CMR Engineering college : Hyderabad

UGC Autonomous

III - B. TECH - I - Semester End Examinations (Regular) - December - 2025

DATA MINING

(CSE) - R22CS504PC

PART-A

(1)

(a) What is Discretization?

A Discretization is the process of transforming continuous data into set of small intervals.

Ex: continuous age values : 22, 25, 37, 60

Discretized into categories

0-25 → "young"

26-50 → "Middle-aged"

50+ → "Senior"

So : 22 → young

37 → Midle-aged

60 → senior

(b) What is data cleaning?

A Data cleaning is defined as Removal of noisy and irrelevant/inconsistent data from data collection.

- cleaning in case of Missing values.
- cleaning noisy data, where noise is a random or variance error.
- Involves handling Missing values, replaced using statistical Methods such as Mean, Median, Mode.

(c) Define frequent item set?

A frequent itemset is a set of items that appear together in transactions or datasets

Ex:
If Many customers buy {Bread, Milk} together often, then {Bread, Milk} is a frequent itemset.

(d) Define Association Rule?

An association rule shows a relationship between two itemsets.

Ex: Bread → Butter

If a customer buys Bread, they are likely to buy Better.

(e) What is a Decision Tree?

A decision tree is a tree-structured classification or prediction model that uses a series of decision rules to split data into subsets based on attribute values.

It consists of root node, Internal nodes, Branches, leaf node.

(f) Define classification?

* classification is a supervised data mining technique used to assign data items to predefined categories or classes based on their features.

* techniques include Decision tree Induction, Naive Bayes, KNN, Bayesian Belief Networks.

(g) Differentiate between classification and clustering.

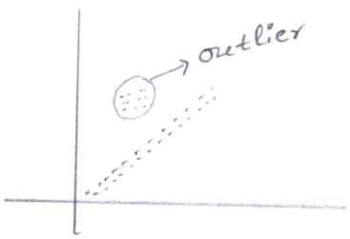
A

classification	clustering
* Supervised * works on labeled data. * Assign data to predefined classes. Algorithms: Decision Tee, Naive Bayes, KNN	* Un supervised * works on unlabled data. * grouping similar data items Algorithms, K-Means, PAM, Hierarchical clustering

(h) What is outlier Detection?

A It is the process of identifying data points that are significantly different from the majority of data.

* Indicates noise, inconsistant data.



Types:
contextual, collective, point Anomalies

(i) Write any two types of web Mining?

A
1) web content Mining
2) web structure Mining
3) web usage Mining

(j) What is Episode Rule Discovery in Text Mining?

A It is a technique used to identify frequent sequences of events (episodes) occur one after another.

* An episode is a set of sequence of events.

