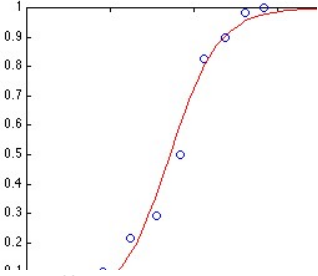


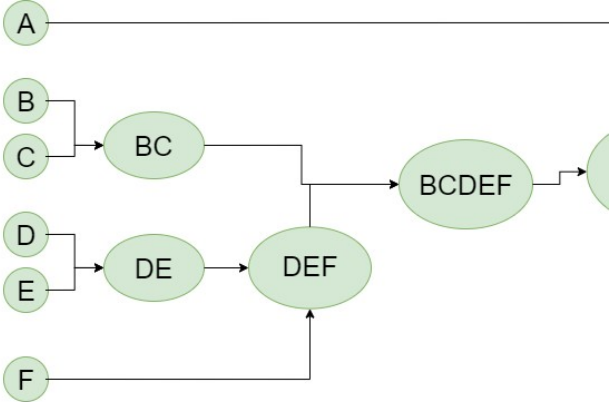
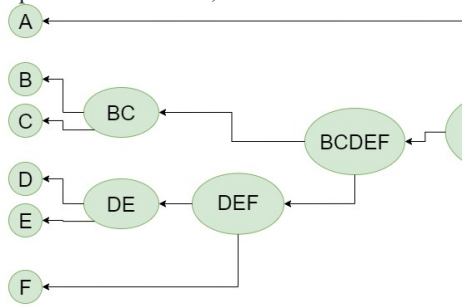
	b) BIRCH c) STING d) DBSCAN					
5	Which is not part of the categories of clustering methods? a) Hierarchical methods b) Density based methods c) Portioning methods d) Rule-based methods	1	1	4	1	1.7.1
6	What are the different ways to classify an Intrusion detection System? a) Zone based b) Host & Network based c) Network & Zone based d) Level based	1	1	5	2	2.7.1
7	Find the outlier in the given data set below. 16, 14, 3, 12, 15, 17, 22, 15, 52 a) 22 b) 12 c) 52 d) 3	1	2	5	2	2.7.1
8	The learning algorithms that can deal with both minimal labelled dataset and large unlabelled dataset together is called _____. a) Supervised b) Unsupervised c) Semi supervised d) Reinforcement	1	1	5	2	2.7.1
9	In customer relationship management, we can detect outlier customers using _____. a) Data sparsity b) Contextual outlier detection c) Collective outlier detection d) Conventional Outlier Detection	1	1	6	2	2.7.1
10	Internet search engine are tasks related to the area of _____. a) Information retrieval b) Information storage c) Information cluster d) Information visualization	1	1	6	8	8.4.2

Part – B (4 x 5 = 20 Marks) Answer any 4 Questions						
11	<p>Outline the K-Medoid Clustering Method.</p> <p>K-Medoids (also called Partitioning Around Medoid) algorithm was proposed in 1987 by Kaufman and Rousseeuw. A medoid can be defined as a point in the cluster, whose dissimilarities with all the other points in the cluster are minimum. The dissimilarity of the medoid(C_i) and object(P_i) is calculated by using $E = P_i - C_i$</p> <p>Algorithm: k-medoids. PAM, a k-medoids algorithm for central objects.</p> <p>Input:</p> <ul style="list-style-type: none">■ k: the number of clusters,■ D: a data set containing n objects. <p>Output: A set of k clusters.</p> <p>Method:</p> <ol style="list-style-type: none">(1) arbitrarily choose k objects in D as initial medoids(2) repeat(3) assign each remaining object to the cluster whose medoid is closest to it(4) randomly select a nonrepresentative object(5) compute the total cost, S, of the current clustering(6) if $S < 0$ then swap o_j with o_i(7) until no change	5	2	4	1	1.7.1
12	<p>List the application of cluster analysis.</p> <ul style="list-style-type: none">• Biology: taxonomy of living things: kingdom, phylum, class, order, family, genus and species• Information retrieval: document clustering• Land use: Identification of areas of similar land use in an earth observation database• Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs• City-planning: Identifying groups of houses according to their house type, value, and geographical location• Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults• Climate: understanding earth climate, find patterns of atmospheric and ocean• Economic Science: market research• Owing to the huge amounts of data collected in databases, cluster analysis has recently become a highly active topic in data mining research.	5	2	4	1	1.7.1
13	<p>How to Measuring Clustering Quality? Explain.</p> <ul style="list-style-type: none">• Dissimilarity/Similarity metric<ul style="list-style-type: none">• Similarity is expressed in terms of a distance	5	2	4	8	8.4.1

