

SET- A Key	
1	<p>The primary use of data cleaning is:</p> <ul style="list-style-type: none"> a. Removing the noisy data b. Correction of the data inconsistencies c. Transformations for correcting the wrong data d. All of the above
2	<p>Which of the following is an essential process in which the intelligent methods are applied to extract data patterns?</p> <ul style="list-style-type: none"> a. Data Mining b. Text Mining c. Data Selection d. Warehousing
3	<p>A graph that uses vertical bars to represent data is called a ____.</p> <ul style="list-style-type: none"> a. Bar graph b. Line graph c. Scatterplot d. All of the mentioned above
4	<p>A class consists of 50 students, out of which 30 are girls. The mean of marks scored by girls in a test is 73 (out of 100), and that of boys is 71. Determine the mean score of the whole class.</p> <ul style="list-style-type: none"> a. 72.9 b. 72.2 b. 74.2 c. 74.9
5	<p>Find the range of the first 10 multiples of 5.</p> <ul style="list-style-type: none"> a. 40 b. 48 c. 45 d. 50
6	<p>What is KDD in data mining?</p> <ul style="list-style-type: none"> a. Knowledge Discovery Database b. Knowledge Discovery Data c. Knowledge Data Definition d. Knowledge Data Discovery
7	<p>Determine the quartile 1 of the data set: 2, 3, 4, 5, 5, 5, 7</p> <ul style="list-style-type: none"> a. 2 b. 3 c. 4 d. 5
8	<p>The initial steps concerned in the process of knowledge discovery is,</p> <ul style="list-style-type: none"> a. Data Selection b. Data Integration c. Data Cleaning

	d. Data Transformation
9	<p>The issues of “Scalability and efficiency of the data mining algorithms” come under:</p> <ul style="list-style-type: none"> a. User Interaction and Mining Methodology Issues b. Diverse Data Types Issues c. Performance Issues d. None of the above
10	<p>The data objects that don’t comply with the general model or behaviour of the available data are</p> <ul style="list-style-type: none"> a. Evolution Analysis b. Outlier Analysis c. Classification d. Prediction

Part-B

11. List the type of attributes and explain any two of it with example.

An attribute is a data field, representing a characteristic or feature of a data object. The nouns attribute, dimension, feature, and variable are often used interchangeably in the literature. The type of an attribute is determined by the set of possible values—nominal, binary, ordinal, or numeric—the attribute can have. In the following subsections, we introduce each type.

Nominal Attributes

Nominal means “relating to names.” The values of a nominal attribute are symbols or names of things. Each value represents some kind of category, code, or state, and so nominal attributes are also referred to as categorical. The values do not have any meaningful order. In computer science, the values are also known as enumerations.

Example: Nominal attributes. Suppose that hair color and marital status are two attributes describing person objects. In our application, possible values for hair color are black, brown, blond, red, auburn, gray, and white. The attribute marital status can take on the values single, married, divorced, and widowed. Both hair color and marital status are nominal attributes. Another example of a nominal attribute is occupation, with the values teacher, dentist, programmer, farmer, and so on.

Ordinal Attributes

An ordinal attribute is an attribute with possible values that have a meaningful order or ranking among them, but the magnitude between successive values is not known. Example: Ordinal attributes. Suppose that drink size corresponds to the size of drinks available at a fast-food restaurant. This nominal attribute has three possible values: small, medium, and large. The values have a meaningful sequence (which corresponds to increasing drink size); however, we cannot tell from the values how much bigger, say, a medium is than a large. Other examples of ordinal attributes include grade (e.g., A+, A, A–, B+, and so on) and professional rank. Professional ranks can be enumerated in a sequential order: for example, assistant, associate, and full for professors, and private, private first class, specialist, corporal, and sergeant for army ranks.

12. Briefly discuss about Quantile plot and Scatter plot with example.

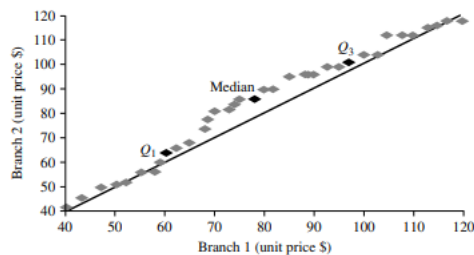
Quantile Plot

A quantile plot, or q-q plot, graphs the quantiles of one univariate distribution against the corresponding quantiles of another. It is a powerful visualization tool in that it allows the user to view whether there is a shift in going from one distribution to another.

Eg. A Set of Unit Price Data for Items Sold at a Branch of All Electronics

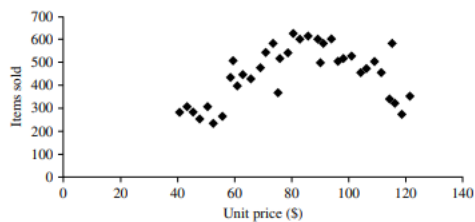
A Set of Unit Price Data for Items Sold at a Branch of *AllElectronics*

Unit price (\$)	Count of items sold
40	275
43	300
47	250
—	—
74	360
75	515
78	540
—	—
115	320
117	270
120	350



Scatter plot

A scatter plot is one of the most effective graphical methods for determining if there appears to be a relationship, pattern, or trend between two numeric attributes. The scatter plot is a useful method for providing a first look at bivariate data to see clusters of points and outliers, or to explore the possibility of correlation relationships. Two attributes, X, and Y, are correlated if one attribute implies the other. Correlations can be positive, negative, or null (uncorrelated). If the plotted points pattern slopes from lower left to upper right, this means that the values of X increase as the values of Y increase, suggesting a positive correlation. If the pattern of plotted points slopes from upper left to lower right, the values of X increase as the values of Y decrease, suggesting a negative correlation.



Part- C

13. Explain in detail about the kinds of pattern that can be mined.

1. Class/Concept Description: Characterization and Discrimination

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. For example, to study the characteristics of software products with sales that increased by 10% in the previous year, the data related to such products can be collected by executing an SQL query on the sales database.

Data discrimination is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes. The target and contrasting classes can be specified by a user, and the corresponding data objects can be retrieved through database queries.

