Internal Question Paper.

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.

Q1. B] 1] Star schema :- A database schema design used in data warehousing that organizes data into fact tables & dimesion tables resembling a star shape.

2] Information Gain :- A metrix used in decision trees to measure the reduction in entropy or uncertainity after splitting a dataset based on an attribute.

3] Regression :- A statistical technique for modeling the relationship between a dependend variable and one or more independent variables to predict continous outcomes.

Date: YOUVA

datapoint to predefined Categories or classes based on their features.

5) Data Cleaning :- The Processe of identifying & Correcting inaccuries, inconsistencies and errors in dataset to improve its quality for analysis.

Q2. Solve any five questions.

1) List the Characteristics of OLAP system.

→ - Multidimesional 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 & dimesion.

- Drill-Down & Roll-up : facilities detailed exploration or summary of data.

= i) Privacy issue
ii) Quality
iii) Scalability
iv)

Q P S 0

Sol 8

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 effectively.
- model Interpretability : understanding & explaning complex models & their predictions.
- overfitting : Avoiding models that perform well on training data but poorly on unseen data.

3] Query processing — Datamining.

1. Retrives specific data based on structural queries — Discovers patterns & insights from large datasets.

2. Uses SQL or similar query languages for data retrival. — Employs algorithms like clustering, classification & accociation rules.

SQL for — Clustering, Classification

3. Focuses on extracting exact predefined information.

3. Focuses on finding hidden patterns & relationships.

4. Generally involves simple & direct queries.

4. 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 : Accessing 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?

Date: YOUVA

→ - Easy to understand : simple and intative 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.
- versutile : Can be used for both classification & regression tasks.
- Handles missing values : Can manage incomplete data effectively.

Q.3. Solve any five Questions.
→ 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 & Customes segmentation.
3] Retail :- Recommendation Systems, inventory management, customer behaviours analysis & pricing optimized.

4] marketing :- Targated advertising sentiment analysis, Comapaign effectiveness & market research.

5] Transportation :- Route optimization predictive maintance & traffic mangement.

6] Social media :- Content relommendation trend analysis & user sentiment analysis.

7] education :- personalized learing student performance prediction & curiculum 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' closeset training examples in the features space to a new data point & assigning the class level based on the majority vote of the these neighbours.

Key aspects of KNN :-

- Distance metrix :- Commonly used Euclidean distance but can use other metrics.
- k-Value :- Detemines 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 eg. KNN doesn't build a model during training but makes decision based on the entire dataset during prediction.

c]

	BI	Datascience
1]	Analyzes historical data	Analyzes & predicts future outcomes.
2]	focuses on reporting & visualisation	focuses on modeling & algorithm.
3]	uses structured data sources	Handles structured & unstructured data
4]	primarily descriptive analytics.	uses predictive and prescriptive analytics.
5]	Tools : power BI, Tableu, SQL.	Tools : Python, R. machine learning libraries.
6]	Geared towards business decision making	Geared towards discovery & innovation

d] major difference between star Schema & Snowflake schema.

→

starschema :- A simple database schema used in data warehousing where a Central fact table is Connected to multiple dimesion 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 quenes and slower performance.

e] Benefits of Building an enterprise Data warehouse :

→

- Centralized Data Storage :
  provides a single source of truth for organisational data
- Improve Query performance :-

optimized for complex queries 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 comprehsive & consitent data.
- Data Integration :
  Combines data from various sources into a unified view.

Q4. Solve any five questions:

(E) A) Data cleaning :

Data cleaning refers to the process of identifying and correcting errors or inconsistenie. ina 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 :

Page No.:
Date:

1) Removing Duplicates :- Identifying & deleting duplicate records to ensure each entry in unique.

2) Handling missing Values :- Techniques like imputation or removing rows/ Columns with missing data.

3) standardizing Data :- Ensuring Consistence 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 or bring it into a common range or format.

B) Machine learning :-

machine learning is a subset of Artificial intelligence that involves traing algorithm to learn pattern from data & make predictions or decision without being explicitly programmed It encompasses :-

i) Supervised learning :- Algorithms learn from labled training data to

predict outcomes for new data.

2] unsupervised learning :- Algorithms identifying patterns or grouping in unlabled data.

3] Reinforcement learning : Algorithms learn by interacting with an enviornment & receiving feedback in the form of rewards or penalties.

4] model Evaluation :- Techniques like Cross validation & performance metrics assess how well model is generalize to new data.

d)

Data analytics	Data Analysis
It extracts actionable insights.	It examines data for understanding.
2. predicts future trends	Describe past events.
uses advanced algorithms	uses statistical methods
focuses on decision making	focuses on data interpretation.