	<ul style="list-style-type: none"> function, typically metric: $d(i, j)$ The definitions of distance functions are usually rather different for interval-scaled, boolean, categorical, ordinal ratio, and vector variables Weights should be associated with different variables based on applications and data semantics Quality of clustering: <ul style="list-style-type: none"> There is usually a separate “quality” function that measures the “goodness” of a cluster. It is hard to define “similar enough” or “good enough” <ul style="list-style-type: none"> The answer is typically highly subjective 					
14	<p>Write about Generalized linear model and Mixed-effect model.</p> <p>Generalized linear models</p> <ul style="list-style-type: none"> allow a categorical response variable (or some transformation of it) to be related to a set of predictor variables similar to the modeling of a numeric response variable using linear regression include logistic regression and Poisson regression  <p>Mixed-effect models</p> <ul style="list-style-type: none"> For analyzing grouped data, i.e. data that can be classified according to one or more grouping variables Typically describe relationships between a response variable and some covariates in data grouped according to one or more factors 	5	2	5	2	2.6.4
15	<p>Explain in detail how data mining used for Retail and Telecommunication Industries.</p> <ul style="list-style-type: none"> Retail industry: huge amounts of data on sales, customer shopping history, e-commerce, etc. Applications of retail data mining <ul style="list-style-type: none"> Identify customer buying behaviors Discover customer shopping patterns and trends Improve the quality of customer service Achieve better customer retention and satisfaction Enhance goods consumption ratios Design more effective goods transportation and distribution policies Telcomm. and many other industries: Share many similar goals and expectations of retail data mining Design and construction of data warehouses Multidimensional analysis of sales, customers, products, time, and region Analysis of the effectiveness of sales campaigns 	5	2	6	2	2.6.4

	<ul style="list-style-type: none"> Customer retention: Analysis of customer loyalty <ul style="list-style-type: none"> Use customer loyalty card information to register sequences of purchases of particular customers Use sequential pattern mining to investigate changes in customer consumption or loyalty Suggest adjustments on the pricing and variety of goods Product recommendation and cross-reference of items Fraudulent analysis and the identification of usual patterns Use of visualization tools in data analysis 					
Part – C (2 x 10 = 20 Marks)						
16	<p>Explain the K-Means algorithm in detail and Apply the K-means algorithm for the following five points (with (x, y) representing locations) into two clusters: A1(3, 10), A2(7, 5), A3(10, 4), A4(5, 9), A5(8, 5). Initial cluster centers are: A1(3, 10), and A4(5, 9)</p> <p><u>-Means Clustering Algorithm-</u></p> <p>K-Means Clustering Algorithm involves the following steps-</p> <p><u>Step-01:</u></p> <p>Choose the number of clusters K.</p> <p><u>Step-02:</u></p> <p>Randomly select any K data points as cluster centers. Select cluster centers in such a way that they are as farther as possible from each other.</p> <p><u>Step-03:</u></p> <p>Calculate the distance between each data point and each cluster center. The distance may be calculated either by using given distance function or by using euclidean distance formula.</p> <p><u>Step-04:</u></p> <p>Assign each data point to some cluster. A data point is assigned to that cluster whose center is nearest to that data point.</p> <p><u>Step-05:</u></p>	10	3	4	1	1.7.1

