemsets Let the minimum confidence threshold is, say 70%. The resulting association rules are shown below, each listed with its confidence. R1: I1 ^ I2 -> I5 Confidence = sc{I1,I2,I5}/sc{I1,I2} = 2/4 = 50%. R1 is Rejected. R2: I1 ^ I5 -> I2 Confidence = sc{I1,I2,I5}/sc{I1,I5} = 2/2 = 100%. R2 is Selected. R3: I2 ^ I5 -> I1 Confidence = sc{I1,I2,I5}/sc{I2,I5} = 2/2 = 100%. R3 is Selected. R4: I1 -> I2 ^ I5 Confidence = sc{I1,I2,I5}/sc{I1} = 2/6 = 33%. R4 is Rejected. R5: I2 -> I1 ^ I5 Confidence = sc{I1,I2,I5}/{I2} = 2/7 = 29% R5 is Rejected. R6: I5 -> I1 ^ I2 Confidence = sc{I1,I2,I5}/ {I5} = 2/2 = 100%. R6 is Selected.					
5	**Explain Naive Baye's Classification.** The **naïve Bayesian** classifier, or **simple Bayesian** classifier, works as follows: **1.** Let D be a training set of tuples and their associated class labels. As usual, each tuple is represented by an n-dimensional attribute vector, $X = (x_1, x_2, \ldots, x_n)$, depicting n measurements made on the tuple from n attributes, respectively, $A_1, A_2, \ldots, A_n$. **2.** Suppose that there are m classes, $C_1, C_2, \ldots, C_m$. Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naïve Bayesian classifier predicts that tuple X belongs to the class C_i if and only if $P(C_i\|X) > P(C_j\|X) \quad \text{for } 1 \le j \le m, j \ne i.$ Thus we maximize $P(C_i\|X)$. The class C_i for which $P(C_i\|X)$ is maximized is called the *maximum posteriori hypothesis*. By Bayes' theorem (Equation (6.10)), $P(C_i\|X) = \frac{P(X\|C_i)P(C_i)}{P(X)}. \quad (6.11)$ **3.** As $P(X)$ is constant for all classes, only $P(X\|C_i)P(C_i)$ need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C_1) = P(C_2) = \cdots = P(C_m)$, and we would therefore maximize $P(X\|C_i)$. Otherwise, we maximize $P(X\|C_i)P(C_i)$. Note that the class prior probabilities may be estimated by $P(C_i) = \|C_{i,D}\|/\|D\|$, where $\|C_{i,D}\|$ is the number of training tuples of class C_i in D. 4. In order to reduce computation in evaluating *P*(*X*j*Ci*), the naïve assumption of class conditional independence is made. $P(X\|C_i) = \prod_{k=1}^{n} P(x_k\|C_i) = P(x_1\|C_i) \times P(x_2\|C_i) \times \cdots \times P(x_n\|C_i).$ (a) If A_k is categorical, then $P(x_k\|C_i)$ is the number of tuples of class C_i in D having the value x_k for A_k, divided by $\|C_{i,D}\|$, the number of tuples of class C_i in D. (b) If A_k is continuous-valued, then we need to do a bit more work, but the calculation is pretty straightforward. A continuous-valued attribute is typically assumed to have a Gaussian distribution with a mean μ and standard deviation σ, defined by $g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad (6.13)$ so that $P(x_k\|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}). \quad (6.14)$	**5**	**L2**	**3**	**2**	**2.6.3**

5. In order to predict the class label of X, $P(X|C_i)P(C_i)$ is evaluated for each class C_i. The classifier predicts that the class label of tuple X is the class C_i if and only if

$$P(X|C_i)P(C_i) > P(X|C_j)P(C_j) \quad \text{for } 1 \le j \le m, j \ne i. \qquad (6.15)$$

In other words, the predicted class label is the class C_i for which $P(X|C_i)P(C_i)$ is the maximum.

