unit-3

Unsupervised Learning

que- Explain Types of unsupervised Learning and challenges in unsupervised Learning.

⇒

Types of unsupervised Learning :-

- There are mainly 3 types of Algorithms which are used for unsupervised dataset.
  1) clustering
  2) Association Rule Learning
  3) Dimensionality Reduction.

1) clustering :-

- clustering in unsupervised machine Learning is the process of grouping unlabeled data into clusters based on their similarities.
- The goal of clustering is to identify patterns and relationships in the data without any prior knowledge of the data's meaning.
- These algorithms are used to process raw, unclassified data objects into groups.
- Some common clustering algorithms :-
  1) K-means clustering
  2) Hierarchical clustering
  3) mean-shift clustering
  4) spectral clustering.

2) Association Rule Learning :-

- Association rule Learning is also known as association rule mining is a common technique used to discover associations in unspervised machine Learning.
- This technique is a rule-based ML technique tha

finds out some very useful relations between parameters of a large data set.
-This technique is basically used for market basket analysis that helps to better understand the relationship between different products.
-Algorithms :-
1] Apriori Algorithm
2] FP-Growth Algorithm
3] Eclat Algorithm
4] Efficient Tree-based Algorithms.

3) Dimensionality Reduction :-
-Dimensionality reduction is the process of reducing the number of features in a dataset while preserving as much information as possible.
-This technique is useful for improving the performance of machine learning algorithms and for data visualization.
-Algorithms :-
1] Principal Component Analysis (PCA)
2] Linear Discriminant Analysis (LDA)
3] Locally Linear Embedding (LLE).

• Challenges of unsupervised Learning :-

1) Evaluation :- Assessing the performance of unsupervised learning algorithm is difficult without predefined labeles or categories.
2) Interpretability :- understanding the decision-making process of unsupervised learning models

is often challenging.

3) Overfitting :- Unsupervised learning algorithms can overfit to the specific dataset used for training, limiting their ability to generalize to new data.

4) Data quality :- unsupervised learning algorithm are sensitive to the quality of the input data. -Noisy or incomplete data can lead to misleading or inaccurate results.

Que- Explain Preprocessing and scaling. Also types of Preprocessing.

⇒

Preprocessing :-
-Preprocessing is a critical step in the data pipeline of machine learning models.
-It involves transforming raw data into a format suitable for modeling, which helps in enhancing the accuracy, efficiency and effectiveness of the learning algorithm.

Why Preprocessing?
-Handling Missing values :- Filling in or removing missing data points.
-Data Normalization and scaling :- Ensuring data features have a similar scale.
-Feature Encoding :- Converting categorical data into numerical format.
-Outlier Detection :- Identifying and handling data

Points that significantly differ from other observations.
- Feature selection and Extraction :- Reducing dimensionality by selecting relevant features or creating new ones.

- Scaling :-
- Scaling is specifically about adjusting the range of independent variable or features of data.
- This step is crucial for algorithms sensitive to the magnitude of data, like gradient descent, neural networks and clustering algorithms.

- Standardization (Z-Score Normalization) :-
- Scales data to have a mean of zero and standard deviation of one.

Formula :- $x' = \frac{x - \mu}{\sigma}$

- useful for algorithms assuming normally distributed data e.g. linear regression, logistic regression.

- Min-max scaling (Normalization)
- Scales data to fit within a specified range usually [0, 1].

Formula :- $x' = \frac{x - x_{min}}{x_{max} - x_{min}}$

- useful when the algorithm does not assume any specific distribution of data.

• Types of Preprocessing Techniques :-

1) Data cleaning :-
• Handing missing Data :- Methods include removal, mean / mode / median or using more advanced imputation techniques like k-nearest neighbors.

• outlier Detection :- Techniques like z-score, IQR method or using algorithms like isolation Forest.

2) Data Transformation :-
- Transformation modifies data into a format that enhances the performance of machine Learning.
  - Normalization
  - standardization
  - Log Transformation

3) Feature Encoding :-
- Feature encoding converts categorical data into a numerical format that can be used by machine Learning.
  - Label Encoding
  - ordinal Encoding
  - Binary Encoding.

4) Feature selection and Extraction :-
- This type of preprocessing focuses on reducing the dimensionality of data to improve model Performance.

5) Dimensionality Reduction:-
- Techniques that reduce the number of Feature in the dataset while retaining the most important information.
- Linear Discriminant Analysis (LDA)
- Factor Analysis.

6) Feature Scaling:-
- scaling adjusts the magnitude of Features to a comparable scale.
- Standard Scaling
- min-max scaling
- Robust scaling.

Que- Explain k-means clustering with example.
⇒

k-means clustering:-
- K-means clustering algorithm, which groups the unlabeled dataset into different clusters.
- The goal of clustering is to divide the population or set of data points into a number of groups so that the data points within each group are more comparable to one another and different from the data points within the other groups.
- It is essentially a grouping of things based on how similar and different they are to one another.