	<p>Re-compute the center of newly formed clusters.</p> <p>The center of a cluster is computed by taking mean of all the data points contained in that cluster.</p> <p><u>Step-06:</u></p> <p>Keep repeating the procedure from Step-03 to Step-05 until any of the following stopping criteria is met</p> <p>Center of newly formed clusters do not change</p> <p>Data points remain present in the same cluster</p> <p>Maximum number of iterations are reached</p> <p>Initial cluster centers are: A1(3, 10), and A4(5, 9)</p> <p>The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as-</p> $P(a, b) = x2 - x1 + y2 - y1 $ <p>Use K-Means Algorithm to find the two cluster centers after the iteration.</p>					
[OR]						
17	<p>Write about hierarchical clustering methods in detail.</p> <p>A Hierarchical clustering method works via grouping data into a tree of clusters. Hierarchical clustering begins by treating every data point as a separate cluster. Then, it repeatedly executes the subsequent steps:</p> <ol style="list-style-type: none"> 1. Identify the 2 clusters which can be closest together, and 2. Merge the 2 maximum comparable clusters. We need to continue these steps until all the clusters are merged together. <p>In Hierarchical Clustering, the aim is to produce a hierarchical series of nested clusters. A diagram called Dendrogram (A Dendrogram is a tree-like diagram that statistics the sequences of merges or splits) graphically represents this hierarchy and is an inverted tree that describes the order in which factors are merged (bottom-up view) or clusters are broken up (top-down view).</p> <p>The basic method to generate hierarchical clustering is</p> <p>1. Agglomerative: Initially consider every data point as an individual Cluster and at every step, merge the nearest pairs of the cluster. (It is a bottom-up method). At first, every dataset is considered as an individual entity or cluster. At every iteration, the clusters merge with different clusters until one cluster is formed.</p> <p>The algorithm for Agglomerative Hierarchical Clustering is:</p> <ul style="list-style-type: none"> • Calculate the similarity of one cluster with all the other clusters (calculate proximity matrix) • Consider every data point as an individual cluster • Merge the clusters which are highly similar or close to each other. • Recalculate the proximity matrix for each cluster • Repeat Steps 3 and 4 until only a single cluster remains. <p>Let's see the graphical representation of this algorithm using a</p>	10	3	4	8	8.4.1

	<p>dendrogram.</p> <p>This is just a demonstration of how the actual algorithm works no calculation has been performed below all the proximity among the clusters is assumed.</p> <p>Let's say we have six data points A, B, C, D, E, and F.</p>  <p>Figure – Agglomerative Hierarchical clustering</p> <ul style="list-style-type: none">• Step-1: Consider each alphabet as a single cluster and calculate the distance of one cluster from all the other clusters.• Step-2: In the second step comparable clusters are merged together to form a single cluster. Let's say cluster (B) and cluster (C) are very similar to each other therefore we merge them in the second step similarly to cluster (D) and (E) and at last, we get the clusters [(A), (BC), (DE), (F)]• Step-3: We recalculate the proximity according to the algorithm and merge the two nearest clusters([(DE), (F)]) together to form new clusters as [(A), (BC), (DEF)]• Step-4: Repeating the same process; The clusters DEF and BC are comparable and merged together to form a new cluster. We're now left with clusters [(A), (BCDEF)].• Step-5: At last the two remaining clusters are merged together to form a single cluster [(ABCDEF)]. <p>2. Divisive:</p> <p>We can say that the Divisive Hierarchical clustering is precisely the opposite of the Agglomerative Hierarchical clustering. In Divisive Hierarchical clustering, we take into account all of the data points as a single cluster and in every iteration, we separate the data points from the clusters which aren't comparable. In the end, we are left with N clusters.</p> 					
18	Interpret the supervised method for detecting the	10	3	5	2	2.4.1

