Internal Examination 2021-22

classmate
Date
Page

Question Paper No. 1

Q.1) A. Choose the correct option

1. b. Selection and Interpretation
2. a. Data cleaning
3. a. Online Analytical Processing
4. a. Rule-based Classification
5. a. Discrete Data.

Q.1) B) 1) Star schema: A database schema design used in data warehousing that organizes data into fact tables & dimension tables, resembling a star shape.

2) Information Gain: A metric used in decision trees to measure the reduction in entropy or uncertainty after splitting a dataset based on an attribute.

3) Regression: A statistical technique for modeling the relationship between a dependent variable and one or more independent variables to predict continuous outcomes.

4) Classification: A supervised learning task that involves assigning datapoints

to predefined categories or classes based on their features.

5) Data cleaning: The process of identifying & correcting inaccuracies, inconsistencies and errors in dataset to improve its quality for analysis.

Q.2) Solve any five Questions

1) List the characteristics of OLAP system.

→ Multidimensional Analysis: Supports analysis of data from multiple perspectives.
- Fast Query Performance: Optimized for quick retrieval of complex queries
- Aggregation: Allows summarization of data at various levels of granularity
- Data Slicing & Dicing: Enables viewing data from different angles & dimensions
- Drill-Down & Roll-Up: Facilities detailed exploration or summary of data.

3	Focuses on extracting exact, predefined information	Focuses on finding hidden patterns & relationships
4	Generally involves simple & direct queries	Involves complex analysis & model building.

4) List out steps of KDD process

→ - Data Selection: choosing relevant data from various sources
- Data Cleaning: Addressing missing, inconsistent or noisy data.
- Data Transformation: Converting data into a suitable format or structure for analysis
- Data Mining: Applying algorithms to discover patterns or models
- Patterns Evaluation: Assessing the discovered patterns for validity & usefulness.
- Knowledge Presentation: Presenting the patterns & insights in a comprehensible manner

5) What are the advantages of using decision tree?

2) Discuss about the major issues in data mining.

→ - Data Quality: Handling missing, noisy, or inconsistent data.
- Data Privacy: Ensuring the protection of sensitive and personal info.
- Scalability: Managing & processing large volumes of data efficiently
- Model Interpretability: Understanding & explaining complex models & their predictions
- Overfitting: Avoiding models that perform well on training data but poorly on unseen data.

3)

	Query Processing	Data Mining
1	Retrives specific data based on structured queries	Discovers patterns & insights from large datasets
2	Uses SQL or similar query languages for data retrival	Employs algorithm like clustering classification & accociation rules

5) Transportation: Route optimization, predictive maintence & traffic management.
6) Social Media: Content recommendation, trend analysis, & user sentiment analysis
7) Education: Personalized learning, student performance prediction & curriculum development.

→ b) K-Nearest Neighbour (KNN) Classifier:
The K-Nearest Neighbour classifier is simple, instance-based learning algorithm used for classification task. It works by finding the 'K' closest training examples in the feature space to a new data point & assigning the class label based on the majority vote of these neighbours.
key aspects of KNN:
- Distance Metric: Commanly used Euclidean distance but can use other metrics.
- K Value: Determines the number of neighbours to consider; a smaller K can be sensitive to noise, while a larger K can smooth out class boundaries.

- Lazy learning : KNN doesn't build a model during training but makes decision based on the entire dataset during prediction.

	BI	Data Science.
	Analyzes historical data	Analyzes & predicts future outcomes.
	focuses on reporting & visualization	Focuses on modeling & algorithms
3	Uses structured data sources	Handles structured & unstructured data
4	Primarily descriptive analytics.	Uses predictive and prescriptive analytics.
5	Tools: Power BI, Tableau, SQL,	Tools: Python, R, machine learning libraries.
6	Geared towards business decision making	Geared towards discovery and innovation.

