20IT604/PE02

Hall Ticket Number:

III/IV B.Tech (Regular) DEGREE EXAMINATION

July/August,2023 — **Information Technology**

Sixth Semester — **Data Warehousing and Data Mining**

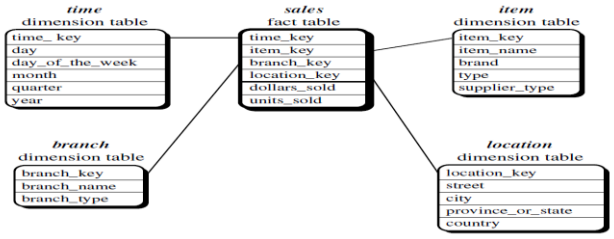
Time: Three Hours — **Maximum:** 70 Marks

Answer question 1 compulsory. **(14X1 = 14Marks)**

Answer one question from each unit. **(4X14=56 Marks)**

			CO	BL	M
1	a)	What is transactional data? Transactional data is a type of data processing that consists of executing a number of transactions occurring concurrently—online banking, shopping, order entry etc	CO1	L1	
	b)	What is meant by concept hierarchy? A concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts.	CO1	L1	
	c)	Define data characterization. Data characterization is a summarization of the general characteristics or features of a target class of data. The data corresponding to the user-specified class are typically collected by a query.	CO1	L1	
	d)	What are outliers? Outlier is a data object that deviates significantly from the rest of the data objects and behaves in a different manner.	CO1	L1	
	e)	What do you mean by tree pruning? Tree pruning attempts to identify and remove such branches, with the goal of improving classification accuracy on unseen data. Pruning reduces the complexity of the final classifier, and hence improves predictive accuracy by the reduction of overfitting.	CO2	L1	
	f)	Define binning. Binning, also known as discretization or bucketing, is a data preprocessing technique used in data mining. It involves dividing a continuous variable into a set of smaller intervals or bins and replacing the original values with the corresponding bin labels.	CO2	L2	
	g)	List any two methods to handle missing data. i)Ignoring the tuple ii) Use a global constant to fill in the missing value iii) Use the attribute mean to fill in the missing value	CO2	L1	
	h)	What is supervised learning. Supervised learning is a machine learning approach that's defined by its use of labeled datasets. These datasets are designed to train or “supervise” algorithms into classifying data or predicting outcomes accurately.	CO2	L1	
	i)	Define Pattern? How are they represented? Pattern mining concentrates on identifying rules that describe specific patterns within the data. These are data points that occur more often in the dataset. Market-basket analysis, which identifies items that typically occur together in purchase transactions.	CO3	L2	
	j)	What is sampling. Sampling is a process in statistical analysis where researchers take a predetermined number of observations from a larger population.	CO3	L1	
	k)	What is partitioning method in clustering? This clustering method classifies the information into multiple groups based on the characteristics and similarity of the data.	CO3	L2	
	l)	Define Dendrogram? A dendrogram is a diagram that shows the hierarchical relationship between objects. The main use of a dendrogram is to work out the best way to allocate objects to clusters.	CO4	L1	
	m)	Define divisive clustering. The divisive clustering algorithm is a top-down clustering approach, initially, all the points in the dataset belong to one cluster and The cluster splitting process repeats until, eventually, each new cluster contains only a single object.	CO4	L1	