For example, a user may want to compare the general features of software products with sales that increased by 10% last year against those with sales that decreased by at least 30% during the same period. The methods used for data discrimination are similar to those used for data characterization. “How are discrimination descriptions output?” The forms of output presentation are similar to those for characteristic descriptions, although discrimination descriptions should include comparative measures that help to distinguish between the target and contrasting classes. Discrimination descriptions expressed in the form of rules are referred to as discriminant rules

2. Mining Frequent Patterns, Associations, and Correlations

Frequent patterns, as the name suggests, are patterns that occur frequently in data. There are many kinds of frequent patterns, including frequent itemsets, frequent subsequences (also known as sequential patterns), and frequent substructures. A frequent itemset typically refers to a set of items that often appear together in a transactional data set—for example, milk and bread, which are frequently bought together in grocery stores by many customers. A frequently occurring subsequence, such as the pattern that customers tend to purchase first a laptop, followed by a digital camera, and then a memory card, is a (frequent) sequential pattern. A substructure can refer to different structural forms (e.g., graphs, trees, or lattices) that may be combined with itemsets or subsequences. If a substructure occurs frequently, it is called a (frequent) structured pattern. Mining frequent patterns leads to the discovery of interesting associations and correlations within data.

Association analysis. Suppose that, as a marketing manager at All Electronics, you want to know which items are frequently purchased together (i.e., within the same transaction). An example of such a rule, mined from the All Electronics transactional database, is $\text{buys}(X, \text{“computer”}) \Rightarrow \text{buys}(X, \text{“software”})$ [support = 1%, confidence = 50%], where X is a variable representing a customer. A confidence, or certainty, of 50% means that if a customer buys a computer, there is a 50% chance that she will buy software as well. A 1% support means that 1% of all the transactions under analysis show that computer and software are purchased together. This association rule involves a single attribute or predicate (i.e., buys) that repeats. Association rules that contain a single predicate are referred to as single-dimensional association rules. Dropping the predicate notation, the rule can be written simply as “computer \Rightarrow software [1%, 50%].”

3. Classification and Regression for Predictive Analysis

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts. The model is derived based on the analysis of a set of training data (i.e., data objects for which the class labels are known). The model is used to predict the class label of objects for which the class label is unknown.

Regression analysis is a statistical methodology that is most often used for numeric prediction, although other methods exist as well. Regression also encompasses the identification of distribution trends based on the available data.

4. Cluster Analysis

Unlike classification and regression, which analyze class-labeled (training) data sets, clustering analyzes data objects without consulting class labels. In many cases, class labeled data may simply not exist at the beginning. Clustering can be used to generate class labels for a group of data. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity.

5. Outlier Analysis

A data set may contain objects that do not comply with the general behavior or model of the data. These data objects are outliers. Many data mining methods discard outliers as noise or exceptions. However, in some applications (e.g., fraud detection) the rare events can be more interesting than the more regularly occurring ones. The analysis of outlier data is referred to as outlier analysis or anomaly mining. Outliers may be detected using statistical tests that assume a distribution or probability model for the data, or using distance measures where objects that are remote from any other cluster are considered outliers. Rather than using statistical or distance measures, density-based methods may identify outliers in a local region, although they look normal from a global statistical distribution view.

14. Discuss in detail about the different strategies of data reduction.

Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results.

a. Overview of Data Reduction Strategies

Data reduction strategies include dimensionality reduction, numerosity reduction, and data compression. Dimensionality reduction is the process of reducing the number of random variables or attributes under consideration. Dimensionality reduction methods include wavelet transforms and principal components analysis, which transform or project the original data onto a smaller space. Attribute subset selection is a method of dimensionality reduction in which irrelevant, weakly relevant, or redundant attributes or dimensions are detected and removed. Numerosity reduction techniques replace the original data volume by alternative, smaller forms of data representation. These techniques may be parametric or non parametric. For parametric methods, a model is used to estimate the data, so that typically only the data parameters need to be stored, instead of the actual data. (Outliers may also be stored.) Regression and log-linear models are examples. Nonparametric methods for storing reduced representations of the data include histograms clustering, sampling, and data cube aggregation.

b. Wavelet Transforms

The discrete wavelet transform (DWT) is a linear signal processing technique that, when applied to a data vector X , transforms it to a numerically different vector, X_0 , of wavelet coefficients. The two vectors are of the same length. When applying this technique to data reduction, we consider each tuple as an n -dimensional data vector, that is, $X = (x_1, x_2, \dots, x_n)$, depicting n measurements made on the tuple from n database attributes.³ “How can this technique be useful for data reduction if the wavelet transformed data are of the same length as the original data?” The usefulness lies in the fact that the wavelet transformed data can be truncated. A compressed approximation of the data can be retained by storing only a small fraction of the strongest of the wavelet coefficients. For example, all wavelet coefficients larger than some user-specified threshold can be retained. All other coefficients are set to 0. The resulting data representation is therefore very sparse, so that operations that can take advantage of data sparsity are computationally very fast if performed in wavelet space.

c. Principal Components Analysis

Suppose that the data to be reduced consist of tuples or data vectors described by n attributes or dimensions. Principal components analysis (PCA; also called the Karhunen-Loeve, or K-L, method) searches for k n -dimensional orthogonal vectors that can best be used to represent the data, where $k \leq n$. The original data are thus projected onto a much smaller space, resulting in dimensionality reduction. Unlike attribute subset selection (Section 3.4.4), which reduces the attribute set size by retaining a subset of the initial set of attributes, PCA “combines” the essence of attributes by creating an alternative, smaller set of variables. The initial data can then be projected onto this smaller set. PCA often reveals relationships that were not previously suspected and thereby allows interpretations that would not ordinarily result.

d. Attribute Subset Selection

Attribute subset selection reduces the data set size by removing irrelevant or redundant attributes (or dimensions). The goal of attribute subset selection is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes. Mining on a reduced set of attributes has an additional benefit: It reduces the number of attributes appearing in the discovered patterns, helping to make the patterns easier to understand.

e. Regression and Log-Linear Models: Parametric Data Reduction

Regression and log-linear models can be used to approximate the given data. In (simple) linear regression, the data are modeled to fit a straight line. For example, a random variable, y (called a response variable), can be modeled as a linear function of another random variable, x (called a predictor variable), with the equation $y = wx + b$, where the variance of y is assumed to be constant. In the context of data mining, x and y are numeric database attributes. The coefficients, w and b (called regression coefficients) specify the slope of the line and the y -intercept, respectively.

f. Log-linear models approximate discrete multidimensional probability distributions.