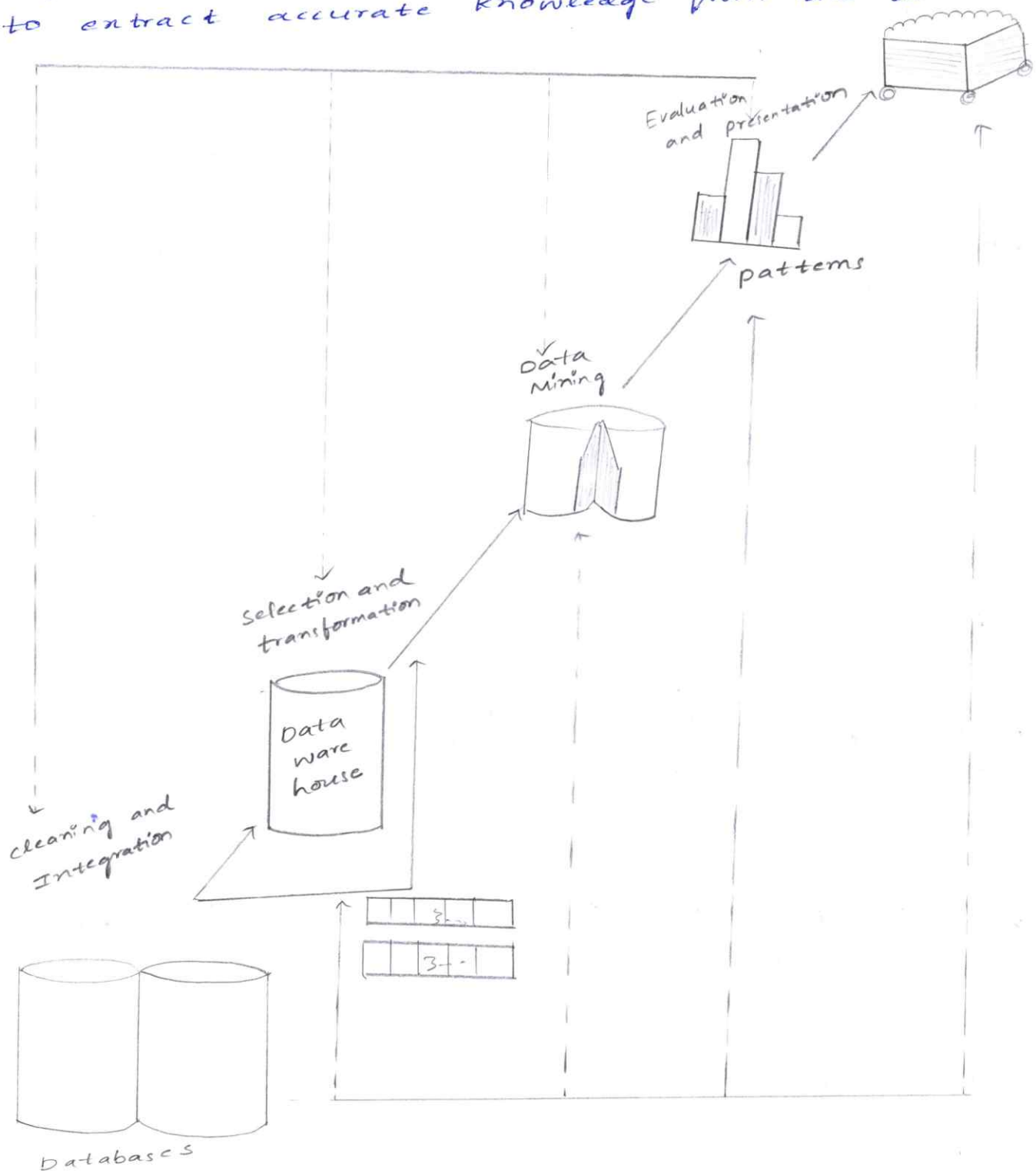
PART- B

(9) A

Explain the KDD process with a neat diagram.

KDD is a process that involves the extraction of useful, previously unknown, and potentially valuable information from large datasets.

The KDD process is an iterative process and it requires multiple iterations of the below steps to extract accurate knowledge from the data.



The following steps are included in KDD process.

1) Data cleaning
2) Data Integration
3) Data selection
4) Data transformation
5) Data Mining
6) pattern Evaluation
7) Knowledge Representation

① Data cleaning:

Data cleaning is defined as Removal of noisy and irrelevant / inconsistent data from data collection.

- cleaning in case of Missing values.
- cleaning noisy data, where noise is a random or variance error.

② Data Integration:

Data Integration is defined as heterogeneous data from multiple data sources combined in a common source (Data Warehouse).

i.e; In this step, multiple data sources may be combined as single data source.

③ Data selection:

Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection. This step in the KDD process is identifying and selecting the relevant data for analysis.

④ Data Transformation:

Data transformation is defined the process of transforming data into appropriate from required by mining procedure. This step involves reducing the data dimensionality, aggregating the data, normalizing it and discretizing it to prepare it for further analysis.

⑤ Data Mining:

This is the heart of the KDD process and involves applying various data mining techniques to the transformed data to discover hidden patterns, trends, relationships, and insights. A few of the most common data mining techniques include clustering, classification, association rule mining, and anomaly detection.

⑥ Pattern Evaluation:

Next step is to evaluate the discovered patterns to determine their usefulness and relevance. This involves assessing the quality of the patterns, evaluating their significance, and selecting the most promising patterns for further analysis.

⑦ Knowledge Representation:

This step involves representing the knowledge extracted from the data in a way humans can easily understand and use. This can be done through visualizations, reports, or other forms of communication that provide meaningful insights into the data.

③ Explain Data preprocessing techniques with suitable examples.

A Data pre-processing is the step in data science and Machine learning where the raw data is cleaned, transformed and prepared so it can be used effectively by models.

It helps in

- Removing noise and irrelevant data.
- Handling missing or inconsistent values.
- Making data suitable for algorithms.

Data Cleaning:

→ Data cleaning is the process of removing or correcting error and inconsistencies in a data set to ensure its quality.

→ It involves handling missing values, which can be replaced using statistical methods such as mean, median or mode or by removing entirely

→ In correct or (irrelevent) irrevalent data points are removed so they don't influence the analysis incorrectly.

ex: In a customer database if some "Age" values are missing, we might replace them with the median age, and if there are duplicate entries of the same customer, we remove them.

Data Integration:

→ Data Integration combines data from multiple sources into a single unified dataset.
→ often data is stored in different systems, such as databases, spread sheets, or external Apis, and needs to be merged for analysis.

Ex: Merging sales records from three different branches of a store into a single dataset for a nationwide sales analysis.

Data Reduction:

Data Reduction in data mining refers to the process of effectively reducing the amount while delivering the same or very similar analytical results.
→ It's an important step in managing huge database that aim to keep the most important information while also simplifying the data.
→ This decrease helps in accelerating data preprocessing and analysis, lowering storage needs and frequently enhancing the precision of mining outcomes.

Techniques for Data Reduction:

It is important for streamlining complex datasets improving algorithm efficiency and enhancing pattern extraction.
- we can remove noise and redundancy

1) Dimensionality Reduction.
2) Data compression.
3) numerosity Reduction
4) Discretization operation
5) Data cube Aggregation.

① Dimension Reduction:

* It is the process of reducing the data by Removing these features from the data,..

Three types:

i) Wavelet transformation in DM

ii) principal component Analysis

iii) feature subset selection or Attribute subset Selection.

② Data compression:

* It reduces the volume of data while preserving its essential information content. It is crucial for managing large datasets, improving storage efficiency, and accelerating data analysis.

i) Lossless compression → Allows for the exact reconstruction of the original data from the compressed version.

ii) Lossy compression → This achieves higher compression ratios by discarding some information deemed to be less important.