→ - Easy to understand: Simple and intuitive representation of decision rules
- No Data Preprocessing Required: Handles both numerical & categorical data without extensive preprocessing.
- Visual Representation: Provides a clear graphical depiction of decision making processes
- Versatile: Can be used for both classification & regression tasks.
- Handles Missing Values: Can manage incompletes data effectively.

Q.3) Solve any five Questions.

1) Application of Data Science.

→ 1) Healthcare: Predictive analytics for patient outcomes, personalized medicine, disease detection & drug discovery
2) Finance: Fraud detection, risk management, algorithmic trading, & customer segmentation.
3) Retail: Recommendation systems, inventory management, customer behavior analysis & pricing optimized
4) Marketing: Targeted advertising, sentiment analysis, campaign effectiveness & market research.

and reporting.

- Historical Data Analysis:
Stores historical data for trend analysis and long-term insights.
- Enhanced Data Quality:
Implements data cleansing & validation processes.
- Better Decision-Making:
Supports strategic decision making with comprehensive & consistent data
- Data Integration:
Combines data from various sources into a unified view.

Q.4) Solve any five Questions.

① A) Data cleaning:
Data cleaning refers to the process of identifying and correcting errors or inconsistencies in a dataset to improve its quality. This step is crucial in preparing data for analysis, as poor quality data can lead to misleading insights and inaccurate results. Methods of data cleaning include:

Major difference between star schema & snowflake Schema

Star Schema: A simple database schema used in data warehousing where a central fact table is connected to multiple dimension tables. The dimension tables are typically denormalized, leading to faster query performance but potentially more redundancy.

Snowflake Schema: A more complex schema where dimension tables are normalized into multiple related tables, resulting in a "snowflake" shape. This reduces redundancy and improves data integrity but can lead to more complex queries and slower performance.

Benefits of Building an Enterprise Data Warehouse:

- Centralized Data Storage:
  Provides a single source of truth for organizational data
- Improve Query Performance:
  Optimized for complex queries

1) Removing Duplicates: Identifying & deleting duplicate records to ensure each entry is unique.
2) Handling Missing Value: Techniques like imputation or removing rows/columns with missing data.
3) Standardizing Data: Ensuring consistency in data formats and units, such as converting dates to a standard format or unifying text case.
4) Correcting Errors: Identifying & fixing inaccuracies, such as type or incorrect entries.
5) Data Transformation: Normalizing or scaling data to bring it into a common range or format.

Machine learning:
Machine learning is a subset of Artificial intelligence that involves training algorithm to learn patterns from data & make predictions or decisions without being explicitly programmed. It encompasses:
1) Supervised learning: Algorithms learn from labled training data to

predict outcomes for new data.
2) Unsupervised learning: Algorithms identify patterns or grouping in unlabled data.
3) Reinforcement learning: Algorithms learn by interacting with an environment & receiving feedback in the form of rewards or penalties.
4) Model Evaluation: Techniques like cross validation & performance metrics assess ~~no~~ how well model is generalize to new data.

③ d)	Data Aanalytics	Data Aanalysis
	It ~~is~~ extracts actionable insights	It Examines data for understanding
2	Predicts future trends	Describe past events
3	Uses advanced algorithms	Uses statistical methods.
4	Focuses on decision making	Focuses on data interpretation

5	Prescriptive and predictive	Descriptive and diagnostic
6	Handles large datasets	Handles structured often smaller datasets
7	Involves complex tools (AI/ML)	Involves simpler tools
8	Often real-time data	Typically historical data.
9	Business & operational focus	Research & reporting focus

(5)

(4) E) Explain Multidimensional View of Data Cube -

→ Multidimensional view:

- Organizes data into multiple dimensions
- Allows analysis from various perspectives
- Dimensions can be attributes like time, location or product
- Facilitates deeper data exploration
- Helps uncover patterns &

trends
- Useful in complex data queries

Data Cube:
- A multidimensional array structures for data storage
- Used in OLAP system
- Each dimension represents a variable
- Cells store aggregated values
- Enables fast and complex data queries
- Simplifies the analysis of large datasets.

