NAME - Himanshu.

Roll no. - 23/4017

Exam Rollno. - 23019582065.

SEM - 5th.

YEAR - 3rd.

Course - B.Sc. Computer Science.

Qs-1 (i) Task: Prediction / Forecasting (Regression).

→ Because we are estimating a numerical value (Sales)(volume) for the future.

~~(ii) Identifying~~

(ii) Task: Association Rule Mining

→ classic Example of Market Basket Analysis.

(iii) Task: Clustering

→ Because customers are grouped into natural segments without predefined labels.

(iv) Task: classification / Anomaly Detection.

→ Fraud detection = classification (Fraud v/s. not fraud).

→ In Some cases, also treated as anomaly detection if fraud is rare.

(v) Task: Prediction / Forecastion (Regression / Time Series).

→ Since rainfall is a continuous variabale predict ~~ov~~ over time.

(vi) Task: Classification.

→ Because the goal is to predict a categorical label (disease: Yes/No).

Qs.2 (i) Ordinal (Since there is a natural ranking : Bronze < Silver < Gold).

(ii) Ratio (count data, true zero and ratios like "twice as many patients" make sense).

(iii) Interval (differences between dates are meaningful but there is no true zero point in the calander).

(iv) Nominal (categories with no ordering! Male, female, Other).

(v) Ordinal (They imply ranking, but the gap between A and B ≠ gap between C and D).

(vi) Nominal (categories with no inherent order: red, blue, black, etc.).

Ans-3 # Definition/Meaning

- Noise → Random error or meaningless data that does not carry useful information.
- Outliner → A data object that deviates significantly from the overall pattern, but may still carry important information.

Examples → Noise Example.

- while measuring people's heights suppose a faulty sensor records one height as -20 cm.

→ This is noise (impossible and meaningless value).

→ Outliner Example

- In the Same dataset, most people are 150 - 190 cm tall, but one person is 220cm.

→ This is an outliner (unusual, but possible and interesting - maybe a professional basketball player).

Ans-4→ Discretization :- Converting continous attributes into discrete categories (intervals / bins).

- Example : Age 0-18 = Young, 19-40 = Adult, 41+ = Senior

→ Binarization : Converting attributes into binary (0/1) values.

- Example : Age ≥ 18 → 1 else 0; Car color Red/Blue/Black → one-hot encoding.

So, Discretization = many categories; Binarization = only 0/1.

Ans-5 1. Filter Methods.

→ Statical Measures.

→ Fast and Simple.

→ Independent of Model.

→ Example :- Correlation, chi-square test.

2. Wrapper Methods

→ Uses Model Performances.

→ More accurate but slow

→ Example: Forward selection, Backward elemination Recursive feature elemination (RFE)

Qs-6 # Scalability.

- Defination : The ability of a data mining algorithm to handle large volumes of data efficiently (in terms of time and memory)

→ Challenge: ~~Asss~~ As data grows (terabytes, petabytes), argorithms may become too slow, memary-intensive or computationally infeasible.

Heterogeneity.

- Defination: Refers to the presence of different types of data (structured, unstructured, images, text, video, categorical, numerical).

→ Challenge: Difficult to integrate, preprocess, and analyze such diverse dato formats together.

Qs-7 Normlization: The process of Scalling numeric data into a specific range (commonly [0,1]).

Formula (Min-Max Normalization)

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}.$$

where,
- X = original value,
- X_{max} = maximum value.
- X_{min} = minimum value,

Given Data

Ages = $\{18, 22, 21, 25\}$.

• $X_{min} = 18$, • $X_{max} = 25$.

normilization:-

1. For 18:

$(18-18)/(25-18) = 0/7 = 0$

2. for 22:

$(22-18)/7 = 4/7 \approx 0.571$.

3. for 21:

$(21-18)/7 = 3/7 \approx 0.429$.

4. for 25:

$(25-18)/7 = 7/7 = 1$.

final normalized ages:

$\{0, 0.571, 0.429, 1\}$.

Q-8 Significance of Dimensionality Reduction.

→ Dimensionality reduction - process of reducing the number of input values (feature) while preserving important information.

(i) Removes noise & redundancy

(ii) Avoids overfitting.

(iii) Improves visualization.

Curse of Dimensionality.

(i) As the number of dimensions (feature) increases, data becomes sparse and distance measures lose meaning.

(ii) Algorithms that rely on distance / similarity (k-NN, clustering) becomes less effective.

Ans-9 Sampling is the process of selecting a small subset of data from a large dataset.

- why it's useful.

(i) Reduces computation cost and memory usage.

(ii) Makes algorithms run faster on large datasets.

Sampling Methods :-

(i) Simple Random Sampling (SRS):

→ Each data item has an equal chance of being selctd

→ Ensures unbaised representation of the dataset

(ii) Stratified sampling:

→ Data is divided into groups (strata) based on some attribute, and then samples are drawn from each group.

→ Ensures all groups are fairly represented.

Ans10 1. Supervised Learning (Techniques).

- Defination: Data mining techniques where the model is trained on a dataset with input features+ known output (labels).

- Goal: Learn a mapping from input → output, then predict labels from new data.

- Example: • classification (e.g. predicting spam vs. non-spam emails)

2. UnSupervised Learning (Techniques).

- Defination: Data mining techniques where the dataset has only input features, no labels.
- Goal: Discover hidden patterns, groups, or associations inthe data.
- Example.
  - clustering (eg. grouping customers by buying behaviour).