	n)	Write the equations for mean and average distance. Mean distance : Average distance :	CO4	L2	
		Unit-I			
2	a)	**Explain the design and construction of Data warehouse.** To design an effective data warehouse we need to understand and analyze business needs and construct a *business analysis framework*. The construction of a large and complex information systemcan be viewed as the construction of a large and complex building, for which the owner, architect, and builder have different views. These views are combined to form a complex framework that represents the top-down, business-driven, or owner's perspective, as well as the bottom-up, builder-driven, or implementor's view of the information system. Four different views regarding the design of a data warehouse must be considered: the *top-down view*, the *data source view*, the *data warehouse view*, and the *business query view*. • The top-down view allows the selection of the relevant information necessary for the data warehouse. This information matches the current and future business needs. • The data source view exposes the information being captured, stored, and managed by operational systems. This information may be documented at various levels of detail and accuracy, from individual data source tables to integrated data source tables. Data sources are often modeled by traditional data modelling techniques, such as the entity-relationship model or CASE (computer-aided software engineering) tools. • The data warehouse view includes fact tables and dimension tables. It represents the information that is stored inside the data warehouse, including precalculated totals and counts, as well as information regarding the source, date, and time of origin, added to provide historical context. • Finally, the business query view is the perspective of data in the data warehouse from the viewpoint of the end user.	CO1	L2	7M
	b)	**On what kinds of data can the data mining can be performed?** **Explanation of any 3 kind of databases can be considered** • A number of different data repositories on which mining can be performed. In principle, data mining should be applicable to any kind of data repository, as well as to transient data, such as data streams. Thus the scope of our examination of data repositories will include relational databases, data warehouses, transactional databases, advanced database systems, flat files, data streams, and the World Wide Web. Advanced database systems include object-relational databases and specific application-oriented databases, such as spatial databases, time-series databases, text databases, and multimedia databases. The challenges and techniques of mining may differ for each of the repository systems. .	CO1	L3	7M
		(OR)			
3	a)	**Explain Star Schema in multidimensional database with a neat diagram**. Explanation- 5M Diagram -2M Star Schema: A star schema is a type of data modeling technique used in data warehousing to represent data in a structured and intuitive way. In a star schema, data is organized into a central fact table that contains the measures of interest, surrounded by dimension tables that describe the attributes of the measures. The fact table in a star schema contains the measures or metrics that are of interest to the user or organization. For example, in a sales data warehouse, the fact table might contain sales revenue, units sold, and profit margins. Each record in the fact table represents a specific event or transaction, such as a sale or order. The dimension tables in a star schema contain the descriptive attributes of the measures in the fact table. These attributes are used to slice and dice the data in the fact table, allowing users to analyze the data from different perspectives. For example, in a sales data warehouse, the dimension tables might include product, customer, time, and location.	CO1	L2	7M



b) **How to identify the interestingness of a pattern.** — CO1 — L3 — 7M

A pattern is **interesting** if it is
(1) *easily understood* by humans,
(2) *valid* on new or test data with some degree of *certainty*,
(3) potentially *useful*, and
(4) *novel*.
A pattern is also interesting if it validates a hypothesis that the user *sought to confirm*. An interesting pattern represents **knowledge**.
Several **objective measures of pattern interestingness** exist. These are based on the structure of discovered patterns and the statistics underlying them. An objective measure for association rules of the form $X \Rightarrow Y$ is rule **support**, representing the percentage of transactions from a transaction database that the given rule satisfies. This is taken to be the probability $P(X \cup Y)$, where $X \cup Y$ indicates that a transaction contains both *X* and *Y*, that is, the union of itemsets *X* and *Y*. Another objective measure for association rules is **confidence**, which assesses the degree of certainty of the detected association. This is taken to be the conditional probability $P(Y|X)$), that is, the probability that a transaction containing *X* also contains *Y*. More formally, support and confidence are defined as

$$\text{Support}(X \Rightarrow Y) = P(X \cup Y)$$
$$\text{Confidence}(X \Rightarrow Y) = P(Y|X)$$

In general, each interestingness measure is associated with a threshold, which may be controlled by the user. For example, rules that do not satisfy a confidence threshold of, say, 50% can be considered uninteresting.

Unit-II

4 a) Consider the dataset below and draw a decision tree using information gain . — CO2 — L3 — 7M

Index	A	B	C	D	E (Class)
1	4.8	3.4	1.9	0.2	positive
2	5	3	1.6	1.2	positive
3	5	3.4	1.6	0.2	positive
4	5.2	3.5	1.5	0.2	positive
5	5.2	3.4	1.4	0.2	positive
6	4.7	3.2	1.6	0.2	positive
7	4.8	3.1	1.6	0.2	positive
8	5.4	3.4	1.5	0.4	positive
9	7	3.2	4.7	1.4	negative
10	6.4	3.2	4.7	1.5	negative
11	6.9	3.1	4.9	1.5	negative
12	5.5	2.3	4	1.3	negative
13	6.5	2.8	4.6	1.5	negative
14	5.7	2.8	4.5	1.3	negative
15	6.3	3.3	4.7	1.6	negative
16	4.9	2.4	3.3	1	negative

Solution can be considered using information gain or Gini index

	b)	**Use these methods to normalize the following group of data: 200, 300, 400, 600,1000** **(i) min-max normalization by setting min = 0 and max =1** **(ii) Z-score normalization** **(a) Min-Max Normalization Data given :- 200, 300, 400, 600, 1000** Min = 200 Max = 1000 V = the respective value of the attribute V1 = 200 V2 = 300 V3 = 400 V4 = 600 V5 = 1000 New max = 1 New min = 0 V' = {V-minA/maxA-minA(new maxA-new minA)} + new min A Therefore, for 200, min max = (200-200(1-0)/1000-200) +0 = 0 for 300 min max = (300-200(1-0)/1000-200) + 0 = 100/800 = 0.125 for 400 min max = (400-200(1-0)/1000-200) + 0 = 200/800 = 0.25 for 600 min max = 600-200/100-200×(1-0)+0 = 400/800 = 0.5 for 1000 min max = 1000-200/1000-200×(1-0)+0	CO2	L3	7M

