Code No: 157BC

**R18**

**JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERABAD**

**B. Tech IV Year I Semester Examinations, January/February - 2023**
**DATA MINING**

**(Common to CSE, IT, ITE)**

**Time: 3 Hours** **Max. Marks: 75**

**Note:** i) Question paper consists of Part A, Part B.
ii) Part A is compulsory, which carries 25 marks. In Part A, answer all questions.
iii) In Part B, Answer any one question from each unit. Each question carries 10 marks and may have a, b as sub questions.

**PART** – A

**(25 Marks)**

1.a) What is data warehouse? [2]
b) List out the applications of data mining. [3]
c) What is meant by association rule mining? [2]
d) Write a short note on SPM algorithm? [3]
e) Why are decision trees useful? [2]
f) List the advantages of using decision trees. [3]
g) Discuss the two approaches to improve quality of hierarchical clustering. [2]
h) List the applications of cluster analysis. [3]
i) Define data stream mining. [2]
j) Give the taxonomy of web mining. [3]

**PART – B**

**(50 Marks)**

2.a) Explain how to integrate data mining system with a data warehouse.
b) "Data preprocessing is necessary before data mining process". Justify your answer. [5+5]

**OR**

3.a) Differentiate between data mining and data
b) warehouse. Discuss the major issues in data mining. [5+5]

4.a) Write a short notes on constraint based association
b) mining. Describe various types of association rules. [5+5]

**OR**

5. Explain in detail about frequent pattern mining in data mining. [10]

6. Describe Bayesian Belief Network with an example. [10]

**OR**

7. a) Briefly explain classification problems and general approaches to solve them.
   b) Explain the merits and de-merits of the lazy learning method. [5+5]

8. Explain the following.
   a) Cluster analysis.
   b) Grid–based methods. [5+5]

**OR**

9.a) How density based method is used for clustering?
   b) Illustrate K-mean algorithm with an example. [4+6]

10. Explain the following.
   a) Spatial data mining.
   b) Text mining. [5+5]

**OR**

11. Discuss various kinds of patterns to be mined from web/server logs in web usage mining. [10]

## Answers

### PART – A

1. **(25 Marks)**

**a) What is data warehouse?**

A data warehouse is a centralized repository that stores large volumes of structured data from various sources. It is designed to support business intelligence activities, such as querying, reporting, and data analysis. Data warehouses enable organizations to consolidate data for better decision-making and insights.

**b) List out the applications of data mining.**

Market Basket Analysis: Identifying product purchase patterns to optimize cross-selling and promotions in retail.

Customer Segmentation: Classifying customers into distinct groups based on purchasing behavior for targeted marketing.

Fraud Detection: Detecting unusual patterns and anomalies in transactions to identify fraudulent activities in finance.

Predictive Maintenance: Analyzing machinery data to predict failures and schedule timely maintenance in manufacturing.

c) **What is meant by association rule mining?**
Association rule mining is a data mining technique used to discover interesting relationships or patterns between variables in large datasets. It identifies rules that highlight how the occurrence of one item is associated with the occurrence of another item. This technique is commonly used in market basket analysis to find product co-purchase behaviors.

d) **Write a short note on SPM algorithm?**
The Sequential Pattern Mining (SPM) algorithm is used to find regular sequences or patterns over time in large datasets. It identifies subsequences that frequently occur in a given order, making it valuable for applications like analyzing customer purchasing sequences or web clickstreams. The algorithm helps uncover temporal relationships and trends in sequential data.

e) **Why are decision trees useful?**
Decision trees are useful because they provide a clear and interpretable model for decision-making by visually mapping out possible outcomes based on different input variables. They handle both numerical and categorical data, making them versatile for various applications. Additionally, they are easy to understand and implement, aiding in transparent and explainable AI.

f) **List the advantages of using decision trees.**
Easy to Understand: Decision trees provide a visual representation that is easy to interpret and explain.
Versatile: They can handle both numerical and categorical data effectively.
No Need for Data Normalization: Decision trees do not require data scaling or normalization.
Handles Non-linear Relationships: They can capture complex, non-linear relationships between variables.

g) **Discuss the two approaches to improve quality of hierarchical clustering.**
**Optimal Number of Clusters:** Use methods like the Elbow Method or Silhouette Analysis to determine the optimal number of clusters, ensuring meaningful groupings.
**Distance Metrics and Linkage Criteria:** Improve clustering quality by selecting appropriate distance metrics (e.g., Euclidean, Manhattan) and linkage criteria (e.g., single, complete, average) that best fit the data characteristics.

h) **List the applications of cluster analysis.**
**Customer Segmentation:** Grouping customers based on purchasing behavior for targeted marketing.
**Market Research:** Identifying distinct consumer segments to tailor products and services.