(5) G) Explain various steps in data pre-processing.

→ Data pre-processing involves preparing & cleaning data before analysis. Steps include:

1) Data Collection:
- Gathering data from various sources

2) Data Integration:
- Combining data from different sources into a cohesive dataset

3) Data Cleaning:
- Identifying and correcting errors or inconsistencies

Regression: A statistical method used to predict a continuous outcome variable based on one or more predictor variables.

Supervised Learning: A type of machine learning where the model is trained on labeled data to predict outcomes for new, unseen data

Classifier: An algorithm that categorizes data into predefined classes or labels.

Association Rule: A rule-based method in data mining used to find relationships or patterns between variables in large datasets.

Solve any five Questions.
Challenges of data science technology:
1) Data Quality & Availability:
Inconsistent, incomplete or unstructured data can affect the accuracy and reliability of data science models.

4) Data Transformation :
- Converting data into a suitable format or structure.
5) Data Reduction : Reducing data volume through techniques like feature selection or dimensionality reduction.
6) Data Splitting :
Dividing data into training, validation & test sets for model development & evaluation.

Internal Examination 2022-23
Question Paper No. 2.

1 A) Choose the correct answer.
1. b. Useful Information
2. c. Metadata
3. b. Snowflake schema
4. b. Squares
5. a. Itemset

B) Define the following terms
1) Clustering : Grouping a set of data points into clusters where objects in the same cluster are more similar to each other than to those in other clusters.

or existing data to forecast future trends or outcomes. It employs machine learning models such as regression, time series forecasting, or classification techniques to anticipate future events such as predicting customer churn, stock prices, product demand, or disease outbreaks.

d) Classification categorization algorithm
1. Decision Trees: A tree-like structure used for decision-making by splitting data into branches based on feature values.
2. Support Vector Machines (SVM): Finds the optimal boundary (hyperplane) that best separates different classes in the data.
3. Naive Bayes: A probabilistic algorithm based on Baye's Theorem, assuming independence between features for class prediction.

e) Features of data warehouse:
1. Subject-Oriented: Organized around key business areas (e.g. sales,

finance) for analysis.
2. Integrated: Combines data from multiple sources into a unified format.
3. Time-Variant: Stores historical data to track changes over time for trend analysis.
4. Non-Volatile: Data is stable & does not change once entered, ensuring consistent reporting.

(Q.3) Solve any five Questions

(c) Bayesian Classification.

1. Baye's Theorem: A probabilistic approach that calculates the likelihood of a class given input features using the formula: $P(C|x) = P[(x|C) * P(C)]/P(x)$, where $P(C|x)$ is the posterior probability of the class.
2. Naive Bayes Assumption: Assumes that features are independent given the class, simplifying computations and making it efficient for large datasets.
3. Advantages: Handels both binary & multi-class problems, works well with small datasets & is computationally

2) Scalability & Computational Power: Handling larger datasets & performing complex computations require significant computational resources & efficient algorithms.
3) Privacy & Ethical Concerns: Protecting sensitive data, ensuring ethical use of data, & complying with regulations like GDPR pose significant challenges.

b) Application of clustering:
1. Customer Segmentation: Grouping customers based on behavior, preferences, or demographics for targeted marketing
2. Anomaly Detection: Identifying outliers or unusual patterns in data, useful in fraud detection or networks security.
3. Image & Document Classification: Grouping similar images or documents for efficient organization, retrieval and processing.

c) Prediction:
Prediction involves using historical

5	Typically more computationally efficient with fewer clusters	Can be more computationally intensive
6	e.g.: single-linkage, complete-linkage	e.g: top-down recursive partitioning.

f) The Star Schema is a data modeling technique used in data warehousing to organize data into a structure that simplifies querying & reporting. It consists of a central fact table connected to multiple dimension tables.