= 1

(b) Z-score normalization

200,300,400,600,1000

Standard derivation = $\sqrt{\sum}$(every individual data - mean of data)^2/n

Now,

mean value of following data = (200+300+400+600+1000)/5

= 2500/5

= 500

Therefore,

Standard derivation = $\sqrt{\{(200-500)^2+(300-500)^2+(400-500)^2+(600-500)^2+(1000-500)^2\}/5}$

$= \sqrt{\{(-300)^2+(-200)^2+(-100)^2+(100)^2+(500)^2\}/5}$

$= \sqrt{(90000+40000+10000+10000+250000)/5}$

$= \sqrt{400000/5}$

$= \sqrt{80000}$

= 282.8

Z score = 200-500/282.8

= -1.06

Z score = 300-500/282.8

= -0.7

Z score = 400-500/282.8

= -0.35

Z score = 600-500/282.8

= 0.35

Z score = 1000-500/282.8

= 1.78

(OR)

5	a)	**How tree pruning in decision tree induction is useful? Explain various methods for pruning decision trees**. When decision trees are built, many of the branches may reflect noise or outliers in the training data.Tree pruning methods address this problem of overfittingthe data. Tree pruning attempts to identify and remove such branches, with the goal of improving classification accuracy on unseen data. Decision trees can suffer from repetition and replication, making them overwhelming to interpret. Repetition occurs when an attribute is repeatedly tested along a given branch of the tree. In replication, duplicate subtrees exist within the tree. These situations can impede the accuracy and comprehensibility of a decision tree. Pruned trees • These tend to be smaller and less complex and, thus, easier to comprehend.	CO2	L3	7M

		• They are usually faster and better at correctly classifying independent test data than unpruned trees. • Pruned trees tend to be more compact than their unpruned counterparts There are two common approaches to tree pruning: 1. prepruning : •In the pre-pruning 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). •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 pre specified 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. 2. post pruning. •The second and more common approach is post pruning, 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. •The cost complexity pruning algorithm used in CART is an example of the post pruning approach. •The basic idea is that the simplest solution is preferred. •Unlike cost complexity, pruning does not require an independent set of tuples. •Post pruning leads to a more reliable tree.			
	b)	**Suppose a group of 12 sales price records has been sorted as follows:** **5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215.** **Partition them into three bins by each of the following methods:** **(i) equal-frequency (equal-depth) partitioning** **(ii) equal-width partitioning** (i) Equal-frequency partitioning Partition the data into equal depthbins of depth 4 Bin 1: 5, 10, 11 13 Bin 2: 15, 35, 50, 55 Bin 3: 72, 92, 204, 215 (ii) Equal -width partitioning Partitioning the data into 3 equal width bin will require the width to be (215-5)/3=70. We get Bin 1: 5, 10, 11 13, 15, 35, 50, 55, 72 Bin 2: 92 Bin 3: 204, 215	CO2	L3	7M
		Unit-III			
6	a)	**"Strong Rules Are Not Necessarily Interesting". Justify**. Whether or not a rule is interesting can be assessed either subjectively or objectively. Ultimately, only the user can judge if a given rule is interesting, and this judgment, being subjective, may differ from one user to another. However, objective interestingness measures, based on the statistics ―behind‖ the data, can be used as one step toward the goal of weeding out uninteresting rules from presentation to the user.	CO3	L3	7M

The support and confidence measures are insufficient at filtering out uninteresting association rules. To tackle this weakness, a correlation measure can be used to augment the support-confidence framework for association rules. This leads to correlation rules of the form

A →B [support, confidence. correlation].