③ Numerosity Reduction:

* The volume of data is reduced by representing in a lower format.

Two types:

i) parametric : It holds assumption.
   1) Log-Linear
   2) regression

ii) Non-parametric: It does not hold assumption of data fitting in the model,

→ Histogram

→ clustering

→ Sampling

④ Discretization operation:

* It is a process of transforming continuous data into set of small intervals.

⑤ Data cube Aggregation:

* It is data Mining is a multi-dimensional array that contains pre-aggregated data for efficient analysis.

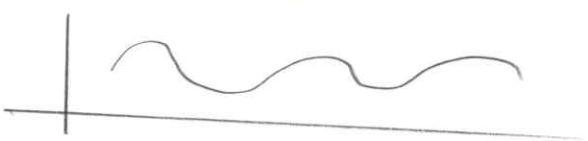
Data Transformation:

* It refers to the process of converting raw data into a format that is suitable for analysis and modeling.

Techniques of Data Transformation:

i) Smoothing:

* Used to remove noise from the data set using some algorithms. It allows for highlighting important features present in the dataset. It helps in predicting patterns.



2) Aggregation:

* It is the method of storing and presenting data in a summary format

Jan → 1000 Feb → 1200 March → 1,300

Monthly sales can be aggregated into to calculate quarterly sales

Total sales = $ 3,500

3) Discretization:

* It is a process of transforming continuous data into set of small intervals.

4) Attribute construction:

* New attributes are created and applied to assist the mining from given set of attributes.

5) Generalization:

* It converts low-level data attributes to high level data attributes using concept hierarchy.

6) Normalization:

* It involves converting all data variables into a given range.

Techniques:

1) Min-Max Normalization
2) z-score Normalization
3) Decimal Scaling

Example:

(1) Min-Max

Salary values: [20000, 50000, 80000]

After Min-Max Normalization (0-1 range):

- 20,000 → 0
- 50,000 → 0.5
- 80,000 → 1

(4) A

Explain the APRIORI algorithm with an example.

Apriori Algorithm most widely used algorithm for frequent itemset mining and association rule learning.

How Apriori algorithm works:

- Bottom-up: starts with finding frequent items, then combines them to find pairs, triplets & so on.
- Apriori property: If a smaller itemset isn't frequent, none of its larger versions can be either. This "prunes" search.
- Support and confidence: Two key measures used to define how often an itemset appears and how strong the association between items is.

Example:

Dataset:

Transaction	List of items
T_1	I_1, I_2, I_3
T_2	I_2, I_3, I_4
T_3	I_4, I_5
T_4	I_1, I_2, I_4
T_5	I_1, I_2, I_3, I_5
T_6	I_1, I_2, I_3, I_4

Min support = 50%.

Min confidence = 80%.

Step 1: (i) candidate set C_1

Item	count
I_1	4
I_2	5
I_3	4
I_4	4
I_5	2

I_5 is not equal meet to the minimum support threshold value is 3

(ii) pruning state: Remove I_5

Item	count
I_1	4
I_2	5
I_3	4
I_4	4

Step-2: Join state

(i) By using C_1, we can add 1 item to the candidate set C_2.

Items	count
I_1 I_2	4
I_1 I_3	3
I_1 I_4	2
I_2 I_3	4
I_2 I_4	3
I_3 I_4	2

(ii) pruning state:

Items	count
I_1 I_2	4
I_1 I_3	3
I_2 I_3	4
I_2 I_4	3

Step-3: Add 2 items to set C_3.

items	count
I_1, I_2, I_3	3
I_1, I_2, I_4	2
I_1, I_3, I_4	1
I_2, I_3, I_4	2

(iii) pruning state

We have to remove (I_1, I_2, I_4) (I_1, I_3, I_4) (I_2, I_3, I_4)

Step-4: confidence

(i) $\text{confidence}(\{I_1, I_2\} \rightarrow I_3) = \dfrac{\text{frequency}(I_1, I_2)}{\text{frequency}(I_3)}$

$= \dfrac{0.5}{0.6} = 0.83 = 83\%$

(ii) confidence $\{(I_2, I_3) \rightarrow I_1\}$ = $\frac{\text{frequency}(I_2 I_3)}{\text{frequency}(I_1)} = \frac{0.5}{0.6}$

= 83%

(iii) confidence $\{(I_1, I_3) \rightarrow I_2\}$

= $\frac{\text{frequency}(I_1, I_3)}{\text{frequency}(I_2)} = \frac{0.5}{0.5} = 100\%$

∴ Above are the confidences of the rule $\{I_1, I_2, I_3\}$.

---

(5) Explain in detail about closed Frequent Itemset with an example.

A A closed frequent itemset is a frequent itemset for which none of its immediate supersets has the same support count.

Example!

Let's consider a small database of four customer transaction and a minimum support threshold of 50%.

