Model Question paper 2

Part A

I,. Answer any four of the following questions. 4×2=8

▼ What are ETL Tools?

ETL stands for Extract, Transform, Load. ETL tools are software applications that facilitate the extraction of data from various sources, transforming it into a desired format, and loading it into a target database for analysis and reporting.

▼ What is Time Series Analysis?

Time Series Analysis involves studying and modeling the patterns, trends, and behaviors in sequential data points over time. It is commonly used in forecasting and understanding the temporal aspects of data.

▼ What is Knowledge Discovery in Databases (KDD)?

Knowledge Discovery in Databases is the process of extracting useful patterns, trends, and knowledge from large volumes of data. It involves various steps such as data selection, data preprocessing, data transformation, and pattern evaluation.

▼ What are Association Rules?

Association rules are relationships or patterns that can be discovered in a dataset, indicating the likelihood of certain events occurring together. They are commonly used in market basket analysis to identify associations between products in transactions.

▼ What is Dendrogram? Give an example.

- A dendrogram is a tree diagram used to represent hierarchical relationships between sets of data. It is often used in hierarchical clustering to visualize the arrangement of clusters.
- **Example:** In hierarchical clustering of species based on genetic similarities, a dendrogram may illustrate how species are grouped into clusters based on their evolutionary relationships.

▼ Define Association Rule and Association Rule Problem.

- **Association Rule Definition:** An association rule is a pattern or

relationship discovered in a dataset that shows the likelihood of one event (or set of events) occurring in conjunction with another event (or set of events).

- **Association Rule Problem:** The association rule problem involves finding interesting relationships or patterns in data. For example, in a retail dataset, discovering that customers who buy bread are also likely to purchase butter represents an association rule.

Part B

II. Answer any four of the following questions. 4x5=20

Explain ETL Process and Pipelining Principle in ETL

- **ETL Process:**
    1. **Extract:** Retrieve data from various sources, such as databases, files, or external systems.
    2. **Transform:** Clean, filter, and structure the extracted data to meet the desired format and quality standards.
    3. **Load:** Load the transformed data into a target data warehouse, database, or analytics platform for analysis.
- **Pipelining Principle in ETL:**
    - Pipelining involves creating a seamless flow of data from one stage of the ETL process to the next. The output of one stage becomes the input for the next, allowing for continuous and efficient processing. This ensures a streamlined and optimized ETL workflow.

Explain the differences Between KDD and Data Mining.

- **Knowledge Discovery in Databases (KDD):**
    - Encompasses the entire process of discovering patterns, trends, and knowledge from large volumes of data.
    - Involves stages such as data selection, preprocessing, transformation, data mining, evaluation, and interpretation.
- **Data Mining:**
    - Specifically refers to the application of algorithms and techniques to extract patterns or knowledge from data.

techniques to extract patterns or knowledge from data.

- A subset of the broader KDD process, focusing on the analysis and modeling stages.

Explain ID3 Algorithm with an example.

- **ID3 Algorithm:**
  - A decision tree algorithm for classification.
  - Selects the best attribute at each node based on information gain.
  - Continues recursively until a stopping criterion is met, forming a tree.
- **Example:**
  - Consider a dataset with weather data (outlook, temperature, humidity) and a target variable (play tennis: yes or no).
  - ID3 selects the attribute with the highest information gain (e.g., Outlook).
  - It creates branches for each value of Outlook (Sunny, Overcast, Rainy) and continues the process.

Explain the simple approach to classification with an example.

- **Simple Approach to Classification:**
  - Assigns a new instance to the class of the majority of its k-nearest neighbors.
  - The number of neighbors (k) is a parameter that influences the decision.
- **Example:**
  - In a dataset of points with features and class labels, when a new point needs classification, the simple approach looks at the class labels of its k-nearest neighbors.
  - If the majority of the neighbors belong to a specific class, the new point is assigned to that class.

Write a short note on Support, Confidence, and Lift.

Explain Fast Update (FUP) Approach. How it works? Explain with an

example.