Part – C
(4 x 10 = 10 Marks)

1	**A database has five transactions. Let the minimum support & confidence, min_sup=2, min_confi=80%**	10	L3	2	1	2.5.2

TID	ITEMS
T1	{1,2,3,4,5,6}
T2	{7,2,3,4,5,6}
T3	{1,8,4,5}
T4	{1,9,0,4,6}
T5	{0,2,2,4,5}

Find the frequent itemsets and generate the association rules using Apriori algorithm

TID	Items
T_1	{1,2,3,4,5,6}
T_2	{7,2,3,4,5,6}
T_3	{1,8,4,5}
T_4	{1,9,0,4,6}
T_5	{0,2,2,4,5}

support $= 60\% = \frac{60}{100} \times 5 = 3$

confidence $= 80\%$

1. Finding Frequent Itemset

Candidate 1 - Itemset

Item	Count
1	3
2	3
~~3~~	~~2~~
4	5
5	4
6	3
~~7~~	~~1~~
~~8~~	~~1~~
~~9~~	~~1~~
~~0~~	~~2~~

L_1

Item	Count
1	3
2	3
4	5
5	4
6	3

Candidate 2 - Itemset

Itemset	Count
~~1,2~~	~~1~~
1,4	3
~~1,5~~	~~2~~
~~1,6~~	~~2~~
2,4	3
2,5	3
~~2,6~~	~~2~~
4,5	4
4,6	3
~~5,6~~	~~2~~

L_2

Item	Count
1,4	3
2,4	3
2,5	3
4,5	4
4,6	3

candidate 3 - Itemset

Item	Count
2,4,5	3
~~4,5,6~~	~~2~~

Frequent Itemset
{2,4,5}

2. Generate Association Rule using Min_confi - 80%.

The frequent Itemset - {2,4,5}

write possible subset of frequent itemset

{2,4}, {4,5}, {2,5}, {2}, {4}, {5}, ~~{}~~, {2,4,5} → Eliminate the empty subset

R_1 $2 \wedge 4 \rightarrow 5$ (Accepted)

$$\text{conf} = \frac{\text{sup}(A \cup B)}{\text{sup}(A)} = \frac{\text{sup}(2,4,5)}{\text{sup}(2,4)} = \frac{3}{3} = 1\ (100\%)$$

R_2 $4 \wedge 5 \rightarrow 2$ (Eliminated)

$$\frac{\text{sup}(4,5,2)}{\text{sup}(4,5)} = \frac{3}{4} = 0.75\ (75\%)$$

R_3 $2 \wedge 5 \rightarrow 4$ (Accepted)

$$\frac{\text{sup}(2,5,4)}{\text{sup}(2,5)} = \frac{3}{3} = 1\ (100\%)$$

R_4 $2 \rightarrow 4 \wedge 5$ (Accepted)

$$\frac{\text{sup}(2,4,5)}{\text{sup}(2)} = \frac{3}{3} = 1\ (100\%)$$

R_5 $4 \rightarrow 2 \wedge 5$ (Eliminated)

$$\frac{\text{sup}(4,2,5)}{\text{sup}(4)} = \frac{3}{5} = 0.6\ (60\%)$$

R_6 $5 \rightarrow 2 \wedge 4$ (Eliminated)

$$\frac{\text{sup}(5,2,4)}{\text{sup}(5)} = \frac{3}{4} = 0.75\ (75\%)$$

Final Rules (That satisfies both supports & confidence)

1. $2 \wedge 4 \rightarrow 5$
2. $2 \wedge 5 \rightarrow 4$
3. $2 \rightarrow 4 \wedge 5$

2	**Write the algorithm steps to generate decision tree from the training tuples of the data partition.**	10	L3	3	2	2.6.2