T_1 : {a, b, c, d}

T_2 : {a, b, c}

T_3 : {a, b, d}

T_4 : {a, b}

Steps to find closed frequent itemsets:

1) Identify all frequent itemsets: After running a standard frequent itemset mining algorithm, we find the following itemsets are frequent (appear in at least 2 transactions):

{a} : 4 transactions
{b} : 4 transactions
{c} : 2 transactions
{d} : 2 transactions
{a, b} : 4 transactions
{a, c} : 2 transactions
{a, d} : 2 transactions
{b, c} : 2 transactions
{b, d} : 2 transactions
{a, b, c} : 2 transactions
{a, b, d} : 2 transactions

2) check for the "closed" property: Now, for each frequent itemset, we check if any of its proper supersets have the exact same support count.

{a, b, c} : support - 2 it is closed frequent itemset.

{a, b, d} : support - 4, closed frequent itemset

{a, b} : support 4. It is a closed frequent itemset.

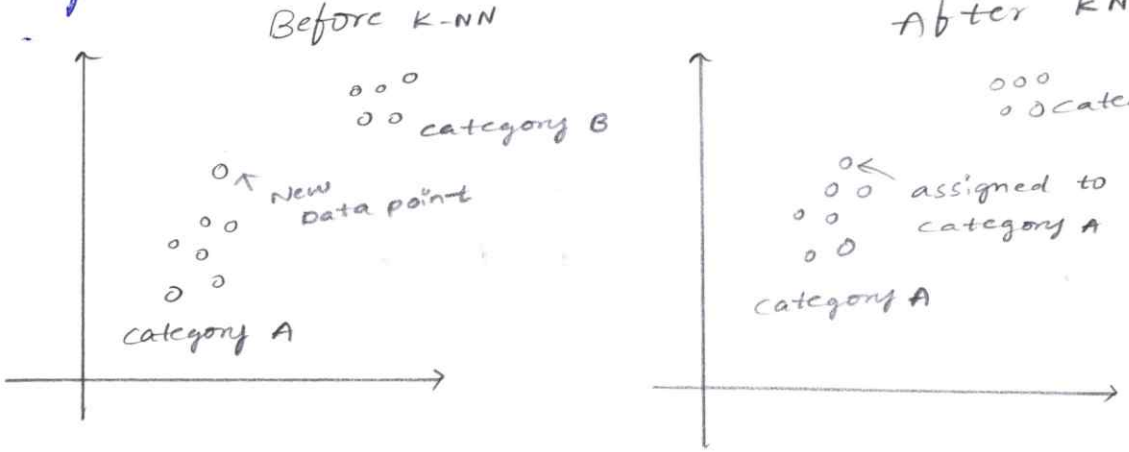
3) Final Result: The closed frequent itemsets are {a, b, c}, {a, b, d}, and {a, b}

(6) compare and contrast KNN and Decision Tree classifiers.

A KNN: K - Nearest Neighbours

* KNN is one of the simplest data mining technique based on supervised learning techniques.

* In KNN, K is just a number that tells the algorithm how many nearby points or neighbours.



Example:

Euclidean Distance : $\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$

Student	Maths	English	Result
A	4	3	F
B	6	7	P
C	7	8	P
D	5	5	F
E	8	8	P

Perform KNN, predict the class for [math = 6, English = 8] where K = 3

Distance:

For A : $\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$

$= \sqrt{(6-4)^2+(8-3)^2}$

$= 5.38$

FOR B

$$= \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$$

$$= \sqrt{(6-6)^2+(8-7)^2} = 1$$

FOR C

$$= \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$$

$$= \sqrt{(6-7)^2+(8-8)^2} = 1$$

FOR D

$$= \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$$

$$= \sqrt{(6-5)^2+(8-5)^2} = 3.16$$

FOR E

$$= \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$$

$$= \sqrt{(6-8)^2+(8-8)^2} = 2$$

Student	Math	English	Result	Distance
A	4	3	F	3.58
B	6	7	P	1
C	7	8	P	1
D	5	5	F	3.16
E	8	8	P	2

First K minimum values 1, 1, 2

∴ Here majority is pass. So for new data, result is pass.

Decision Tree

The Decision tree algorithm is a method used in supervised learning for solving both classification and regression problems.

It consists of root node, internal node, branches, leaf nodes.

Key points:

The tree splits data based on attributes using criteria such as

- Information Gain (ID3)
- Gain Ratio (C4.5)
- Gini Index (CART)

Comparison Table - Differences

KNN	Decision Tree
• Supervised	• Supervised
• Lazy Learner	• Eager Learner
• NO explicit model build	Builds a clear, rule-based tree model.
• Based on distance from neighbours.	• Based on attribute selection and splitting.
• Works with small datasets	• Works with large datasets.
• Highly sensitive to noisy	• Less sensitive due to tree structure.
• Needs future scaling.	• scaling is not required
• Euclidean distance $\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$	• Entropy = $-\sum P_i * \log_2(P_i)$ Gain = Entropy(S) $-\sum \frac{\lvert S_v \rvert}{S}$ Entropy(S_v) Gini Index = $-\sum (P_i)^2$
• Lazy Learner	• Eager Learner
• classification technique	• classification technique.

(7) Explain the K-Means clustering algorithm with a suitable example.

A

- K-Means number of groups or number of clusters are present.
- K-Means algorithm is an unsupervised learning Algorithm.
- For every cluster there is an center point it called as centroid-based Algorithm.