- **Fast Update (FUP) Approach:**
  - A technique to efficiently update association rules when the underlying data changes.
  - Instead of recalculating rules from scratch, it updates existing rules based on the changes in the dataset.
- **Working:**
  - Identify the subset of rules affected by the data changes.
  - Update the support and confidence values of the affected rules without reprocessing the entire dataset.
- **Example:**
  - Consider a retail dataset where the support for a particular item changes due to new transactions.
  - Instead of recalculating support for all items, the FUP approach selectively updates the rules associated with the changed item, making the process faster and more efficient.

Part C

III. Answer any four of the following questions.4x8=32

What are Similarity Measures in data mining? Explain.

- **Similarity Measures:**
  - Quantify the degree of similarity or dissimilarity between two objects or instances in a dataset.
  - Commonly used in clustering, classification, and recommendation systems.
- **Examples:**
  - **Euclidean Distance:** Measures the straight-line distance between two points in multidimensional space.
  - **Cosine Similarity:** Measures the cosine of the angle between two vectors, often used for text data.
  - **Jaccard Similarity:** Measures the intersection over the union of sets, suitable for binary data.

- Sets, suitable for binary data.

What is Descriptive Data Mining Tasks? Explain the types of Descriptive Data Mining Tasks.

- **Descriptive Data Mining Tasks:**
  - Involve summarizing and describing the general properties and patterns in a dataset.
- **Types:**
  - **Clustering:** Group similar instances into clusters based on shared characteristics.
  - **Association Rule Mining:** Discover relationships and associations between variables.
  - **Sequential Pattern Mining:** Identify patterns in sequences, such as time series or ordered data.
  - **Summary Statistics:** Calculate and present descriptive statistics for the datas

Explain the metrics to evaluate the performance of the classification algorithm.

Metrics for Classification Evaluation:

1. **Accuracy:** Proportion of correctly classified instances.
2. **Precision:** Proportion of true positives among instances predicted as positive.
3. **Recall (Sensitivity):** Proportion of true positives among actual positive instances.
4. **F1 Score:** Harmonic mean of precision and recall.
5. **Confusion Matrix:** Tabulates true positives, true negatives, false positives, and false negatives.

Explain K-Means Clustering Algorithm. How it works? Write K-Means Clustering Algorithm.

- **K-Means Clustering Algorithm:**
  1. **Input:** Set of data points and the desired number of clusters (k).
  2. **Initialization:** Randomly select k data points as initial cluster

centroids.

3. **Assignment:** Assign each data point to the nearest centroid, forming k clusters.
4. **Update Centroids:** Recalculate the centroids as the mean of points in each cluster.
5. **Repeat Steps 3 and 4:** Iterate until convergence (minimal change in centroids).

- **Example:**
  - Consider a dataset of customer purchase behavior. K-Means can group customers into k clusters based on their spending patterns, with each cluster representing a distinct customer segment.

Explain the divisive clustering algorithm.

- **Divisive Clustering Algorithm:**
  - Starts with a single cluster containing all data points and recursively divides it into subclusters.
  - Continues until each data point is in its own cluster or a predefined stopping criterion is met.
- **Working:**
  - Begin with one cluster encompassing all data points.
  - Identify the most dissimilar subset of data points and split the cluster.
  - Repeat the process recursively for each subcluster until the stopping criterion is satisfied.

What is Sampling Algorithm? How it works? Explain with an example.

- **Sampling Algorithm:**
  - Involves selecting a subset of instances from a larger dataset to represent its characteristics.
  - Used for efficiency in data analysis when working with large datasets.
- **Working:**

    - Randomly select instances or use systematic sampling techniques.
    - Ensure that the sample is representative of the overall population.
    - Analyze the sample to draw conclusions about the entire dataset.
- **Example:**
    - In a population of one million records, a sampling algorithm might randomly select 10,000 instances for analysis. The conclusions drawn from analyzing the sample are then extrapolated to make inferences about the entire population.