Reg. No.															

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

Batch - 1

DEPARTMENT OF COMPUTING TECHNOLOGIES
SRM Nagar, Kattankulathur – 603203, Chengalpattu District, Tamilnadu

Academic Year: 2022-2023(ODD)

Test: CLAT-2 **Date:** 14.10.2022
Course Code & Title: 18CSE355T - Data Mining and Analytics **Duration:** 2 Periods
Year & Sem: III Year & 05th 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	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 20 minutes (MCQ Answer sheet will be collected after 20 minutes)

Q. No	Question	Marks	BL	CO	PO	PI Code
1	What is the relation between a candidate and frequent item sets? a) A candidate item set is always a frequent item set **b) A frequent item set must be a candidate item set** c) No relation between these two d) Strong relation with transactions	1	1	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	1	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	2	2	1	1.7.1
4	If {A, B, C, D} is a frequent item set, candidate rule	1	2	2	1	1.7.1

	which is no possible a) C –> A **b) D –> ABCD** C) A –> BC d) B–> ADC					
5	Which of the following is the direct application of frequent item set mining? a)Social Network Analysis **b)Market Basket Analysis** c)Outlier Detection d) Intrusion Detection	**1**	**1**	**2**	**1**	**1.7.1**
6	Which of the following statement is true about the classification? a) Market basket analysis b) similar group generation **c) find unknown class with trained model** d) identify the data object which does not comply with the general behavior	**1**	**1**	**3**	**2**	**2.7.1**
7	_____predicts categorical labels a) Prediction b) Back Propagation **c) Classification** d) Data trends	**1**	**1**	**3**	**2**	**2.7.1**
8	The class label of training data is unknown in _____. a)Supervised Learning b**) Unsupervised Learning** c) Machine Learning d) NLP	**1**	**1**	**3**	**2**	**2.7.1**
9	_____involves scaling all values for a given attribute so that they fall within a small specified range **a) Normalization** b) Regression c) Prediction d) Classification	**1**	**1**	**3**	**2**	**2.7.1**
10	_____refers to the ability to construct the classifier or predictor efficiently given large amounts of data a) Robustness b) **Scalability** c) Speed d) Interpretability	**1**	**1**	**3**	**2**	**2.7.1**

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

11

Explain Rule based classification with suitable example.
Explanation 2 Marks + Rule 2 Mark + Example 1 Mark.

Rule-based classifier makes use of a set of IF-THEN rules for classification.
We can express a rule in the following from −
IF condition THEN conclusion

Rule Extraction

Here we will learn how to build a rule-based classifier by extracting IF-THEN rules from a decision tree.

To extract a rule from a decision tree −

- One rule is created for each path from the root to the leaf node.
- To form a rule antecedent, each splitting criterion is logically AND.
- The leaf node holds the class prediction, forming the rule consequent.

Let us consider a rule R1,
R1: IF age = youth AND student = yes

THEN buy_computer = yes

Rule-Based Classifier classify records by using a collection of "if…then…" rules.
(Condition) → Class Label
e.g: (BloodType = Warm) ∧ (LayEggs = Yes) → Birds
(TaxableIncome < 50K) ∨ (Refund = Yes) → Evade = No

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

12

Form the Association rules for the frequent item set for given transaction with confidence 35%.

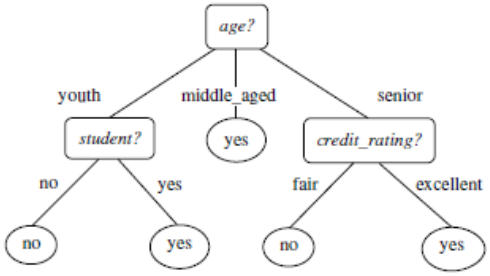
Items
I1,I2,I3
I1,I2,I4
I1,I3,I4
I2,I3,I4

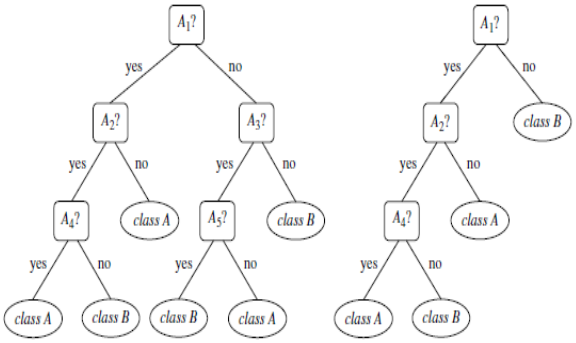
Generate Association Rules: From the frequent item set discovered above the association could be:
{I1, I2} => {I3}