Algorithm:

- Input the dataset and decide the number of clusters K.
- Initialize K centroids randomly from the dataset.
- Assign each data points to the nearest centroid.
- Re-calculate the centroids as mean of all points in each cluster.
- check if centroids have changed:
  - If Yes → Repeat steps 4 and 5
  - If NO → stop
- Output final clusters with their centroids.

Euclidean Distance:

$$d\{(x_2, x_1), (y_2, y_1)\} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Data points	Distance to (185,72)	Distance to (170,56)	cluster	New cluster
(185, 72)				
(170,56)				
(168,60)				
(179,68)				
(182,72)				
(188,77)				

∴ K=2

The initial centroids

$c_1 = (185, 72)$

$c_2 = (170, 56)$

⇒ From $c_1(185, 72)$

At point (185, 72) $= \sqrt{(185-185)^2+(72-72)^2} = 0$

At point (170, 56) $= \sqrt{(170-185)^2+(56-72)^2} = 21.93$

At point (168, 60) $= \sqrt{(179-185)^2+(68-72)^2} = \sqrt{36+16} = 7.21$

At point (179, 68) $= \sqrt{(179-185)^2+(68-72)^2} = 7.21$

At point (182, 72) $= \sqrt{(182-185)^2+(72-72)^2} = 3$

At point (188, 77) $= \sqrt{(188-185)^2+(77-72)^2} = 5.83$

⇒ From $c_2(170, 56)$

At point (185, 72) $= \sqrt{(185-170)^2+(72-56)^2} = 21.93$

At point (170, 56) $= 0$

At point (168, 60) $= \sqrt{(168-170)^2+(60-56)^2} = 4.47$

At point (179, 68) $= \sqrt{(179-170)^2+(68-56)^2} = 15$

At point (182, 72) $= \sqrt{(182-170)^2+(72-56)^2} = 20$

At point (188, 77) $= \sqrt{(188-170)^2+(77-56)^2} = \sqrt{324+21} = 27.66$

Data points	Distance to (185, 72)	Distance to (170, 56)	cluster	New cluster
(185, 72)	0	21.93	c_1	
(170, 56)	21.93	0	c_2	
(168, 60)	20.81	4.47	c_2	
(179, 68)	7.21	15	c_1	
(182, 72)	3	20	c_1	
(188, 77)	5.83	27.66	c_1	

We have to select less distance

classify clusters c_1 or c_2 with less distance.

Second centroid values

C_1 :

$$\frac{\Sigma x_i}{n} = \frac{185+179+182+188}{4} = 183.5$$

C_2 : $\frac{\Sigma y_i}{n} = \frac{72+68+72+77}{4} = 72.25$

$(183.5, 72.25)$

C_2 : $\frac{\Sigma x_i}{n} = \frac{(170+168)}{2} = 169$

$\frac{\Sigma y_i}{n} = \frac{(56+60)}{2} = 58$ $(169, 58)$

dist for new centroids

⇒ From C_1 $(183.5, 72.25)$

At point $(185, 72) = \sqrt{(185-183.5)^2+(72-72.25)^2} = 1.52$

At point $(170, 56) = \sqrt{(170-183.5)^2+(56-72.25)^2} = 21.13$

At point $(168, 60) = \sqrt{(168-183.5)^2+(60-72.25)^2} = 19.76$

At point $(179, 68) = \sqrt{(179-183.5)^2+(68-72.25)^2} = 6.19$

At point $(182, 72) = \sqrt{(182-183.5)^2+(72-72.25)^2} = 1.52$

At point $(188, 77) = \sqrt{(188-183.5)^2+(77-72.25)^2} = 6.54$

⇒ From C_2 $(169, 58) = \sqrt{(185-169)^2+(72-58)^2} = 22.02$

At point $(185, 72) = \sqrt{(170-169)^2+(56-58)^2} = 2.83$

At point $(170, 56) = \sqrt{(168-169)^2+(60-58)^2} = 2.00$

At point $(168, 60) = \sqrt{(179-169)^2+(68-58)^2} = 14.87$

At point $(179, 68) = \sqrt{(182-169)^2+(72-58)^2} = 19.80$

At point $(182, 72) = \sqrt{(188-169)^2+(77-58)^2} = 27.59$

At point $(188, 77) = \sqrt{\quad}$

Data points	Distance to (183.5, 72.5)	Distance to (169, 58)	cluster	New cluster
(185, 72)	1.52	22.02	C_1	C_1
(170, 56)	21.13	2.83	C_2	C_2
(168, 60)	19.76	2.00	C_2	C_2
(179, 68)	6.19	14.87	C_1	C_1
(182, 72)	1.52	19.80	C_1	C_1
(188, 77)	6.54	27.59	C_1	C_1

∴ Final clusters:

C_1 : {(185, 72), (179, 68), (182, 72), (188, 77)}

C_2 : {(170, 56), (168, 60)}

Issues!

1) choice of K
2) sensitivity of initial centroids
3) Empty clusters
4) outliers and noise
5) Non-convex shapes
6) Different clusters sizes
7) Distance measure dependency
8) High-dimensional data problems.

(9) Explain the steps involved in Agglomerative Hierarchical clustering.