<table>
<tr><td></td><td>Algorithm: Generate_decision_tree. Generate a decision tree from the training tuples of data partition D.
Input:
■ Data partition, D, which is a set of training tuples and their associated class labels;
■ attribute_list, the set of candidate attributes;
■ Attribute_selection_method, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a splitting_attribute and, possibly, either a split point or splitting subset.
Output: A decision tree.
Method:
(1) create a node N;
(2) if tuples in D are all of the same class, C then
(3) return N as a leaf node labeled with the class C;
(4) if attribute_list is empty then
(5) return N as a leaf node labeled with the majority class in D; // majority voting
(6) apply Attribute_selection_method(D, attribute_list) to find the "best" splitting_criterion;
(7) label node N with splitting_criterion;
(8) if splitting_attribute is discrete-valued and
multiway splits allowed then // not restricted to binary trees
(9) attribute_list ← attribute_list − splitting_attribute; // remove splitting_attribute
(10) for each outcome j of splitting_criterion
// partition the tuples and grow subtrees for each partition
(11) let D_j be the set of data tuples in D satisfying outcome j; // a partition
(12) if D_j is empty then
(13) attach a leaf labeled with the majority class in D to node N;
(14) else attach the node returned by Generate_decision_tree(D_j, attribute_list) to node N;
endfor
(15) return N;

Basic algorithm for inducing a decision tree from training tuples.</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>3</td><td>a) What are the advantages of FP-Growth algorithm?
b) Discuss the applications of association analysis.
Ans a) Advantages of FP Growth Algorithm
o This algorithm needs to scan the database twice when compared to Apriori, which scans the transactions for each iteration.
o The pairing of items is not done in this algorithm, making it faster.
o The database is stored in a compact version in memory.
o It is efficient and scalable for mining both long and short frequent patterns.
o No candidate generation, no candidate test
o Uses compact data structure called FP-Tree
o Eliminates repeated database scan
o Basic operation is counting and FP-tree building
b) Applications of association analysis
The association rule learning is the important technique of machine learning, and it is employed in Market Basket analysis, Web usage mining, continuous production, etc. In market basket analysis, it is an adequate used by several big retailers to find the relations among items.
Association rules were originally transformed from point-of-sale data that represent what products are purchased together. Although its roots are in linking point-of-sale transactions, association rules can be used external the retail market to find relationships among types of "baskets."
There are various applications of Association Rule which are as follows −
• Items purchased on a credit card, such as rental cars and hotel rooms, support insight into the following product that customer are likely to buy.
• Optional services purchased by tele-connection users (call waiting, call forwarding, DSL, speed call, etc.) support decide how to bundle these functions to maximize revenue.
• Banking services used by retail users (money industry accounts, CDs, investment services, car loans, etc.)</td><td>10</td><td>L1</td><td>2</td><td>1</td><td>1.7.1</td></tr>
</table>

	recognize users likely to needed other services. • Unusual group of insurance claims can be an expression of fraud and can spark higher investigation. • Medical patient histories can supports expressions of likely complications based on definite set of treatments.					
4	**Illustrate any two of the attribute selection measure with an example.** **Attribute Selection Measures** • An attribute selection measure is a heuristic for selecting the splitting criterion that "best" separates a given data partition, *D, of class-labeled training tuples into individual* classes. • If we were to split *D into smaller partitions according to the outcomes of the* splitting criterion, ideally each partition would be pure (i.e., all the tuples that fall into a given partition would belong to the same class). • Attribute selection measures are also known as splitting rules because they determine how the tuples at a given node are to be split. • Three popular attribute selection measures - *information gain* - *gain ratio*, and - *gini index*. **Information gain** • ID3 uses information gain as its attribute selection measure. • The attribute with the highest information gain is chosen as the splitting attribute for node *N*. • Where pi is the probability that an arbitrary tuple in D belongs to class Ci and is estimated by |Ci,D|/|D|. • A log function to the base 2 is used, because the information is encoded in bits. • Info(D) is just the average amount of information needed to identify the class label of a tuple in D. • *Info(D)* is also known as the entropy of *D*. • The expected information needed to classify a tuple in *D is given by* $Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i),$ • How much more information would we still need (after the partitioning) in order to arrive at an exact classification? This amount is measured by $Info_A(D) = \sum_{j=1}^{v} \frac{\lvert D_j \rvert}{\lvert D \rvert} \times Info(D_j).$ • The term |*Dj* |/|*D*| acts as the weight of the *j*th partition. • *InfoA*(*D*) is the expected information required to classify a tuple from *D* based on the partitioning by *A*. • The smaller the expected information (still) required, the greater the purity of the partitions. • Information gain is defined as the difference between the original information requirement (i.e., based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on *A*). $Gain(A) = Info(D) - Info_A(D).$ • The class label attribute, *buys computer*, has two distinct values (namely, {*yes, no*}). • There are two distinct classes (that is, *m* = 2). • Let class *C*1 correspond to *yes* and class *C*2 correspond to *no.* • There are nine tuples of class *yes* and five tuples of class *no*. A (root) node *N* is created for the tuples in *D*.	**10**	**L2**	**3**	**2**	**2.5.2**