K-means clustering Working :-

1] First, we randomly initialize k points, called means or cluster centroids.
2] we categorize each item to its closest mean, and we update the mean's coordinates, which are the averages of the items categorized in that cluster so far.
3] We repeat the process for a given number of iterations and at the end; we have our clusters.

- K-means clustering is an unsupervised machine learning algorithm used for partitioning a dataset into k distinct, non-overlapping clusters based on feature similarity.
- It aims to minimize the variance within each cluster while maximizing the variance between clusters.

• Example :-
DataPoints :-

A_1	2	10
A_2	2	5
A_3	8	4
B_1	5	8
B_2	7	5
B_3	6	4
C_1	1	2
C_2	4	9

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

1) $C_1 = 2, 10$ (~~x_2 y_2~~)
$C_2 = 5, 8$
$C_3 = 1, 2$
} centroids

Distance to				classes		
2	10	5	8	1	2	C_1	C_2	C_3
0		3.61		8.06		1		
5		4.24		3.16		3		
8.49		5		7.28		2		
3.61		0		7.21		2		
7.07		3.61		6.71		2		
7.21		4.12		5.39		2		
8.06		7.21		0		3		
2.24		1.41		7.62		2		

2) $C_1 = (2, 10)$
$C_2 = (6, 6)$
$C_3 = (1.5, 3.5)$
} By mean

Distance to						classes		
2	10	6	6	1.5	3.5	C_1	C_2	C_3
0		5.66		6.52		1	1	
5		4.12		1.58		3	3	
8.49		2.95		6.52		2	2	
3.61		2.13		5.70		2	2	
7.07		1.56		5.70		2	2	
7.21		2.83		4.53		2	2	
8.06		6.39		1.58		3	3	
2.24		3.54		6.04		2	1	

3] $C_1 = (3, 9.5)$
$C_2 = (6.5, 5.25)$
$C_3 = (1.5, 3.5)$
} By mean

Distance to						classes		
3	9.5	6.5	5.25	1.5	3.5	C_1	C_2	C_3
1.12		6.54		6.52		1	1	1
4.51		4.51		1.58		3	3	3
7.43		1.95		6.52		2	2	2
2.50		3.13		5.70		2	2	1
0.02		0.56		5.70		2	2	2
0.26		1.35		4.53		2	2	2
7.76		6.39		1.58		3	3	3
1.124		4.51		6.04		2	1	1

Que- Explain Agglomerative clustering with example.

=>

Agglomerative clustering :-

- Agglomerative clustering is a type of hierarchical clustering technique used to group data points into clusters based on their similarities.
- It's a "bottom-up" approach, where each data points starts as it own cluster, and pairs of clusters are merged as one moves up the hierarchy.

Hierarchical Approach :-

- Agglomerative clustering starts with each data Point as an individual cluster.
- In each step, the closest two clusters are merged.
- This process continues until all points are merged into a single cluster or until a stopping criterion is met. (e.g. desired number of clusters)

Example :-

18, 22, 25, 27, 42, 43

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

43 to 42 = min distance = 1 (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

25 to 27 = min distance = 2 (25,27)

step 3 :-

	18	22	25,27	42,43
18	0	4	7	24
22	4	0	3	20
25,27	7	3	0	17
42,43	24	20	17	0

25,27 to 22 = min distance = 3 (22,25,27)

step 4 :-

	18	22,25,27	42,43
18	0	4	24
22,25,27	4	0	20
42,43	24	20	0

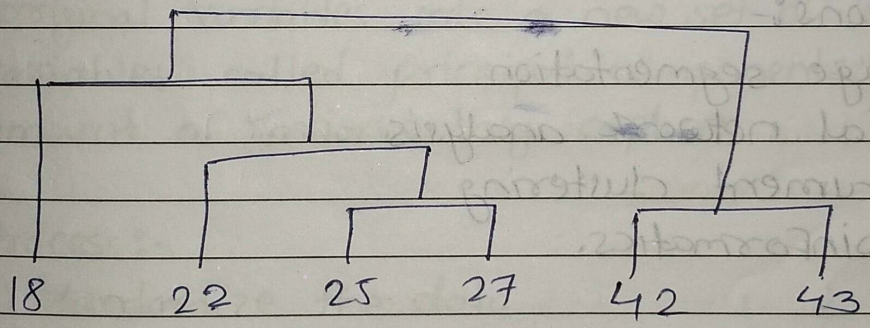
22,25,27 to 18 = min distane = 4
(18,22,25,27)