That is, a correlation rule is measured not only by its support and confidence but also by the correlation between itemsets A and B. There are many different correlation measures from which to choose. In this section, we study various correlation measures to determine which would be good for mining large data sets.

b) Consider the following transactional data for a commercial shop. CO3 L3 7M

TID	List of Items with Ids
T1	i2, i4
T2	i1, i2, i5
T3	i2, i3
T4	i1, i3
T5	i1, i2, i4
T6	i2, i3
T7	i1, i3
T8	i1, i2, i3
T9	i1, i2, i3, i5

Generate all the frequent itemsets using apriori algorithm. Consider the minimum support count is 2. Clearly show your computational steps

Transactions List

TID	List of Items IDs
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

1-item Sets	Frequency
I1	6
I2	7
I3	5
I4	4
I5	2

Frequent 1-item Sets	Frequency
I1	6
I2	7
I3	5
I4	4

Transactions List

TID	List of Items IDs
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

2-item Sets	Frequency
I1, I2	4
I1, I3	4
I1, I4	1
I1, I5	2
I2, I3	3
I2, I4	2
I2, I5	2
I3, I4	0
I3, I5	1
I4, I5	0

Frequent 2-item Sets	Frequency
I1, I2	4
I1, I3	4
I1, I5	2
I2, I3	3
I2, I4	2
I2, I5	2

		(OR)			
7	a)	**Discuss FP Growth algorithm with an example.** The FP-Growth Algorithm is an alternative way to find frequent item sets without using candidate generations, thus improving performance. For so much, it uses a divide-and-conquer strategy. The core of this method is the usage of a special data structure named frequent-pattern tree (FP-tree), which retains the item set association information. **This algorithm works as follows:** o First, it compresses the input database creating an FP-tree instance to represent frequent items. o After this first step, it divides the compressed database into a set of conditional databases, each associated with one frequent pattern. o Finally, each such database is mined separately. Using this strategy, the FP-Growth reduces the search costs by recursively looking for short patterns and then concatenating them into the long frequent patterns	CO3	L2	7M

Transaction ID	Items
T1	{E,K,M,N,O,Y}
T2	{D,E,K,N,O,Y}
T3	{A,E,K,M}
T4	{C,K,M,U,Y}
T5	{C,E,I,K,O,O}

The frequency of each individual item is computed:-

Transaction ID	Items
T1	{E,K,M,N,O,Y}
T2	{D,E,K,N,O,Y}
T3	{A,E,K,M}
T4	{C,K,M,U,Y}
T5	{C,E,I,K,O,O}

Item	Frequency
A	1
C	2
D	1
E	4
I	1
K	5
M	3
N	2
O	3
U	1
Y	3

Item	Frequency
A	1
C	2
D	1
E	4
I	1
K	5
M	3
N	2
O	3
U	1
Y	3

A **Frequent Pattern set (L)** is built which will contain all the elements whose frequency is greater than or equal to the minimum support.

As minimum support be 3.

These elements are stored in descending order of their respective frequencies.

After insertion of the relevant items, the set L looks like this:- **L = {K : 5, E : 4, M : 3, O : 3, Y : 3}**

Now, for each transaction, the respective **Ordered-Item set** is built.

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

Transaction ID	Items	Ordered-Item Set
T1	{E,K,M,N,O,Y}	{K,E,M,O,Y}
T2	{D,E,K,N,O,Y}	{K,E,O,Y}
T3	{A,E,K,M}	{K,E,M}
T4	{C,K,M,U,Y}	{K,M,Y}
T5	{C,E,I,K,O,O}	{K,E,O}