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

$$Info(D) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940 \text{ bits.}$$

- Compute the expected information requirement for each attribute.
- *Age* category *youth – 2 yes &3 no, Middle aged – 4 yes & 0 no, Senior – 3 yes & 2 no.*

$$\begin{aligned} Info_{age}(D) = {} & \frac{5}{14} \times \left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) \\ & + \frac{4}{14} \times \left(-\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}\right) \\ & + \frac{5}{14} \times \left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right) \\ = {} & 0.694 \text{ bits.} \end{aligned}$$

$$Gain(age) = Info(D) - Info_{age}(D) = 0.940 - 0.694 = 0.246 \text{ bits.}$$

- Compute *Gain*(*income*) = 0.029 bits, *Gain*(*student*) = 0.151 bits, and *Gain*(*credit rating*) = 0.048 bits.
- Age has the highest information gain among the attributes, it is selected as the splitting attribute.

Gain ratio

- It prefers to select attributes having a large number of values.
- C4.5, a successor of ID3, uses an extension to information gain known as *gain ratio.*
- It applies a kind of normalization to information gain using a "split information" value defined analogously with *Info*(*D*) as

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right).$$

- It differs from information gain, which measures the information with respect to classification that is acquired based on the same partitioning.
- The gain ratio is defined as

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}.$$

$$SplitInfo_A(D) = -\frac{4}{14} \times \log_2\left(\frac{4}{14}\right) - \frac{6}{14} \times \log_2\left(\frac{6}{14}\right) - \frac{4}{14} \times \log_2\left(\frac{4}{14}\right).$$

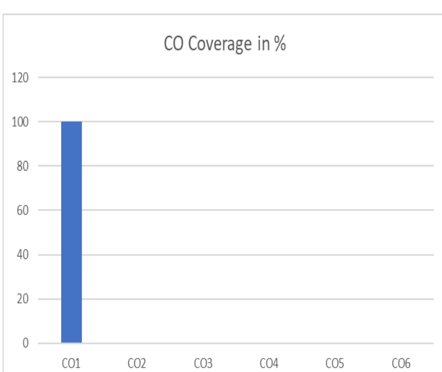
$$= 0.926.$$

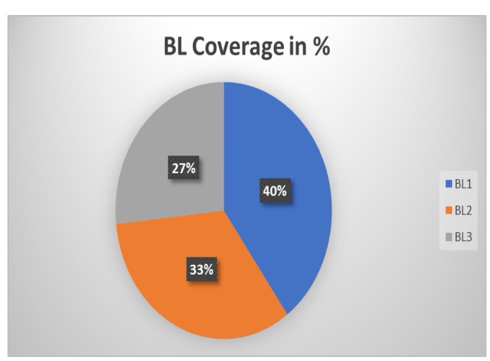
$$Gain(income) = 0.029.$$
$$GainRatio(income) = 0.029/0.926 = 0.031.$$

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

***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





Approved by the Audit Professor/Course Coordinator