Given a set of tuples in n dimensions (e.g., described by n attributes), we can consider each tuple as a point in an n -dimensional space. Log-linear models can be used to estimate the probability of each point in a multidimensional space for a set of discretized attributes, based on a smaller subset of dimensional combinations. This allows a higher-dimensional data space to be constructed from lower-dimensional spaces. Log-linear models are therefore also useful for dimensionality reduction (since the lower-dimensional points together typically occupy less space than the original data points) and data smoothing (since aggregate estimates in the lower-dimensional space are less subject to sampling variations than the estimates in the higher-dimensional space).

g. Histograms

Histograms use binning to approximate data distributions and are a popular form of data reduction. A histogram for an attribute, A , partitions the data distribution of A into disjoint subsets, referred to as buckets or bins. If each bucket represents only a single attribute–value/frequency pair, the buckets are called singleton buckets. Often, buckets instead represent continuous ranges for the given attribute.

h. Clustering

Clustering techniques consider data tuples as objects. They partition the objects into groups, or clusters, so that objects within a cluster are “similar” to one another and “dissimilar” to objects in other clusters. Similarity is commonly defined in terms of how “close” the objects are in space, based on a distance function.

i. Sampling

Sampling can be used as a data reduction technique because it allows a large data set to be represented by a much smaller random data sample (or subset)

j. Data Cube Aggregation

Imagine that you have collected the data for your analysis. These data consist of the all Electronics sales per quarter, for the years 2008 to 2010. You are, however, interested in the annual sales (total per year), rather than the total per quarter. Thus, the data can be aggregated so that the resulting data summarize the total sales per year instead of per quarter.

Academic Year: 2023-24 (ODD SEMESTER)

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

Test: CLAT-1

Course Code & Title: 18CSE355T & DATA MINING AND ANALYTICS

Year & Sem: III & V /IV & VII

Date: 09.08.2023

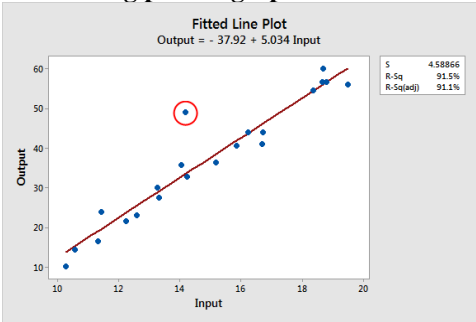
Duration: 1 Period

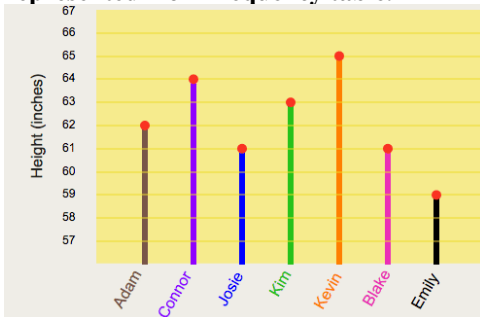
Max. Marks: 25

PART A

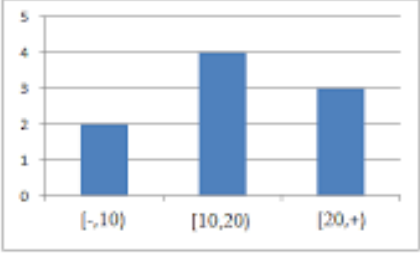
(10 x 1 = 10 Marks)

Instructions: Answer all Questions

1	Under which category does Sensor data belongs to? a. Time Series b. Data streams c. Spatial d. Networked	1	B2	2	2	2.6.2
2	Find out / Relate the type of Data Analysis on the data points, where a single data point which was encircled from the following plotted graph sketch. 	1	B2	3	1	2.7.1
3	Recall the data object(s), where noise reduction techniques exist on a. text b. image c. audio d. video Note: Both the Options are correct	1	B2	2	2	2.6.2
4	Match the standard / form of attribute noise from the cells with the string as N/A in the given attached table document.	1	B1	2	1	1.7.1

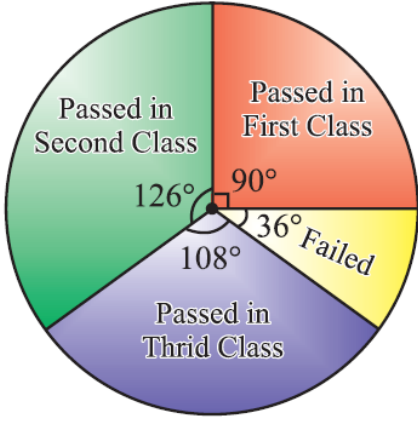
	<table><tr><th>PARTNERS</th><th>JAN 2017</th><th>FEB 2017</th><th>MAR 2017</th></tr><tr><td>West</td><td>200</td><td>100</td><td>200</td></tr><tr><td>East</td><td>N/A</td><td>600</td><td>N/A</td></tr><tr><td>South</td><td>300</td><td>N/A</td><td>400</td></tr><tr><td>North</td><td>600</td><td>300</td><td>200</td></tr></table> <p>a. Erroneous Values b. Missing Values c. Don't Care Values d. Borderline Values</p>	PARTNERS	JAN 2017	FEB 2017	MAR 2017	West	200	100	200	East	N/A	600	N/A	South	300	N/A	400	North	600	300	200					
PARTNERS	JAN 2017	FEB 2017	MAR 2017																							
West	200	100	200																							
East	N/A	600	N/A																							
South	300	N/A	400																							
North	600	300	200																							
5	<p>Find out / Recognize the mode from the given frequency distribution table</p> <table><tr><th>Age</th><th>Frequency</th></tr><tr><td>54</td><td>3</td></tr><tr><td>55</td><td>1</td></tr><tr><td>56</td><td>1</td></tr><tr><td>57</td><td>2</td></tr><tr><td>58</td><td>2</td></tr><tr><td>60</td><td>2</td></tr></table> <p>a. 57, 58, 60 b. 55 and 56 c. 54 d. N/A</p>	Age	Frequency	54	3	55	1	56	1	57	2	58	2	60	2	1	B1	2	1	1.7.1						
Age	Frequency																									
54	3																									
55	1																									
56	1																									
57	2																									
58	2																									
60	2																									
6	<p>Select / Calculate the value of IQR (Interquartile Range) Q2 from the given attribute values: attr 1: {2, 3, 4, 5, 6, 7, 8, 9}</p> <p>a. 2 b. 3 c. 4 d. 5</p>	1	B2	2	2	2.6.2																				
7	<p>Identify which chart have to be chosen for dataset, as represented from frequency table.</p>  <p>a. Pie Chart b. Bar Chart c. Histogram d. Spider Chart</p>	1	B1	2	1	1.7.1																				
8	<p>Interpret the method which would be used, data can be visualized for various age groups instead of a single age or single individuals.</p> <p>a. Binning b. Word Clouds c. Donut Charts d. Network Diagrams</p>	1	B3	3	2	2.6.2																				

9	<p>Name the data mining goal, where the system autonomously finds new patterns for presenting in an understandable form.</p> <ul style="list-style-type: none"> a. Verification b. Discovery c. Prediction d. Description 	1	B2	3	1	2.7.1
---	---	---	----	---	---	-------