	b)	**Give a short example to show that items in a strong association rule actually may be negatively correlated.** The support and confidence measures are insufficient at filtering out uninteresting association rules. To tackle this weakness, a correlation measure can be used to augment the support-confidence framework for association rules. This leads to *correlation rules* of the form $A \rightarrow B$ [*support, confidence*. *correlation*]. That is, a correlation rule is measured not only by its support and confidence but also by the correlation between itemsets *A* and *B*. There are many different correlation measures from which to choose. In this section, we study various correlation measures to determine which would be good for mining large data sets. Lift is a simple correlation measure that is given as follows. The occurrence of itemset *A* is independent of the occurrence of itemset *B* if $P(A \cup B) = P(A)P(B)$; otherwise, itemsets *A* and *B* are dependent and correlated as events. This definition can easily be extended to more than two itemsets. The lift between the occurrence of *A* and *B* can be measured by computing $lift(A, B) = P(A \cup B)/P(A)P(B)$ If the resulting value is less than 1, then the occurrence of *A* is *negatively correlated with* the occurrence of *B*. If the resulting value is greater than 1, then *A* and *B* are *positively correlated*, meaning that the occurrence of one implies the occurrence of the other. If the resulting value is equal to 1, then *A* and *B* are *independent* and there is no correlation between them. It assesses the degree to which the occurrence of one "lifts" the occurrence of the other. For example, if *A* corresponds to the sale of computer games and *B* corresponds to the sale of videos, then given the current market conditions, the sale of games is said to increase or "lift" the likelihood of the sale of videos by a factor of the value returned by Equation	CO3	L4	7M
		Unit-IV			
8	a)	**Explain agglomerative and divisive hierarchical clustering algorithm** . Agglomerative hierarchical clustering: This bottom-up strategy starts by placing each object in its own cluster and then merges these atomic clusters into larger and larger clusters, until all of the objects are in a single cluster or until certain termination conditions are satisfied. Most hierarchical clustering methods belong to this category. They differ only in their definition of intercluster similarity. Divisive hierarchical clustering: This top-down strategy does the reverse of agglomerative hierarchical clustering by starting with all objects in one cluster. It subdivides the cluster into smaller and smaller pieces, until each object forms a cluster on its own or until it satisfies certain termination conditions, such as a desired number of clusters is obtained or the diameter of each cluster is within a certain threshold.	CO4	L2	7M
	b)	**Explain about DBSCAN algorithm.** DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a densitybased clustering algorithm. The algorithm grows regions with sufficiently high density into clusters and discovers clusters of arbitrary shape in spatial databases with noise. It defines a cluster as a maximal set of *density-connected* points. The basic ideas of density-based clustering involve a number of new definitions. We intuitively present these definitions, and then follow up with an example. • The neighborhood within a radius e of a given object is called the e-neighborhood of the object. • If the e-neighborhood of an object contains at least a minimum number, *MinPts*, of objects, then the object is called a core object. • Given a set of objects, *D*, we say that an object ***p*** is directly density-reachable from object ***q*** if ***p*** is within the e-neighborhood of ***q***, and ***q*** is a core object. • An object ***p*** is density-reachable fromobject ***q*** with respect to e and*MinPts* in a set of objects, *D*, if there is a chain of objects ***p*****1**, : : : , ***pn***, where ***p*****1** = ***q*** and ***pn*** = ***p*** such that ***pi*****+1** is directly density-reachable from ***pi*** with respect to e and *MinPts*, for 1 _ *i* _ *n*, ***pi*** 2 *D*. • An object ***p*** is density-connected to object ***q*** with respect to e and *MinPts* in a set of objects, *D*, if there is an object **o** 2 *D* such that both ***p*** and ***q*** are density-reachable from ***o*** with respect to e and *MinPts*. • Density reachability is the transitive closure of direct density reachability, and this relationship is asymmetric. Only core objects are mutually density reachable. Density	CO4	L2	7M

		connectivity, however, is a symmetric relation.			
		(OR)			
9	a)	**Differentiate between density based and grid-based methods**. **Density-based methods:** Most partitioning methods cluster objects based on the distance between objects. Such methods can find only spherical-shaped clusters and encounter difficulty at discovering clusters of arbitrary shapes. Other clustering methods have been developed based on the notion of *density*. Their general idea is to continue growing the given cluster as long as the density (number of objects or data points) in the "neighborhood" exceeds some threshold; that is, for each data point within a given cluster, the neighborhood of a given radius has to contain atleast a minimum number of points. Such a method can be used to filter out noise (outliers) and discover clusters of arbitrary shape. DBSCAN and its extension, OPTICS, are typical density-based methods that grow clusters according to a density-based connectivity analysis. DENCLUE is a method that clusters objects based on the analysis of the value distributions of density functions. **Grid-based methods:** Grid-based methods quantize the object space into a finite number of cells that form a grid structure. All of the clustering operations are performed on the grid structure (i.e., on the quantized space). The main advantage of this approach is its fast processing time, which is typically independent of the number of data objects and dependent only on the number of cells in each dimension in the quantized space. STING is a typical example of a grid-based method. WaveCluster applies wavelet transformation for clustering analysis and is both grid-based and density-based.	CO4	L2	7M
	b)	**Both k-means and k-medoids algorithms can perform effective clustering. Illustrate the strength and weakness of k-means in comparison with k-medoids.**	CO4	L3	7M

K-means takes the mean of data points to create new points called centroids.	K-medoids uses points from the data to serve as points called medoids.
Centroids are new points previously not found in the data.	Medoids are existing points from the data.
K-means can only by used for numerical data.	K-medoids can be used for both numerical and categorical data.
K-means focuses on reducing the sum of squared distances, also known as the sum of squared error (SSE).	K-medoids focuses on reducing the dissimilarities between clusters of data from the dataset.
K-means is not sensitive to outliers within the data.	K-medoids is outlier resistant and can reduce the effect of outliers.
K-means is less costly to implement.	K-medoids is more costly to implement.
K-means is faster.	K-medoids is comparatively not as fast.

Scrutiny Prepared by:
P.Ravi Kumar,
Dept of IT,
BEC

HOD,IT