**Image Segmentation:** Dividing images into regions for analysis in computer vision.
**Anomaly Detection:** Detecting outliers or unusual patterns in datasets for fraud detection and monitoring.

i) **Define data stream mining.**
Data stream mining is the process of extracting knowledge and patterns from continuous, rapid data streams in real-time. It focuses on analyzing and processing data as it arrives, rather than storing it for later analysis. This technique is essential for applications requiring immediate insights, such as online monitoring and fraud detection.

j) **Give the taxonomy of web mining.**
Web mining is categorized into three main types: Web Content Mining*(extracting useful information from web page content), Web Structure Mining (analyzing the structure of hyperlinks within the web), and Web Usage Mining (analyzing user interaction data, such as logs, to understand behavior). Each type focuses on different aspects of web data for various analytical purposes.

**2.a) Explain how to integrate data mining system with a data warehouse.**

Integrating a data mining system with a data warehouse involves several steps:

1. **Data Preparation:** Extract relevant data from the data warehouse, ensuring it's clean, formatted, and suitable for analysis.

2. **Data Transformation:** Transform the extracted data into a format compatible with the data mining tools, such as converting categorical variables or handling missing values.

3. **Model Building:** Use data mining algorithms to build models and extract patterns, trends, or insights from the prepared data.

4. **Evaluation:** Assess the effectiveness and accuracy of the models generated using techniques like cross-validation or holdout testing.

5. **Deployment:** Deploy the models back into the data warehouse environment, ensuring they are accessible for ongoing analysis and decision-making.

6. **Monitoring:** Continuously monitor the performance of the deployed models to ensure they remain accurate and relevant over time.

7. **Feedback Loop:** Incorporate feedback from model performance into the data mining process to refine and improve future analyses.

8. **Collaboration:** Foster collaboration between data mining experts and data warehouse administrators to optimize the integration process and maximize the value of insights derived from the data.

**b) "Data preprocessing is necessary before data mining process". Justify your answer.**

Data preprocessing serves as a critical foundation for the data mining process, playing a pivotal role in ensuring the accuracy, efficiency, and effectiveness of subsequent analyses. Firstly, it involves cleaning the data to remove inconsistencies, errors, and outliers that can skew results and undermine the reliability of patterns extracted during mining. By addressing missing values, duplicates, or inaccuracies, preprocessing enhances the quality of the dataset, fostering more reliable and insightful outcomes. Moreover, data preprocessing facilitates the transformation of raw data into a suitable format for analysis, including converting categorical variables, scaling numerical features, or normalizing distributions. This standardization ensures that disparate data sources can be effectively integrated and compared, enabling comprehensive analysis across the entire dataset. Additionally, preprocessing reduces computational complexity by eliminating redundant or irrelevant attributes, thus streamlining the data mining process and improving efficiency. Furthermore, preprocessing enables feature selection and dimensionality reduction, which not only enhance the interpretability of results but also mitigate the curse of dimensionality, especially crucial for high-dimensional datasets. In essence, data preprocessing acts as a crucial preparatory step that optimizes data quality, compatibility, and efficiency, laying the groundwork for meaningful insights and actionable outcomes in the subsequent data mining process.

**3.a b) Differentiate between data mining and data warehouse. Discuss the major issues in data mining.**

a)

| Aspect | Data Mining | Data WareHouse |
|---|---|---|
| **Purpose** | Discovers patterns, trends, and insights in data to extract actionable information. | Stores large volumes of structured data from various sources for analysis. |
| **Focus** | Analyzes data to uncover hidden patterns, relationships, or anomalies. | Consolidates data from multiple sources nto a centralized repository. |
| **Process** | Involves applying algorithms and techniques to large datasets to extract knowledge. | Primarily focuses on data storage, retrieval, and management. |
| **Outcome** | Generates actionable insights and knowledge to support decision-making processes. | Provides a unified view of organizational data for reporting and analysis. |