10	<p>Interpret the method which is used to divide data into bins.</p>  <p>a. Equal Width Binning b. Equal Frequency Binning c. Supervised Binning d. Unsupervised Binning</p>	1	B3	3	2	2.6.2
----	---	---	----	---	---	-------

PART B (1 sx 5 = 5 Marks)

11	<p>Differentiate between KDD & Data Mining in terms of objectives & workflow.</p> <p><u>Answer:</u></p> <table><tr><td></td><td>KDD</td><td>Data Mining</td></tr><tr><td>Objective</td><td>Ensure useful high-level knowledge indeed it is executed experimentally: with feedback / corrections at each step</td><td>Extract patterns, no matter the quality. Indeed, Mining Algorithm could be blinded leads ‘Data dredging’</td></tr><tr><td>Optional Answer / point not mandatory</td><td>KDD – is a process, a methodology for extracting data leads to Knowledge</td><td>Data Mining is a step of KDD Workflow</td></tr><tr><td>Workflow</td><td>End to End Process Workflow Including 9 steps and different tasks</td><td>Application of specific and different algorithms for extracting patterns from data.</td></tr><tr><td colspan="3">Explanations with necessary Points could be awarded marks</td></tr></table>		KDD	Data Mining	Objective	Ensure useful high-level knowledge indeed it is executed experimentally: with feedback / corrections at each step	Extract patterns, no matter the quality. Indeed, Mining Algorithm could be blinded leads ‘Data dredging’	Optional Answer / point not mandatory	KDD – is a process, a methodology for extracting data leads to Knowledge	Data Mining is a step of KDD Workflow	Workflow	End to End Process Workflow Including 9 steps and different tasks	Application of specific and different algorithms for extracting patterns from data.	Explanations with necessary Points could be awarded marks			5	B3	2	2	2.8.2			
	KDD	Data Mining																						
Objective	Ensure useful high-level knowledge indeed it is executed experimentally: with feedback / corrections at each step	Extract patterns, no matter the quality. Indeed, Mining Algorithm could be blinded leads ‘Data dredging’																						
Optional Answer / point not mandatory	KDD – is a process, a methodology for extracting data leads to Knowledge	Data Mining is a step of KDD Workflow																						
Workflow	End to End Process Workflow Including 9 steps and different tasks	Application of specific and different algorithms for extracting patterns from data.																						
Explanations with necessary Points could be awarded marks																								
(OR)																								
12	<p>Display / Draw a Pie Chart from the given Frequency Table & Demonstrate the categorical Data / Variables with a mathematical formula.</p> <table><tr><td>Result</td><td>Percentage of students</td><td>Central angle</td></tr><tr><td>Passed in first class</td><td>25%</td><td>$\frac{25}{100} \times 360^\circ = 90^\circ$</td></tr><tr><td>Passed in second class</td><td>35%</td><td>$\frac{35}{100} \times 360^\circ = 126^\circ$</td></tr><tr><td>Passed in third class</td><td>30%</td><td>$\frac{30}{100} \times 360^\circ = 108^\circ$</td></tr><tr><td>Failed</td><td>10%</td><td>$\frac{10}{100} \times 360^\circ = 36^\circ$</td></tr><tr><td>Total</td><td>100%</td><td>360°</td></tr></table> <p><u>Solution:</u></p>	Result	Percentage of students	Central angle	Passed in first class	25%	$\frac{25}{100} \times 360^\circ = 90^\circ$	Passed in second class	35%	$\frac{35}{100} \times 360^\circ = 126^\circ$	Passed in third class	30%	$\frac{30}{100} \times 360^\circ = 108^\circ$	Failed	10%	$\frac{10}{100} \times 360^\circ = 36^\circ$	Total	100%	360°	5	B1	2	1	1.2.2
Result	Percentage of students	Central angle																						
Passed in first class	25%	$\frac{25}{100} \times 360^\circ = 90^\circ$																						
Passed in second class	35%	$\frac{35}{100} \times 360^\circ = 126^\circ$																						
Passed in third class	30%	$\frac{30}{100} \times 360^\circ = 108^\circ$																						
Failed	10%	$\frac{10}{100} \times 360^\circ = 36^\circ$																						
Total	100%	360°																						

	<p>1. <u>Pie Chart</u> - 2 Marks</p>  <p>2. <u>Mathematical Formula</u>: 1 Mark</p> $\frac{\theta_A}{360} = \frac{\# \text{ data points in category } A}{\text{Total number of data points}}$ <p>3. <u>Explanations on relative prepositions of data in different categories</u> – 2 Mark</p>					
--	---	--	--	--	--	--

PART C (1 x 10 = 10 Marks)

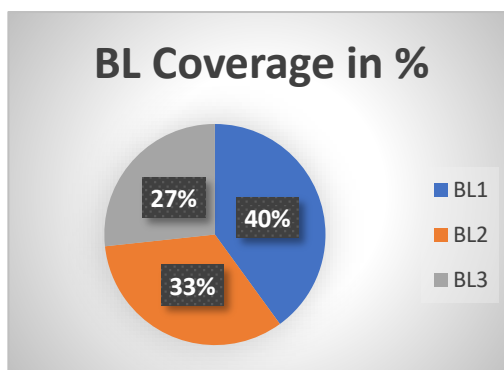
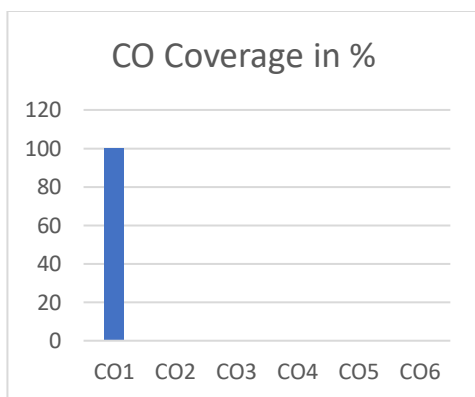
13	<p>Apply the Mathematical Formula & Employ the measuring Central Tendency on the following:</p> <p><u>Dataset:</u></p> <p>TABLE : Median Salary when n or N is an even number</p> <table><thead><tr><th colspan="2">Bi-weekly Salary</th></tr></thead><tbody><tr><td>1</td><td>2710</td></tr><tr><td>2</td><td>2755</td></tr><tr><td>3</td><td>2850</td></tr><tr><td>4</td><td>2880</td></tr><tr><td>5</td><td>2880</td></tr><tr><td>6</td><td>2890</td></tr><tr><td>7</td><td>2920</td></tr><tr><td>8</td><td>2940</td></tr><tr><td>9</td><td>2950</td></tr><tr><td>10</td><td>3050</td></tr><tr><td>11</td><td>3130</td></tr><tr><td>12</td><td>3325</td></tr></tbody></table> <p><u>Statistical Methods:</u></p> <p>i.) Mean ii.) Median iii.) Mode</p> <p><u>Solution:</u></p> <p>i.) <u>Mean:</u></p> <p>Formula: 1 Mark</p>	Bi-weekly Salary		1	2710	2	2755	3	2850	4	2880	5	2880	6	2890	7	2920	8	2940	9	2950	10	3050	11	3130	12	3325	10	B3	2	2	2.8.2
Bi-weekly Salary																																
1	2710																															
2	2755																															
3	2850																															
4	2880																															
5	2880																															
6	2890																															
7	2920																															
8	2940																															
9	2950																															
10	3050																															
11	3130																															
12	3325																															