1. Central fact table: Contains quantitative data & foreign keys linking to dimension tables.
2. Dimension Tables: Store descriptive attributes related to the facts, enabling detailed data analysis
3. Star-like Structure: The fact table is at the center, and dimension tables surround it, resembling a star. This design simplifies queries & reporting.
4. Query Efficiency: The schema enhances performance by reducing complex

5) Spatial Data : Represents the physical location shape & arrangement of objects.
e.g : maps, satellite imagery, GPS data.
6) Textual data: Data consisting of text that requires natural language processing for analysis
e.g: customer reviews, social media posts.

e)	Agglomerative Clustering	Divisive Clustering
1	Starts with each data point as a separate cluster	Starts with all data points in a single cluster
2	Merges clusters iteratively	Splits clusters iteratively
3	Commonly used with hierarchical clustering	Less common but also hierarchical
4	Works bottom-up (builds up from individual points)	Works top-down (breaks down from a single cluster)

efficient even with high-dimensional data.

4. Application: Commonly used in spam detection, sentiment analysis, text classification & medical diagnosis due to its simplicity & effectiveness.

d) Types of data:

1. Structured data: Organized in row & columns, typically in relational databases.
e.g. Spreadsheets, SQL, databases.

2. Unstructured Data: Lacks a predefined format making it challenging to analyze.
e.g.: text documents, images, videos, social media posts.

3. Semi-Structured data: Data that is not organized in rigid format but has some identifiable structure
e.g. XML files, JSON files, NoSQL database

4. Time-Series data: Data that is collected or recorded at specific time intervals, used to identify trends & patterns over time.
e.g: stock market data, weather data, sensor readings.

8	Read / Update Access	Mostly Read
9	No data redundancy	Redundancy present
10	Database size 100 MB - 100 GB	Database size 100 GB - few TB

d) Different types of data used in cluster analysis

1. Numerical data: Continuous data types like integers and real numbers that can be used with distance-based methods such as k-means
2. Categorical Data: Discrete data with a limited set of values, typically used in algorithms like k-modes.
3. Binary Data: Data with only two possible values used in specific clustering techniques such as binary clustering.
4. Ordinal Data: Data that represents order or ranking but the intervals between values are not necessarily uniform.

joins & improving retrieval times, making it ideal for OLAP systems.

Q4. Solve any five Questions.

b)

	OLTP	Data Warehouse
	Application Oriented	Subject Oriented
1	Used to run business	Used to analyzed business
2	Detailed data	Summerized & refined data
3	Current up to date	History to current
4	Repetitive data	Ad-hoc access
5	Clerical User	Knowledge User
6	Performance Sensitive	Performance relaxed
7	Few records accessed at a time	Large volumes access at a time

5. Mixed-Type Data: Datasets that contain a combination of numerical, categorical & ordinal data, requiring specialized algorithms like k-prototypes to handle multiple types.

f) Decision tree classifier with an example:

A decision tree is a machine learning algorithm used for classification. It builds a tree-like model where each node represents a decision based on feature values, branches represent outcomes of those decisions, and leaves represent class lables. It splits the data into subsets to create a tree structure that helps in predicting the class of new instances by following the decision paths from the root to the leaves.

Example: Dataset

Weathers	Temperature	Buy icecream
Sunny	Hot	Yes
Sunny	mild	Yes
Rainy	Hot	No
Rainy	Mild	No

Decision tree:

1. Root Node : Split on Weather
   - If Weather is Sunny → Buy icecream (Yes)
   - If weather is Rainy → Go to next node
2. Next Node: Split on Temperature (for rainy)
   - If temperature is Hot → No
   - If temperature is Mild → No

g) Data Reduction and its techniques:

Data Reduction is the process of simplifying and reducing data volume while preserving its essential features.

Techniques for Data Reduction

1. Dimensionality Reduction: Reducing the number of features or dimension
2. Data Compression: Encoding the data in a more compact form.
3. Aggregation: Summarizing or consolidating data (e.g. calculating the average sales of a month rather than daily sales)
4. Numerosity Reduction: Reducing data volume through approximation.
5. Sampling : selecting a representative subset of the data instead of analyzing the entire dataset.