Confidence = support {I1, I2, I3} / support {I1, I2} = (3/ 4)* 100 = 75%
{I1, I3} => {I2}
Confidence = support {I1, I2, I3} / support {I1, I3} = (3/ 3)* 100 = 100%
{I2, I3} => {I1}

5	**2**	**2**	**8**	**8.4.2**

	Confidence = support {I1, I2, I3} / support {I2, I3} = (3/ 4)* 100 = 75% {I1} => {I2, I3} Confidence = support {I1, I2, I3} / support {I1} = (3/ 4)* 100 = 75% {I2} => {I1, I3} Confidence = support {I1, I2, I3} / support {I2 = (3/ 5)* 100 = 60% {I3} => {I1, I2} Confidence = support {I1, I2, I3} / support {I3} = (3/ 4)* 100 = 75% This shows that all the above association rules are strong if minimum confidence threshold is 60%.					
13	**Briefly discuss about Maximal and closed frequent item set. Give difference between apriori and FP growth algorithm.** **Maximal and closed frequent item set (2 Marks)** A closed pattern is a frequent pattern. So it meets the minimum support criteria. In addition to that, all super-patterns of a closed pattern are less frequent than the closed pattern. A Max Pattern is a frequent pattern. So it also meets the minimum support criteria like closed pattern In addition, but unlike closed pattern, all super-patterns of a max pattern are NOT frequent patterns. ***Let's see some examples*** Consider the database containing the transactions T1 : {a1, …, a3}, T2 : {a2, …, a4}. Let minsup = 1. What fraction of all frequent patterns is max frequent patterns? 1-itemsets: {a1}:1, {a2}:2, {a3}:2, {a4}:1 2-itemsets: {a1, a2}:1, {a2, a3}:2, {a1, a3}:1, {a2, a4}:1, {a3, a4}:1 3-itemsets: {a1, a2, a3}:1, {a2, a3, a4}:1 **max frequent patterns:** {a1, a2, a3}, {a2, a3, a4} **closed frequent patterns:** {a1, a2, a3}, {a2, a3, a4}, {a2, a3} Thus, {all frequent patterns} >= {closed frequent patterns} >= {max frequent patterns} **Difference between Apriori and FP Growth (Any 3 points – 3Marks)** **Apriori** 1. It is an array based algorithm. 2. It uses Join and Prune technique. 3. Apriori uses a breadth-first search 4. Apriori utilizes a level-wise approach where it generates patterns containing 1 item, then 2 items, then 3 items, and so on. 5. Candidate generation is extremely slow. Runtime increases exponentially depending on the number of different items. 6. Candidate generation is very parallelizable. 7. It requires large memory space due to large number of candidate generation. 8. It scans the database multiple times for generating candidate sets.	**5**	**2**	**2**	**1**	**1.7.1**

	FP Growth 1. It is a tree based algorithm. 2. It constructs conditional frequent pattern tree and conditional pattern base from database which satisfy minimum support. 3. FP Growth uses a depth-first search 4. FP Growth utilizes a pattern-growth approach means that, it only considers patterns actually existing in the database. 5. Runtime increases linearly, depending on the number of transactions and items 6. Data are very interdependent, each node needs the root. 7. It requires less memory space due to compact structure and no candidate generation. 8. It scans the database only twice for constructing frequent pattern tree.					
14	**Write any five rules from the following Decision tree** Rule 1: If (age=youth) and (student = no) Then class No Rule 2: If (age=youth) and (student = yes) Then class yes Rule 3: If (age=middle aged) Then class yes Rule 4: If (age=senior) and (credit rating = fair) Then class No Rule 5: If (age=senior) and (credit rating = excellent) Then class yes 	**5**	**2**	**3**	**8**	**8.4.2**
15	**Define tree pruning. Explain the process and types of tree pruning with example.** **Pruning (1 Mark)** When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers. Tree pruning methods address this problem of *overfitting* the data. Such methods typically use statistical measures to remove the least-reliable branches. Pruned trees tend to be smaller and less complex and, thus, easier to comprehend. An unpruned tree and a pruned version of it are shown in Figure.	**5**	**2**	**3**	**2**	**2.7.1**



Pre-printing approach (2 Marks)

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.

Post pruning Approach (2 Marks)

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.