	$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ $= \frac{x_1 + x_2 + \dots + x_{12}}{n}$ <p><u>Computation – 2.5 Marks</u> $= \frac{2710 + 2755 + 2850 + 2880 + 2880 + 2890 + 2920 + 2940 + 2950 + 3050 + 3130 + 3325}{12}$ $= 2940$</p> <p>ii.) <u>Median</u> <u>Formula:</u> 1 Mark A given data set of N distinct values is sorted in numerical order If N is Even, the median is the average of the middle two values N is Even Number Median Observation lies between 6th Observation and 7th Observation</p> <p><u>Computation – 2.5 Marks</u> Median Salary = $2890 + 2920 / 2$ $= 5810 / 2$ $= 2905$</p> <p>iii.) <u>Mode:</u> <u>Key Formulation</u> – 1 Mark The Mode for a set of data is the value that occurs <u>most frequently</u> in the set.</p> <p><u>Computation / Observation</u> – 2 Marks Frequently occurred data in the Dataset: 2880(4th Observation and 5th Observation)</p>					
(OR)						
14	<p>Perform smoothing on Sorted price values as on the following methods: <u>Data Points in the Dataset:</u> 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34 <u>Methods:</u> i.) Partition into three (equi-depth) bins ii.) Smoothing by bin means iii.) Smoothing by bin boundaries</p> <p><u>Solutions</u> i.) <u>Partition into three (equi-depth) bins</u> – 3 Marks Bin 1: 4, 8, 9, 15 Bin 2: 21, 21, 24, 25 Bin 3: 26, 28, 29, 34</p> <p>ii.) <u>Smoothing by bin means</u> – 3.5 Marks <u>Mean:</u> Bin 1 $\Rightarrow 4+8+9+15 \Rightarrow 36/4 \Rightarrow 9$ Bin 2 $\Rightarrow 21+21+24+25 \Rightarrow 91/4 \Rightarrow 22.75 \sim 23$ Bin 3 $\Rightarrow 26+28+29+34 \Rightarrow 117/4 \Rightarrow 29.25 \sim 29$</p> <p>Bin 1: 9, 9, 9, 9 Bin 2: 23, 23, 23, 23 Bin 3: 29, 29, 29, 29</p> <p>iii.) <u>Smoothing by bin boundaries</u> – 3.5 Marks Bin 1: 4, 4, 4, 15 (Min: 4 Max: 15) Bin 2: 21, 21, 25, 25 (Min: 21 Max: 25)</p>	10	B3	3	4	4.4.2

	Bin 3: 26, 26, 26, 34(Min: 26 Max: 34)					
	Bin 1: 4, 8, 9, 15					
	Bin 2: 21, 24, 24, 25					
	Bin 3: 26, 28, 29, 34					

***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



SET- C Key	
1	<p>Which of the following process uses intelligent methods to extract data patterns?</p> <p>a. Data mining b. Text mining c. Warehousing d. Data selection</p>
2	<p>Rahul's family drove through 7 states on summer vacation. The prices of Gasoline differ from state to state. Calculate the median of gasoline cost.</p> <p>1.79, 1.61, 2.09, 1.84, 1.96, 2.11, 1.75</p> <p>a. 1.61 b. 1.84 b. 1.75 c. 1.96</p>
3	<p>Smoothing is one of the strategies of</p> <p>a. Data Reduction b. Data Transformation c. Data Cleaning d. Data integration</p>
4	<p>Ramu downloaded a cloud dataset that has no class labels and he wants to generate the class label for the group of data. Which model he will choose?</p> <p>a. Classification b. Regression c. Logistic Regression d. Clustering</p>
5	<p>The classification or mapping of a class using a predefined class or group is called:</p> <p>a. Data Sub Structure b. Data Set c. Data Discrimination d. Data Characterisation</p>
6	<p>Which graph/chart is used to visually examine the relationship between two quantitative variables.</p> <p>a. Bar graph b. Scatterplot c. Line graph d. Pie chart</p>
7	<p>The weight of 8 boys in kgs are 54, 49, 51, 58, 61, 52, 54, 60. Find the median weight.</p> <p>a. 49 b. 51 c. 54 d. 58</p>
8	<p>Quartiles divide the entire set into</p> <p>a. 2 equal parts</p>

	b. 3 equal parts c. 4 equal parts 5 equal parts
9	The analysis that measures how strongly one attribute implies the other, based on the available data is a. Correlation analysis b. Variance c. Business analysis d. Performance analysis
10	The process of reducing the number of random variables or attributes under consideration is a. Numerosity reduction b. Dimensionality reduction c. Data Compression d. None of the above

Part-B

11. Write Short notes on kinds of data that can be mined.

As a general technology, data mining can be applied to any kind of data as long as the data are meaningful for a target application. The most basic forms of data for mining applications are database data, data warehouse data, and transactional data.

Database Data

- A database system, also called a database management system (DBMS), consists of a collection of interrelated data, known as a database, and a set of software programs to manage and access the data.
- The software programs provide mechanisms for defining database structures and data storage; for specifying and managing concurrent, shared, or distributed data access; and for ensuring consistency and security of the information stored despite system crashes or attempts at unauthorized access.
- A relational database is a collection of tables, each of which is assigned a unique name. Each table consists of a set of attributes (columns or fields) and usually stores a large set of tuples (records or rows).
- Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute values.
- A semantic data model, such as an entity-relationship (ER) data model, is often constructed for relational databases. An ER data model represents the database as a set of entities and their relationships.

Data Warehouses

- A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site. Data warehouses are constructed via a process of data cleaning, data

integration, data transformation, data loading, and periodic data refreshing.

- A data warehouse is usually modeled by a multidimensional data structure, called a data cube, in which each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure such as count or sum(sales amount). A data cube provides a multidimensional view of data and allows the precomputation and fast access of summarized data

Transactional Data

- In general, each record in a transactional database captures a transaction, such as a customer's purchase, a flight booking, or a user's clicks on a web page.
- A transaction typically includes a unique transaction identity number (trans ID) and a list of the items making up the transaction, such as the items purchased in the transaction.
- A transactional database may have additional tables, which contain other information related to the transactions, such as item description, information about the salesperson or the branch, and so on.