5.	Prescriptive and predictive	Decriptive and diagnostic.
	Handles large datasets	Handles structured often smaller datasets.
	Involves Complex tools (AI/ML)	Involves simpler tools.
	often real-time data	Typically historically data.
	Business & operation al focus	Research & reporting focus.

remaining

Explain multidimesional view of Data Cube.

→ Multidimensional view :-

- organizes data into multiple dimensions
- allows analysis from various persepectives.
- Dimensions can be attribute like time location or product.
- facilitates deeper data exploration
- Helps uncover patterns & trends.
- useful in complex data queries

Data cubes:-
- A multidimensional array structure 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.

6] 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.

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

Ishel
24/9/24

Page No.:
Date:

Internal Examination 22-23

Question paper No-2

Q.1 A) Choose the correct answer.

1. b. Useful Information
2. c. Metaldata
3. b. Snowflake Schema
4. b. Squares
5. a. Itemset

B] Define the following terms.

1) Clustering :- Grouping a set of dots points into clusters where objects in the same cluster are more similar to each other than to those in other clusters.

2) Regression :- A statistical method used to predict a continous outcome variable based on one or more predictor variables.

3] Supervised learning :- A type of machine learning where the model is trained on labled data to predict outcomes for new, unseen data.

4] classifier :- An algorithm that categorizes data into predefined classes or lables.

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

Q.2. Solve any five Questions.

a) challenges of data Science technology :-

1] Data quality & Availability :- In Consistent, incomplete or unstructured data can affect the accuracy and reliability of data science models.

2] scalability & Computational power. Handling larger datasets & performing Complex computations require significant Computational resources & efficient algorithms

3] Privacy & ethical Concerns :- protecting sensitive data, & Complying with regulations like GDPR pose significant challenge.

Page No.:
Date:

b) Application of clustering :-

1. Customer segmentation :- Grouping customers based on behaviour, preference 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 organisation retrival and processing.

c) Prediction :-

Prediction involves using historical finance for analysis.

2. Integrated :- Combines data from multiple sources into a unified format.

3. Time - Variant :- Stores historical data to tract changes over time for tend 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 likehood 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 : Handles both binary & multi-class problems works well with small datasets & is computationally.

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 element or disease outbreaks.

M T W T F S S
Page No.:
Date:
YOUVA

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 optional boundary (hyperplane) that best separates differents classes in the data.

3. Naive Bayes :- A probabilistic algorithm based on Bay'e'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, 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 & cloums typically in relational database.

e.g. Spreadsheets, SQL, databases.

2. Unstructured Data: lacks a predefined format making it challenging to analyze.
eg: 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, No SQL database.

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

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. social media posts.

e) Agglomerative clustering | Divisive clustering

1. Starts with each data point as a separate cluster. | 1. Starts with all data points in a single cluster.

2. merges clusters iteratively | 2. splits clusters iteratively.

Commonly used with hiemarchical clustering | less common but also hiemachial

Works bottom-up (builds up from individual points) | Works top-down. (breaks down from a single cluster)

Typically more computationally efficient with fewer cluster | Can be more computationally intensive.

e.g. Single-linkage, Complete-linkage. | e.g: top-down, recursive.

Page No.:
Date:

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

1. Central fact table :- Contains quantitative data & foreign keys linking to dimesion 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 dimesion tables surround it, resembling a star. This designs simplifies queries & reporting.
4. Query efficiency :- The schema enhances performance by reducing complex joins & improving retrieval times, making it ideal for OLAP systems.

Q.4. Solve any five questions.

OLTP	Data warehouse
Application oriented	Subject oriented.
used to run business	used to analyzed business.
Detailed data	summerized & refined data.
Current up to date	History to current.
repetitive data	Ad-hoc access.
clerical user	knowledge user.
performance sensitive.	performance relaxed.
few records accessed at a time	Large volumes access at a time.
Read/update.	mostly Read.

No data Redundancy	Redundancy present.
Database size. 100 MB - 100 GB	Database Size 100 GB - Few TB

d) Different types of data used in cluster analysis.

1. Numerical data: Continous 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
4. Ordinal Data: Data that represents orders or ranking but the intervals between values are not necessarily uniform.
5. mixed-Type Data: Datasets that contain a combination of numerical, categorical & ordinal data requiring specialized algorithms like k-prototypes to

handle multiple types uniform.

8) Decision tree classifier with 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 represents outcomes of those decisions and leaves ~~represent outcomes of those decisions and leaves represent~~ represent class tables. 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 dateset.

Weather	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 (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.

Bhola
24/9/24