	<p>outlier.</p> <p>Supervised methods model data normality and abnormality. Domain professionals tests and label a sample of the basic data. Outlier detection can be modeled as a classification issue. The service is to understand a classifier that can identify outliers.</p> <p>The sample can be used for training and testing. In various applications, the professionals can label only the normal objects, and several objects not connecting the model of normal objects are documented as outliers. There are different methods model the outliers and consider objects not connecting the model of outliers as normal.</p> <ul style="list-style-type: none"> Modeling outlier detection as a classification problem <ul style="list-style-type: none"> Samples examined by domain experts used for training & testing Methods for Learning a classifier for outlier detection effectively: <ul style="list-style-type: none"> Model normal objects & report those not matching the model as outliers, or Model outliers and treat those not matching the model as normal Challenges <ul style="list-style-type: none"> Imbalanced classes, i.e., outliers are rare: Boost the outlier class and make up some artificial outliers Catch as many outliers as possible, i.e., recall is more important than accuracy (i.e., not mislabeling normal objects as outliers) 					
[OR]						
19	<p>Explain in detail, how data mining algorithms can be used for Intrusion detection and prevention.</p> <ul style="list-style-type: none"> Majority of intrusion detection and prevention systems use <ul style="list-style-type: none"> Signature-based detection: use signatures, attack patterns that are preconfigured and predetermined by domain experts Anomaly-based detection: build profiles (models of normal behavior) and detect those that are substantially 	10	3	5	8	8.4.1

	<p>deviate from the profiles</p> <ul style="list-style-type: none"> • What data mining can help <ul style="list-style-type: none"> • New data mining algorithms for intrusion detection • Association, correlation, and discriminative pattern analysis help select and build discriminative classifiers • Analysis of stream data: outlier detection, clustering, model shifting • Distributed data mining <p>Visualization and querying tools</p> <p>How does data mining help in Intrusion detection and prevention</p> <p>Modern network technologies require a high level of security controls to ensure safe and trusted communication of information between the user and a client. An intrusion Detection System is to protect the system after the failure of traditional technologies. Data mining is the extraction of appropriate features from a large amount of data. And, it supports various learning algorithms, i.e. supervised and unsupervised. Intrusion detection is basically a data-centric process so, with the help of data mining algorithms, IDS will also learn from past intrusions, and improve performance from experience along with find unusual activities. It helps in exploring the large increase in the database and gather only valid information by improving segmentation and help organizations in real-time plan and save time. It has various applications such as detecting anomalous behavior, detecting fraud and abuse, terrorist activities, and investigating crimes through lie detection. Below list of areas in which data mining technology can be carried out for intrusion detection.</p> <ul style="list-style-type: none"> • Using data mining algorithms for developing a new model for IDS: Data mining algorithm for the IDS model having a higher efficiency rate and lower false alarms. Data mining algorithms can be used for both signature-based and anomaly-based detection. In signature-based detection, training information is classified as either “normal” or “intrusion.” A classifier can then be derived to discover acknowledged intrusions. Research on this place has included the software of clarification algorithms, association rule mining, and cost-sensitive modeling. Anomaly-primarily based totally detection builds models of normal behavior and automatically detect s massive deviations from it. Methods consist of the software of clustering, outlier analysis, and class algorithms, and statistical approaches. The strategies used have to be efficient and scalable, and able to dealing 					
--	--	--	--	--	--	--

	<p>with community information of excessive volume, dimensional, and heterogeneity.</p> <ul style="list-style-type: none"> • Analysis of Stream data: Analysis of stream data means is analyzing the data in a continuous manner but data mining is basically used on static data rather than Streaming due to complex calculation and high processing time. Due to the dynamic nature of intrusions and malicious attacks, it is more critical to perform intrusion detection withinside the records stream environment. Moreover, an event can be ordinary on its own but taken into consideration malicious if regarded as a part of a series of activities. Thus, it's far essential to look at what sequences of activities are regularly encountered together, locate sequential patterns, and pick out outliers. Other data mining strategies for locating evolving clusters and constructing dynamic class models in records streams also are essential for real-time intrusion detection. • Distributed data mining: It is used to analyze the random data which is inherently distributed into various databases so, it becomes difficult to integrate processing of the data. Intrusions may be launched from numerous distinctive places and focused on many distinctive destinations. Distributed data mining strategies can be used to investigate community data from numerous network places to detect those distributed attacks. • Visualization tools: These tools are used to represent the data in the form of graphs which helps the user to get a visual understanding of the data. These tools are also used for viewing any anomalous patterns detected. Such tools may encompass capabilities for viewing associations, discriminative patterns, clusters, and outliers. Intrusion detection structures must actually have a graphical user interface that permits safety analysts to pose queries concerning the network data or intrusion detection results. 					
--	--	--	--	--	--	--

***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