12. Find the mean, median, mode and range for the given data:

90, 94, 53, 68, 79, 94, 53, 65, 87, 90, 70, 69, 65, 89, 85, 53, 47, 61, 27,

80

Mention the formula for all the types.

Given,

90, 94, 53, 68, 79, 94, 53, 65, 87, 90, 70, 69, 65, 89, 85, 53, 47, 61, 27, 80

Number of observations = 20

Mean = (Sum of observations)/ Number of observations

= (90 + 94 + 53 + 68 + 79 + 94 + 53 + 65 + 87 + 90 + 70 + 69 + 65 + 89 + 85 + 53 + 47 + 61 + 27 + 80)/20

= 1419/20

= 70.95

Therefore, mean is 70.95.

Median:

The ascending order of given observations is:

27, 47, 53, 53, 53, 61, 65, 65, 68, 69, 70, 79, 80, 85, 87, 89, 90, 90, 94, 94

Here, $n = 20$

Median = $\frac{1}{2} [(n/2) + (n/2 + 1)]$ th observation

= $\frac{1}{2} [10 + 11]$ th observation

= $\frac{1}{2} (69 + 70)$

= $139/2$

= 69.5

Thus, the median is 69.5.

Mode:

The most frequently occurred value in the given data is 53.

Therefore, mode = 53

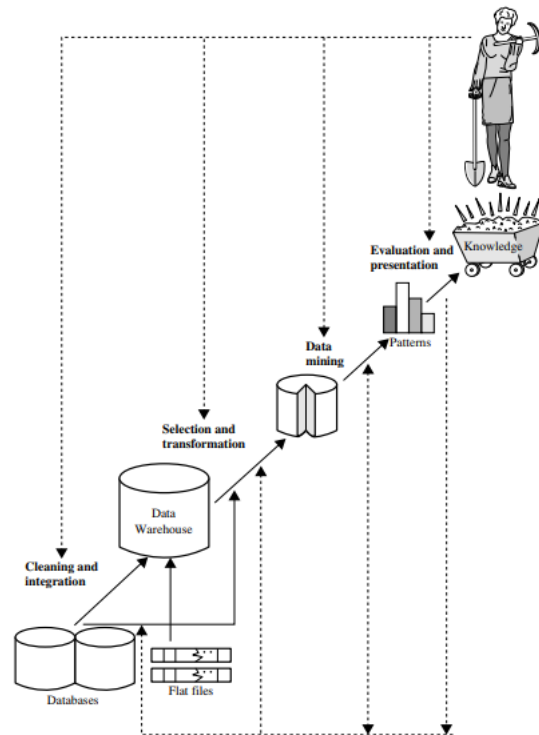
Range = Highest value – Lowest value

= $94 - 27$

= 67

Part- C

13. Explain the steps in the process of knowledge discovery in databases



1. Data cleaning (to remove noise and inconsistent data)
2. Data integration (where multiple data sources may be combined)
3. Data selection (where data relevant to the analysis task are retrieved from the database)
4. Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)
5. Data mining (an essential process where intelligent methods are applied to extract data patterns)
6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on interestingness measures)
7. Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users)

14. What is data cleaning? Explain the different methods of data cleaning.

Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

1. Missing Values
 - a. Ignore the tuple

- b. Fill in the missing value manually
- c. Use a global constant to fill in the missing value
- d. Use a measure of central tendency for the attribute (e.g., the mean or median) to fill in the missing value
- e. Use the attribute mean or median for all samples belonging to the same class as the given tuple
- f. Use the most probable value to fill in the missing value

2. Noisy Data “

Noise is a random error or variance in a measured variable. Some basic statistical description techniques (e.g., boxplots and scatter plots), and methods of data visualization can be used to identify outliers, which may represent noise. The following are the data smoothing techniques

- a. Binning
- b. Regression
- c. Outlier analysis

3. Data Cleaning as a Process

The first step in data cleaning as a process is discrepancy detection. Discrepancies can be caused by several factors, including poorly designed data entry forms that have many optional fields, human error in data entry, deliberate errors (e.g., respondents not wanting to divulge information about themselves), and data decay (e.g., outdated addresses). Discrepancies may also arise from inconsistent data representations and inconsistent use of codes. Other sources of discrepancies include errors in instrumentation devices that record data and system errors. Errors can also occur when the data are (inadequately) used for purposes other than originally intended. There may also be inconsistencies due to data integration (e.g., where a given attribute can have different names in different databases).

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

Test: CLAT-1

Course Code & Title: 18CSE355T & DATA MINING AND ANALYTICS

Year & Sem: III & V /IV & VII

Date: 09.08.2023

Duration: 1 Period

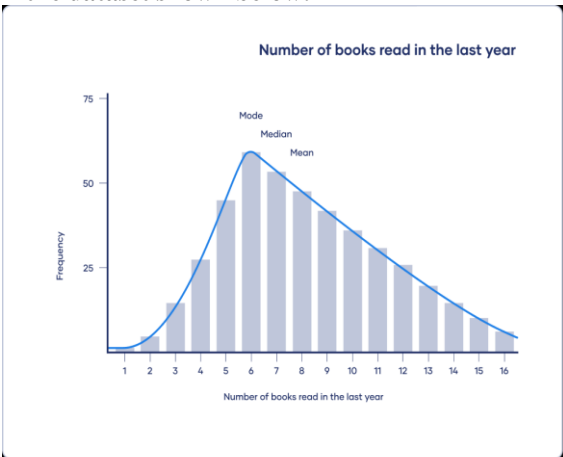
Max. Marks: 25

PART A

(10 x 1 = 10 Marks)

Instructions: Answer all Questions:

1	Infer & choose the notation of values more than numeric attribute x on k th q-quantile for a given data distribution <u>Note:</u> k is an integer such that $0 < k < q$ a. k / q b. $(q - k) / q$ c. 0 d. $(k - q) / q$	1	B2	2	3	3.5.4
2	Infer & choose the method / rules from the following one, which is used in classification techniques as for representing a derived model. a. if-then Statement b. if-then-else Statement c. switch Statement d. for Statement	1	B2	2	3	3.5.4
3	State the Graphic Display type from the option, which compare data across categories. a. Boxplot b. Bar Chart c. Histogram d. Scatter Plot	1	B1	2	1	1.7.1
4	Select the percentile in Q3 Quartile from distribution of equal-sized consecutive subsets of data. a. 25% b. 50% c. 75% d. 100%	1	B2	2	2	2.6.2
5	Record the type of Quadrant as present of negative values in x- axis and y- axis on the plotting graph. a. Quadrant I b. Quadrant II c. Quadrant III d. Quadrant IV	1	B1	2	1	1.7.1