A
- It is a bottom-up approach
- starts with each object as a single cluster
- step by step, merges the two closest clusters.
- continues until all objects are in one big cluster.

Steps:
- start with each data point as its own cluster.
- calculate distance between all clusters.
- Merge the two closest clusters.
- Re-calculate distances between new clusters.
- Repeat until only one cluster remains.

Linkage Methods:

i) Single Linkage: Min dist b/w points of clusters
ii) complete Linkage: Maximum distance b/w clusters.
iii) Average Linkage: Average dist b/w all points in clusters.
iv) centroid Linkage: Distance between cluster centroids.

Example:

consider 18, 22, 25, 42, 27, 43

- Apply single Linkage method
- Merge clusters with Min distance and update Matrix accordingly.

step-1:

	18	22	25	27	42	43
18	0	4	7	9	24	25
22	4	0	3	5	20	21
25	7	3	0	2	17	18
27	9	5	2	0	15	16
42	24	20	17	15	0	1
43	25	21	18	16	1	0

42 & 43 have min distance (42, 43)

step-2:

	18	22	25	27	42,43
18	0	4	7	9	24
22	4	0	3	5	20
25	7	3	0	2	17
27	9	5	2	0	15
42,43	24	20	17	15	0

27 and 25 have min dist so we will merge them

step-3:

	18	22	25,27	42,43
18	0	4	7	24
22	4	0	3	20
25,27	7	3	0	15
42,43	24	20	15	0

22, & 25, 27 have the minimum distance ((27, 25), 22)
∴ merge them

step-4:

	18	22,25,27	42,43
18	0	4	24
22,25,27	4	0	15
42,43	24	15	0

18 & 22, 25, 27 have the minimum distance ((27,25)22)18)

step-5:

	18, 22, 25, 27	42, 43
18, 22, 25, 27	0	15
42, 43	15	0

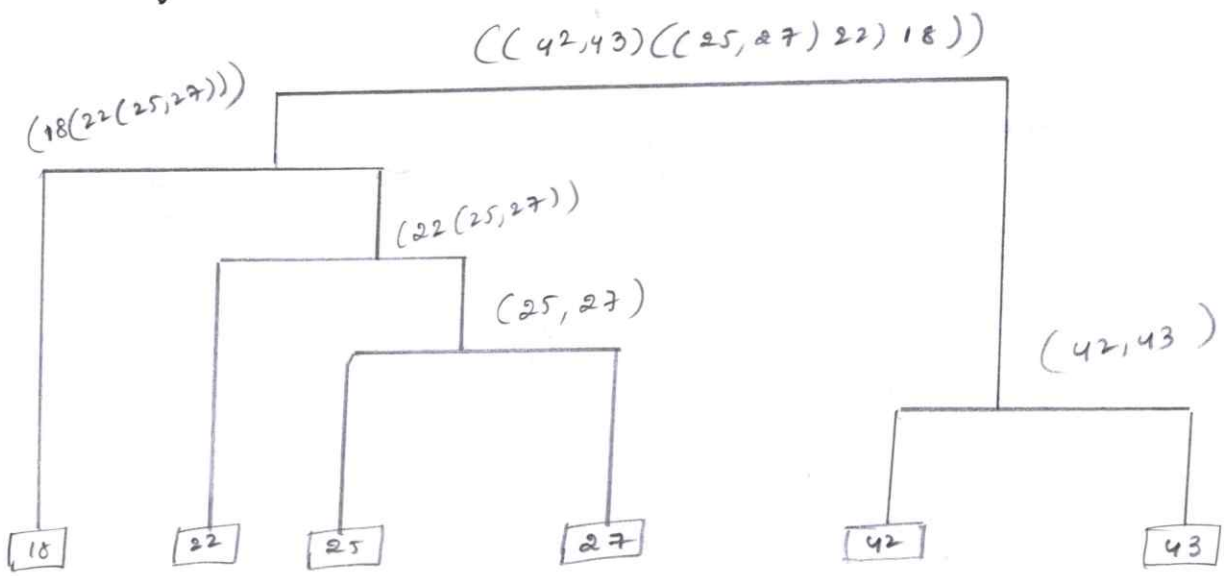
18 & 22, 25, 27 have the min dist ((27, 25) 22) 18)

step-6:

	18, 22, 25, 27, 42, 43
18, 22, 25, 27, 42, 43	0

step-7:

Dendogram:



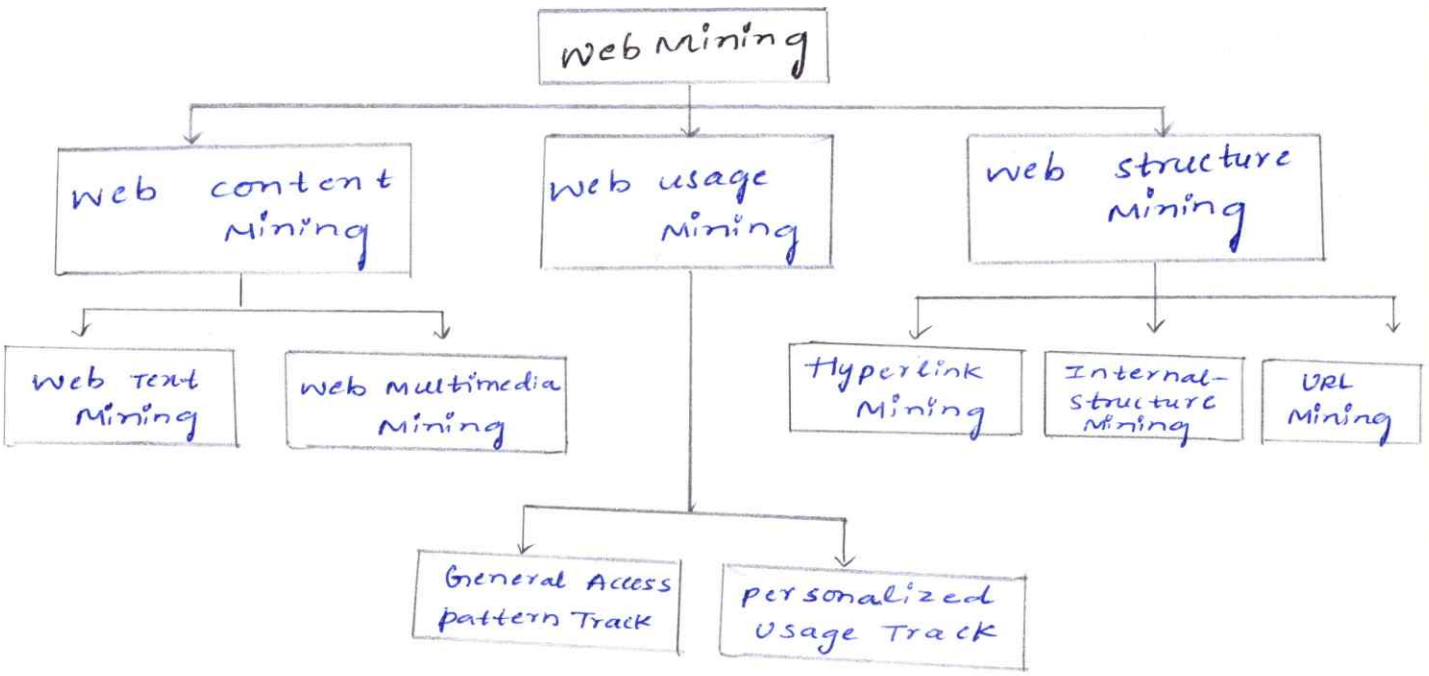
Single linkage agglomerative clustering forms clusters by repeatedly merging the two groups with the smallest inter-point distance, often resulting in long chain-like clusters.

(10) Explain the architecture and process of web Mining with suitable examples.

A

Web Mining:

- Web mining is the process of discovering useful information and patterns from the world wide web.
- It helps organizations understand how people use the internet and how web data can be analyzed to improve service.
- In simple words web mining = Data Mining + the web.



Types of web Mining:

1) Web content mining
2) web structure mining
3) web usage mining

① Web content Mining:

Extracting useful information from web pages, such as text, images, and multimedia content.

② Web structure mining:

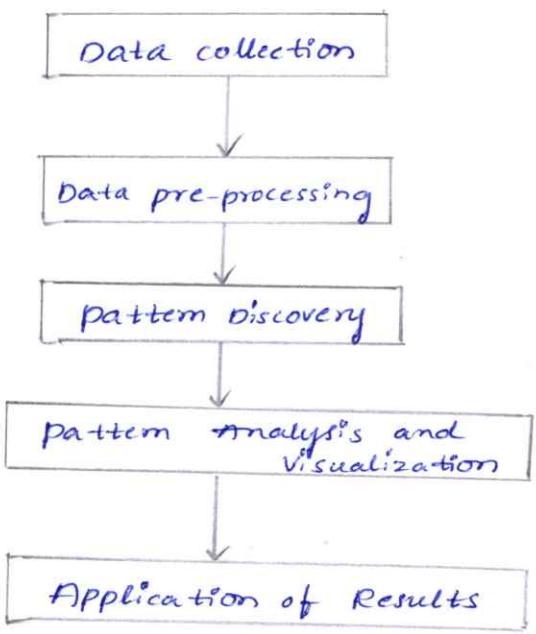
Analyzing the link structure of the web to identify patterns and relationships.

③ Web usage Mining:

Analyzing user behavior and interactions with websites.

Web usage mining is the application of identifying or discovering interesting usage patterns from large data sets.

Architecture:



Step-1 : Data collection :

Data is gathered from :

- Web servers
- Web pages
- Browser interaction
- cookies and clickstreams

Step-2 : Data preprocessing

collected data is cleaned and formatted :

- Removing noise, bots, duplicates
- session identification

Step-3 : Pattern Discovery

Applying mining techniques such as :

- clustering, classification, Association Rules, Link Analysis.

Step-4 : pattern Analysis & Visualization

Meaningful patterns are extracted and visualized using :

- Dashboards
- Graphs
- Visualization

Step-5 : Application of Results

Results are applied in Recommendation Systems, website optimization, Fraud detection.

Ex : E-commerce website.

(11) Illustrate the challenges and solutions in handling unstructured text during Text Mining.