Part – B
(2 x 10 = 20 Marks)

16	**Illustrate the method for generating association rules from transactional database. Justify the method of your choice.** **Generating association rules (4 Marks)** Association rule learning is a type of unsupervised learning technique that checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable. It tries to find some interesting relations or associations among the variables of dataset. It is based on different rules to discover the interesting relations between variables in the database. The association rule learning is one of the very important concepts of machine learning, and it is employed in Market Basket analysis, Web usage mining, continuous production, etc. Here market basket analysis is a technique used by the various big retailer to discover the associations between items. We can understand it by taking an example of a supermarket, as in a supermarket, all products that are purchased together are put together.	**10**	**3**	**2**	**1**	**2.7.1**

For example, if a customer buys bread, he most likely can also buy butter, eggs, or milk, so these products are stored within a shelf or mostly nearby.

Association rule learning works on the concept of If and Else Statement, such as if A then B.

Here the If element is called **antecedent**, and then statement is called as **Consequent**. These types of relationships where we can find out some association or relation between two items is known *as single cardinality*. It is all about creating rules, and if the number of items increases, then cardinality also increases accordingly. So, to measure the associations between thousands of data items, there are several metrics. These metrics are given below:

- **Support**
- **Confidence**
- **Lift**

Justify the method of your choice. (4 Marks, Any one algorithm with justification)

Apriori Algorithm

This algorithm uses frequent datasets to generate association rules. It is designed to work on the databases that contain transactions. This algorithm uses a breadth-first search and Hash Tree to calculate the itemset efficiently.

It is mainly used for market basket analysis and helps to understand the products that can be bought together. It can also be used in the healthcare field to find drug reactions for patients.

Eclat Algorithm

Eclat algorithm stands for Equivalence Class Transformation. This algorithm uses a depth-first search technique to find frequent itemsets in a transaction database. It performs faster execution than Apriori Algorithm.

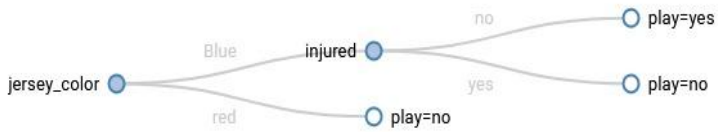
F-P Growth Algorithm

The F-P growth algorithm stands for Frequent Pattern**,** and it is the improved version of the Apriori Algorithm. It represents the database in the form of a tree structure that is known as a frequent pattern or tree.

	The purpose of this frequent tree is to extract the most frequent patterns.					
	[OR]					
17	**Consider the Database for finding frequency pattern using Apriori Algorithm** **TID** | **Items** **T1** | **Hot Dogs, Buns, Ketchup** **T2** | **Hot Dogs, Buns** **T3** | **Hot Dogs, Coke, Chips** **T4** | **Chips, Coke** **T5** | **Chips** **T6** | **Hot Dogs, Coke , Chips** **Find the frequent item sets and generate association rules on this. Assume that minimum support threshold (s = 33.33%) and minimum confident threshold (c = 60%)** Find the frequent item sets and generat**e** association rules on this. Assume that minimum support threshold (s = 33.33%) and minimum confident threshold (c = 60%) Support threshold =33.34% => threshold is at least 2 transactions. Applying Apriori **Frequent item sets** (4 Marks) Pass (k) | Candidate k-itemsets and their support | Frequent k-itemsets k=1 | HotDogs(4), Buns(2), Ketchup(2), Coke(3), Chips(4) | HotDogs, Buns, Ketchup,Coke, Chips k=2 | {HotDogs, Buns}(2), ~~{HotDogs, Ketchup}~~(1), {HotDogs, Coke}(2), {HotDogs, Chips}(2), ~~{Buns, Ketchup}~~(1), ~~{Buns, Coke}~~(0), ~~{Buns, Chips}~~(0), ~~{Ketchup, Coke}~~(0), ~~{Ketchup, Chips}~~(1), {Coke, Chips}(3) | {HotDogs, Buns}, {HotDogs, Coke}, {HotDogs, Chips}, {Coke, Chips} k=3 | {HotDogs, Coke, Chips}(2) | {HotDogs, Coke, Chips} k=4 | {} | Note that {HotDogs, Buns, Coke} and {HotDogs, Buns, Chips} are not candidates when k=3 because their subsets {Buns, Coke} and {Buns, Chips} are not frequent. Note also that normally, there is no need to go to k=4 since the longest transaction has only 3 items. All Frequent Itemsets: {HotDogs}, {Buns}, {Ketchup}, {Coke}, {Chips}, {HotDogs, Buns}, {HotDogs, Coke}, {HotDogs, Chips}, {Coke, Chips}, {HotDogs, Coke, Chips}. All Frequent Itemsets: {HotDogs}, {Buns}, {Ketchup}, {Coke}, {Chips}, {HotDogs, Buns}, {HotDogs, Coke}, {HotDogs, Chips}, {Coke, Chips}, {HotDogs, Coke, Chips}. **Association rules: (4 Marks)S**	**10**	**3**	**2**	**8**	**8.4.2**