6	<p>Make use of which attribute type does miles per hour (Speed = 20) are categorized as Fast / Slow?</p> <p>a. Ordinal b. Nominal c. Binary d. Interval</p>	1	B2	3	2	1.7.1																
7	<p>Specify the correct option from the following</p> <p>a. Noise data unnecessarily decreases the storage capacity b. Outliers are not present in the data c. Outliers do belong to the range of the majority datapoint in the dataset d. cause barrier in the result obtained</p>	1	B1	2	1	1.2.2																
8	<p>Show the median value from the given dataset observations for 7 participants on a computer task.</p> <table><tr><td>Participant</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td></tr><tr><td>Speed</td><td>Medium</td><td>Slow</td><td>Fast</td><td>Fast</td><td>Medium</td><td>Fast</td><td>Slow</td></tr></table> <p>a. Medium b. Slow c. Fast d. 0</p>	Participant	1	2	3	4	5	6	7	Speed	Medium	Slow	Fast	Fast	Medium	Fast	Slow	1	B1	2	1	1.7.1
Participant	1	2	3	4	5	6	7															
Speed	Medium	Slow	Fast	Fast	Medium	Fast	Slow															
9	<p>Match the correct option from the observations of data as, If a data set had values of 2, 4 and 6, the normalized value of the first data point is</p> <p>a. 0 b. 0.5 c. 1 d. 1.5</p>	1	B1	2	1	1.7.1																
10	<p>Name the type of distribution in the given plotted graph from the dataset shown below:</p> <div><p>Number of books read in the last year</p></div> <p>a. Normal Distribution b. Positively Skewed Distribution c. Negatively Skewed Distribution d. Zero Skewed Distribution</p>	1	B2	3	1	2.7.1																

PART B (1 x 5 = 5 Marks)

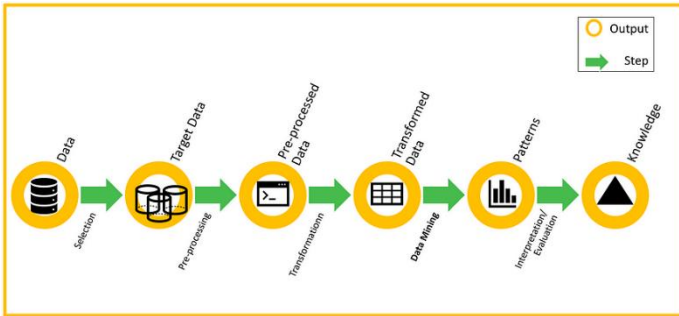
11	<p>Apply the concept of Inter Quartile Range (IQR) for the given scenario as: Suppose the distribution of math scores in a class of 19 students in ascending order is: 59, 60, 65, 65, 68, 69, 70, 72, 75, 75, 76, 77, 81, 82, 84, 87, 90, 95, 98 Calculate median (Q2) from the observations of data in the dataset. Solution:</p>	5	B2	2	2	2.6.2
----	--	---	----	---	---	-------

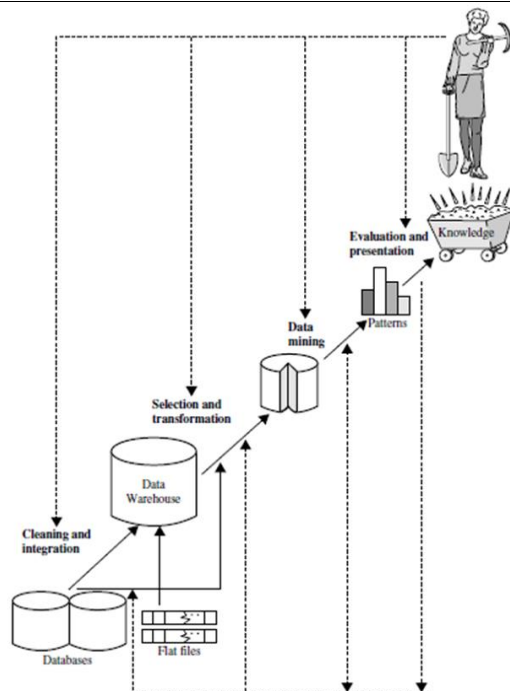
	<p>First, mark down the median, Q2, which in this case is the 10th value: 75.</p> <p>Q1 is the central point between the smallest score and the median. In this case, Q1 falls between the first and fifth score: 68(5th Observation)</p> <p>Note: that the median can also be included when calculating Q1 or Q3 for an odd set of values. If We were to include the median on either side of the middle point, then Q1 will be the middle value between the first and 10th score, which is the average of the fifth and sixth score— (fifth + sixth)/2 = (68 + 69)/2 = 68.5.</p> <p>Q3 is the middle value between Q2 and the highest score: 84(15th Observation) (Or if We include the median, Q3 = (82 + 84)/2 = 83).</p> <p>IQR = Q3 - Q2 = 84 - 68 = 16</p> <p>Quartiles:</p> <p>Interpret their numbers. A score of 68 (Q1) represents the first quartile and is the 25th percentile. Sixty-eight is the median of the lower half of the score set in the available data—that is, the median of the scores from 59 to 75.</p> <p>Q1 tells us that 25% of the scores are less than 68 and 75% of the class scores are greater. Q2 (the median) is the 50th percentile and shows that 50% of the scores are less than 75, and 50% of the scores are above 75. Finally, Q3, the 75th percentile, reveals that 25% of the scores are greater and 75% are less than 84.</p>					
(OR)						

12	<p>Analyze the sequence of operations on calculating the following for the given dataset: x= 2, 7, 3, 12, 9</p> <p>Methods: i.) Mean ii.) Standard Deviation</p> <p>Solution:</p> <p>Step 1: Calculate the <u>mean</u> 2 Marks $= 2 + 7 + 3 + 12 + 9 / 5$ $= 33 / 5$ $= \mathbf{6.6}$</p> <p>Step 2: Take each value in the data set, subtract the mean and square the difference.</p> <p>Formula: - 1 Mark</p> $\text{Variance} = \sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$ $\text{Standard Deviation} = \sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$	5	B2	2	2	2.6.5
----	--	---	----	---	---	-------