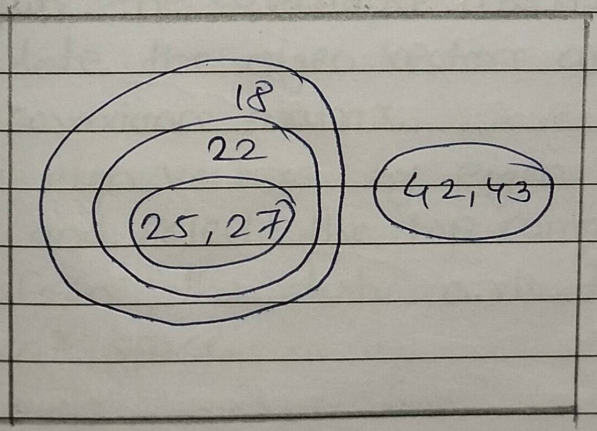
step 5 :-

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

42,43 to 18,22,25,27 = min distance = 24
(18,22,25,27,42,43)

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





• dendogram :-
- dendogram is a type of tree diagram showing hierachical relationship between set of data.

- just by looking at dendogram we tell that how the cluster is formed.

Advantages :-
1) Interpretability
2) Flexibility
3) No need for the number of clusters in advance.

Disadvantages :-
1) Computationally Expensive
2) sensitive to noise and outliers
3) Irreversible merges.

Applications :-
1) Image segmentation
2) social network analysis
3) Document clustering
4) Bioinformatics.

ue- Explain Dimensionality Reduction and Feature Extraction.

⇒

Dimensionality Reduction:-
- Dimensionality reduction is the process of reducing the number of input variables in a dataset.
- This helps in addressing issues like overfitting, reducing computational cost and improving visualization for high-dimensional data.

• methods of Dimensionality Reduction:-

1] Principal component Analysis (PCA):-
- PCA is a linear technique that transforms the original variables into a new set of uncorrelated variables called principal components, ordered by the amount of variance they capture.

Process:-
1) standardize the data
2) compute the covariance matrix of the data
3) calculate the eigen vectors and eigen values of the covariance matrix.
4) sort eign vectors by eigen values in descending order and select the top components.
5) Transform the data onto the new principal component space.

• Advantages of Dimensionality Reduction :-
1) Reduces overfitting
2) Improves model performance
3) Enhances visualization

• Disadvantages :-
1) Information Loss
2) Interpretability.

• Feature Extraction :-
- Feature Extraction involves transforming the data into a set of new features that better capture the underlying structure or pattern of the data.
- It's type of dimensionality reduction that focuses on creating new, more informative features from the original ones.

Key Techniques of Feature Extraction :-

1) Manual Feature Extraction :-
- Involves manually selecting or creating new features based on domain knowledge.

2) Principal component Analysis (PCA) :-
- Not only reduces dimensionality but also extracts features that capture the most variance.

3) Independent component Analysis (ICA) :-
- ICA focuses on finding components that are statistically independent, unlike PCA, which

Focuses on Variance.

- Advantages :-
1) Enhances Model Accuracy
2) Reduces Noise
3) Automates complex Feature Engineering.

- Disadvantages :-
1) complexity
2) Dependence on Algorithms.

FaceLook
D D M M Y Y Y Y

que- Explain Dimentionality reduction. define PCA, NMF, t-SNF

⇒

Dimentionality Reduction:-
-Dimensionality reduction is a process used in machine learning and data analysis to reduce the number of input variables or features while retaining as much of the essential information as possible.

1) PCA (Principal component Analysis):-
-PCA is a linear dimensionality reduction technique that transforms a dataset into a set of orthogonal variables, called Principal components.
-The principal components are ordered such that the first few retain most of the variation present in the original dataset.

How it works:-
-PCA identifies directions where data variance is maximal and projects the data along these directions.
-It is often used when reducing features in continuous data.

Applications:- Feature extraction, image compression, noise reduction.

2] NMF (Non-negative matrix Factorization):-
- NMF is a matrix Factorization technique that reduces dimensionality by factorizing the original data matrix into two lower-dimensional matrices with the constraint that both metrices are non-negative.
- This is particularly useful when data are non-negative, such as images or document-term matrices.

How it works:-
- It decomposes a matrix into two matrices whose product approximates the original matrix, often used for applications that involve parts-based representations.

Applications:- Text mining, topic modeling, image processing.

3] t-SNE (t-Distributed stochastic Neighbor Embedding):-
- t-SNE is a non-linear dimensionality reduction technique mainly used for visualizing high-dimensional data.
- It aims to minimize the divergence between two distributions:- one representing pairwise similarities in high-dimensional space and one in low-dimensional space.

How it works:-
- t-SNE places similar data points closer together in lower-dimensional space, preserving

local structures of the data, which is useful for visualizing complex, high-dimensional relationship.

Applications :- Data visualization, clustering, image recognition,