A Challenges:

1) Noise in Text

* Unstructed text contains spelling mistakes, stopwords, emojis, advertisements, HTML tags, and irrelevant information.

This noise reduces the quality of mining results.

2) Ambiguity of words

* Words may have multiple meanings, or different words may have the same meaning.

ex: bank (river bank / financial bank).

3) High Dimensionality

* Text data produces thousands of unique terms, creating very large feature spaces.
* This increases computational time and reduces algorithm efficiency.

4) Sparsity of Data

Most documents contain only a few words from a huge vocabulary.

So the document-term matrix becomes extremely sparse, making analysis complex.

5) Informal and Inconsistent Text

User-generated content contains slang, abbreviations, mixed languages, short forms (like "u", "gr8"), making understanding and classification difficult.

Solutions for Handling Unstructured Text

1) NLP-Based preprocessing

Natural Language processing techniques help clean and prepare text.

- Tokenization
- stopword removal
- stemming & Lemmatization

These reduce noise and ambiguity.

2) Text Normalization

converts text into a consistent format using:

- Lowercasing
- spelling correction
- Removing punctuation, emojis, HTML tags

This improves text quality.

3) Feature Extraction Techniques

Transforms text into numerical features for mining:

- Bag of words
- TF-IDF

These reduce dimensionality and capture meaningful patterns.

4) Dimensionality Reduction

Techniques like PCA, LSA and SVD help reduce high dimensional text data, removing redundant terms and improving performance.

5) semantic processing

Advanced NLP techniques such as:

- Named Entity Recognition (NER)
- part-of-speech (pos)
- Topic Modeling

These help understand content, meaning, and relationships between words.

Handling unstructured text is challenging due to noise, ambiguity, and high dimensionality, but NLP preprocessing, normalization, and feature extraction methods make text mining accurate efficient.

Examples of unstructured text Data:

1) Text Documents
2) Emails
3) Images
4) Audio Files
5) Video Files
6) Log Files
7) sensor Data
8) social Media posts

Subject: Data Mining
III B.Tech I Semester End Examinations Regular-December 2025

Part-A

Scheme of Evaluation for 1 Marks Questions:

S.No	Question	Expected Answer (Key Points for 1 Mark)	Marks
1.a	**What is Discretization?**	Converting continuous data into discrete intervals or categories.	1
b	**What is Data Cleaning?**	Process of detecting and correcting errors, missing values, or inconsistencies in data.	1
c	**Define frequent itemset.**	A set of items that appear together frequently in transactions or datasets.	1
d	**Define Association Rule.**	A rule that shows relationships or correlations between itemsets in transactional data.	1
e	**What is a Decision Tree?**	A tree-structured model used for classification or prediction based on decision rules.	1
f	**Define Classification.**	Assigning data items to predefined classes or categories.	1
g	**Differentiate between Classification and Clustering.**	Classification: supervised learning; Clustering: unsupervised grouping of data.	1
h	**What is Outlier Detection?**	Identifying unusual data points that deviate significantly from the rest.	1
i	**List any two types of Web Mining.**	Web Content Mining, Web Structure Mining, Web Usage Mining (any two).	1
j	**What is Episode Rule Discovery in Text Mining?**	Identifying frequently occurring sequential patterns or event sequences in text.	1

Part-B

Scheme of Evaluation for 10 Marks Questions:

S.No	Theory	Marks Split-up	Total
2	**Explain the KDD process with a neat diagram.**	• Definition of KDD – 2 • Steps (Selection, Preprocessing, Transformation, Data Mining, Interpretation) – 6 • Diagram – 2	10
3	**Explain Data Preprocessing techniques with suitable examples.**	• Need for Preprocessing – 2 • Techniques (Cleaning, Integration, Reduction, Transformation) – 6 • Examples – 2	10
4	**Explain the APRIORI algorithm with an example.**	• Need for frequent pattern mining – 1 • Steps of Apriori algorithm – 3 • Example (candidate generation, pruning, support counts) – 1	5
5	**Explain in detail about Closed Frequent Itemset with an example.**	• Definition – 2 • Properties – 3 • Example – 5	10
6	**Compare and Contrast KNN and Decision Tree classifiers.**	• Explanation of KNN – 3 • Explanation of Decision Tree – 3 • Comparison table/differences – 4	10
7	**Discuss various attribute selection measures used in Decision Tree Construction.**	• Need for attribute selection – 2 • Information Gain – 3 • Gain Ratio – 3 • Gini Index – 2	10
8	**Explain the K-Means clustering algorithm with a suitable example.**	• Purpose of K-Means – 2	10

		• Algorithm steps – 5 • Example – 3	
9	**Explain the steps involved in Agglomerative Hierarchical Clustering.**	• Introduction – 2 • Steps (initial clusters, distance calculation, merging, dendrogram) – 6 • Example/diagram – 2	**10**
10	**Explain the architecture and process of Web Mining with suitable examples.**	• Definition & components – 2 • Types (Content, Structure, Usage) – 5 • Architecture/process flow – 3	**10**
11	**Illustrate the challenges and solutions in handling Unstructured Text during Text Mining.**	• Challenges (noise, ambiguity, high dimensionality, sparsity) – 5 • Solutions (NLP, preprocessing, feature extraction, normalization) – 5	**10**