	{HotDogs, Buns} would generate: HotDogs □ Buns (2/6=0.33, 2/4=0.5) and Buns □ HotDogs (2/6=0.33, 2/2=1); {HotDogs, Coke} would generate: HotDogs □ Coke (0.33, 0.5) and Coke □ HotDogs (2/6=0.33, 2/3=0.66);{HotDogs, Chips} would generate: HotDogs □ Chips (0.33, 0.5) and Chips □ HotDogs (2/6=0.33, 2/4=0.5); {Coke, Chips} would generate: Coke □ Chips (3/6=0.5, 3/3=1) and Chips □ Coke (3/6=0.5, 3/4=0.75); {HotDogs, Coke, Chips} would generate: HotDogs □ Coke ^ Chips (2/6=0.33, 2/4=0.5), Coke □ Chips ^ HotDogs (2/6=0.33, 2/3=0.66), Chips □ Coke ^ HotDogs (2/6=0.33, 2/4=0.5), HotDogs ^ Coke □ Chips(2/6=0.33, 2/2=1), HotDogs ^ Chips □ Coke(2/6=0.33, 2/2=1) and Coke ^ Chips □ HotDogs(2/6=0.33, 2/3=0.66). With the confidence threshold set to 60%, the Strong Association Rules are (sorted by confidence): **(2 Marks)** 1. Coke □ Chips (0.5, 1)5. Chips □ Coke (0.5, 0.75); 2. Buns □ HotDogs (0.33, 1);6. Coke □ HotDogs (0.33, 0.66); 3. HotDogs ^ Coke □ Chips(0.33, 1) 7. Coke □ Chips ^ HotDogs (0.33, 0.66) 4. HotDogs ^ Chips □ Coke(0.33, 1) 8. Coke ^ Chips □ HotDogs(0.33, 0.66).					
18	**Construct Decision tree for the basketball players dataset using Information Gain** 	Person	Jersey Color	Offense/Defense	Injured	play
---	---	---	---	---
John	Blue	Offense	No	Yes
Steve	Red	Offense	No	No
Sarah	Blue	Defense	No	Yes
Rachel	Blue	Offense	Yes	No
Richard	Red	Defense	No	No
Alex	Red	Defense	Yes	No | **Calculation = 6 Marks** Info(D) = 0.917 Information Gain for Jersey Color $info_A(D) = 0.458$ $gain_{(Jersey\ Color)} = 0.459$ Information Gain for Offense/Defence $info_A(D) = 0.917$ $gain_{(offense/defence)} = 0$ Information Gain for Injured $info_A(D) = 0.666$ $gain_{(injured)} = 0.251$	**10**	**3**	**3**	**2**	**2.5.2**

Decision Tree (4 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. Cold, 2.Temperature and 3.fever or Not. The attribute "Cold" has the values "Yes" and "No". Attribute temperature has "Less than 99" and "Greater than 99".

B) A data table

Sugar	Bread	Butter	Cake
Yes	Yes	NO	No
Yes	NO	No	Yes
No	Yes	Yes	Yes
Yes	Yes	Yes	No
Yes	No	Yes	Yes

i) Justify the mining task chosen.
ii) The algorithm you prefer to do the task.
iii) The information which can be derived.

10 **3** **3** **8** **8.4.2**

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.

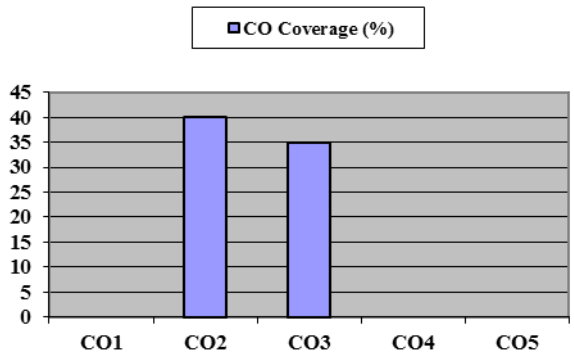
ii) The algorithm you prefer to do the task: (any one)

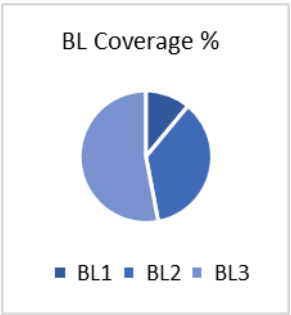
1. Decision Trees
2. Bayesian Classifiers
3. Neural Networks
4. K-Nearest Neighbour

	5. Support Vector Machines 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





Approved by the Audit Professor/Course Coordinator