	<p>Variance:</p> <p><u>Computation</u> – 2 Marks</p> <p>For instances, 5 values:</p> $(2 - 6.6)^2 = 21.16$ $(7 - 6.6)^2 = 0.16$ $(3 - 6.6)^2 = 12.96$ $(12 - 6.6)^2 = 29.16$ $(9 - 6.6)^2 = 5.76$ <p>Squared Differences for all values are <u>added</u>:</p> $21.16 + 0.16 + 12.96 + 29.16 + 5.76 = 69.20$ <p>Sum is then divided by number of data points:</p> $69.20 \div 5 = 13.84$ <p>Variance = 13.84</p> <p><u>Standard Deviation</u></p> <p>Square Root of the variance</p> <p>Standard Deviation = 3.72</p>					
--	--	--	--	--	--	--

PART C (1 x 10 = 10 Marks)

13	<p>Draw a sketch & arrange the 9 Steps of Main Workflow in Knowledge Discovery of Databases. Discuss briefly about it.</p> <p><u>Answer:</u></p> <p>Schematic Sketch / Diagram – 3 Marks</p>  <p>(Or)</p>	10	B1	2	1	1.2.2
----	---	----	----	---	---	-------



Explanations: 7 Marks

1. Developing an understanding of the application domain - preparatory step for understanding what should be done with many decisions

if set wrong, can lead to false interpretations and negative impacts on the end-user

2. Selecting and creating a data set: this includes finding out what data is available and select a subset on which discovery will be performed

3. Pre-processing and cleaning: in this stage data reliability is enhanced, it includes data cleaning such as handling missing values and removal of noise or outliers, redundant and low-quality data from the data set in order to improve the reliability of the data

4. Data transformation: in this stage the generation of better data for the data mining is prepared and developed

5. Choosing the appropriate Data Mining task: we are ready to decide on which type of data mining to use, for example, classification, regression or clustering

6. Choosing the data mining algorithm: this stage includes selecting the specific method and so algorithm to be used for searching patterns in the data.

7. Employing the data mining algorithm: finally, the implementation of the data mining algorithm is reached, and algorithms are applied in order to extract data patterns.

8. Evaluation of mined patterns: in this stage we evaluate and interpret the mined patterns with respect to the goals defined in the first step

9. Using the discovered knowledge: we are now ready to incorporate the knowledge into another system for further action

(OR)

14	Interpret the methods & elaborate the working of normalization in data mining with each following example: Method: i.) Min-Max Normalization Dataset:	10	B3	3	4	4.5.1
----	---	----	----	---	---	-------

Employee Name	Years of Experience
ABC	8
XYZ	20
PQR	10
MNO	15

Solution:

Minimum value - 8

Maximum value - 20

As this formula scales the data between 0 and 1,

The new min is 0

The new max is 1

V - respective value of the attribute, i.e., 8, 10, 15, 20

Formula : 1 Mark

$$v' = \frac{v - \min_F}{\max_F - \min_F} (\text{new_max}_F - \text{new_min}_F) + \text{new_min}_F ,$$

Computation : 4 Mark

For 8 years of experience: $v' = (8 - 8 / 20 - 8) (1 - 0) + 0 = 0$

For 10 years of experience: $v' = (10 - 8 / 20 - 8) (1 - 0) + 0$
 $= 0.16$

For 15 years of experience: $v' = (15 - 8 / 20 - 8) (1 - 0) + 0$
 $= 0.58$

For 20 years of experience: $v' = (20 - 8 / 20 - 8) (1 - 0) + 0$
 $= 1$

Method:

ii.) Z-Score Normalization

Dataset: 3,5,5,8,9,12,12,13,15,16,17,19,22,24,25,134

Solution:

Step 1: Find out Mean

$$= 3 + 5 + 5 + 8 + 9 + 12 + 12 + 13 + 15 + 16 + 17 + 19 + 22 + 24 + 25 + 134 / 16$$

$$= 339 / 16$$

$$= 21.18 \sim 21.2$$

Mean = 21.2

$$\text{Variance} = \sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

$$\text{Standard Deviation} = \sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

Step 2:

Calculation on Standard Deviation – 2 Marks

Variance:

For instances, 16 values

$$\begin{aligned}
 (3 - 21.2)^2 &= 331.24 \\
 (5 - 21.2)^2 &= 262.44 \\
 (5 - 21.2)^2 &= 262.44 \\
 (8 - 21.2)^2 &= 29.16 \\
 (9 - 21.2)^2 &= 174.24 \\
 (12 - 21.2)^2 &= 84.64 \\
 (12 - 21.2)^2 &= 84.64 \\
 (13 - 21.2)^2 &= 67.24 \\
 (15 - 21.2)^2 &= 38.44 \\
 (16 - 21.2)^2 &= 27.04 \\
 (17 - 21.2)^2 &= 17.64 \\
 (19 - 21.2)^2 &= 4.84 \\
 (22 - 21.2)^2 &= 0.64 \\
 (24 - 21.2)^2 &= 7.84 \\
 (25 - 21.2)^2 &= 14.44 \\
 (134 - 21.2)^2 &= 12723.84
 \end{aligned}$$

The **squared differences** for all values are added

$$\text{Variance} = 14130.76 / 16 = \mathbf{883.17}$$

Standard Deviation:

$$\text{Square Root of } 883.17 = 29.718 \sim \mathbf{29.8}$$

Step 3:

Z-Score Normalization:

Formula: - 1 Mark

$$Z = \frac{X - \mu}{\sigma}$$

Z-Score Formula

where:

- X is the data point
- μ is the mean of the attribute values
- σ is the standard deviation of the attribute values

$$\text{New value} = (3 - 21.2) / 29.8$$

$$\therefore \text{New value} = \mathbf{-0.61}$$

Similarly for other Data,

Computation – 2 Marks

Z- Score Normalized Value are

$$\begin{aligned}
 (5 - 21.2) / 29.8 &= \mathbf{-0.54} \\
 (5 - 21.2) / 29.8 &= \mathbf{-0.54} \\
 (8 - 21.2) / 29.8 &= \mathbf{-0.44} \\
 (9 - 21.2) / 29.8 &= \mathbf{-0.41} \\
 (12 - 21.2) / 29.8 &= \mathbf{-0.31} \\
 (12 - 21.2) / 29.8 &= \mathbf{-0.31} \\
 (13 - 21.2) / 29.8 &= \mathbf{-0.28} \\
 (15 - 21.2) / 29.8 &= \mathbf{-0.21} \\
 (16 - 21.2) / 29.8 &= \mathbf{-0.17}
 \end{aligned}$$

	$(17 - 21.2) / 29.8 = -0.14$ $(19 - 21.2) / 29.8 = -0.07$ $(22 - 21.2) / 29.8 = 0.03$ $(24 - 21.2) / 29.8 = 0.09$ $(25 - 21.2) / 29.8 = 0.13$ $(134 - 21.2) / 29.8 = 3.79$					
--	---	--	--	--	--	--

***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