b) Data mining, despite its immense potential, grapples with several significant challenges. Firstly, ensuring data quality remains paramount, as poor-quality data can yield misleading or inaccurate insights. Secondly, managing large volumes of data poses scalability issues, requiring robust infrastructure and algorithms capable of handling massive datasets efficiently. Additionally, data privacy and security concerns arise due to the sensitive nature of the information being analyzed, necessitating stringent measures to protect confidentiality and compliance with regulations. Moreover, the curse of dimensionality presents challenges in high-dimensional datasets, leading to increased computational complexity and decreased algorithm performance. Interpretability of complex models is another issue, as black-box algorithms may obscure the understanding of how predictions are made, hindering trust and adoption. Furthermore, bias and discrimination in data and algorithms can perpetuate inequalities, emphasizing the need for fairness-aware mining techniques. Handling evolving data streams in real-time poses further challenges, requiring adaptive algorithms capable of continuous learning and adaptation. Lastly, ethical considerations regarding the responsible use of mined insights and the potential societal impacts underscore the importance of ethical frameworks and governance in data mining practices. Addressing these issues requires a multifaceted approach, encompassing technological advancements, regulatory frameworks, and ethical guidelines to harness the full potential of data mining while mitigating its inherent risks.

**4. a) Write a short notes on constraint based association mining.**

**b) Describe various types of association rules.**

a) Constraint-based association mining focuses on discovering associations between items in a dataset while adhering to specified constraints. Unlike traditional association mining, which explores all possible itemsets, constraint-based approaches use predefined rules or conditions to guide the mining process. These constraints can include minimum support thresholds, minimum confidence levels, or specific item relationships. By incorporating constraints, this method reduces the search space, leading to more efficient and targeted mining, thus improving the quality and relevance of discovered associations. Constraint-based association mining techniques include Apriori-based algorithms and FP-growth, offering flexibility in defining constraints to suit diverse mining objectives and datasets.

b) Association rules are patterns that describe relationships among items in a dataset. Several types of association rules exist, each with its own characteristics and applications:

**Frequent Itemsets**: These rules identify sets of items that frequently occur together in a dataset. They are essential for market basket analysis and understanding customer purchasing patterns.

**Closed Itemsets:** Closed itemsets are maximal frequent itemsets where no super-pattern has the same support. They offer a more concise representation of frequent itemsets, which can improve efficiency in pattern mining algorithms.

**Maximal Itemsets:** Maximal itemsets are frequent itemsets that are not subsets of any other frequent itemset. They provide a comprehensive view of all potential associations in the dataset, without redundancy.

**Association Rules:** These rules consist of an antecedent (or premise) and a consequent (or conclusion), indicating an implication or correlation between sets of items. They are expressed in the form "if {A} then {B}" and are used for decision-making and recommendation systems.

**Sequential Rules**: Sequential rules capture temporal associations between sequences of events or transactions. They are prevalent in analyzing time-stamped data such as web clickstreams or transaction histories.

**Constraint-Based Rules:** Constraint-based rules incorporate predefined constraints such as minimum support or confidence thresholds to guide the rule generation process. They help filter out less relevant or insignificant associations, focusing on those that meet specific criteria.

**5. Explain in detail about frequent pattern mining in data mining.**

Frequent pattern mining is a fundamental task in data mining aimed at discovering sets of items (or itemsets) that frequently occur together in a dataset. Here's a detailed explanation in 8 points:

1. Frequent pattern mining involves identifying itemsets with a support value exceeding a predefined threshold. Support represents the frequency with which an itemset appears in the dataset. Higher support indicates greater significance and relevance of the itemset.

2. The Apriori principle is a fundamental concept in frequent pattern mining. It states that any subset of a frequent itemset must also be frequent. This principle forms the basis for efficient algorithms in mining frequent patterns by reducing the search space.

3. Various algorithms are used for frequent pattern mining, including the Apriori algorithm, FP-growth algorithm, and Eclat algorithm. These algorithms employ different strategies for traversing the search space efficiently and identifying frequent itemsets.

4. The Apriori algorithm iteratively generates candidate itemsets and prunes those that fail to meet the minimum support threshold. It starts by finding frequent individual items (singletons) and progressively extends to larger itemsets, exploiting the Apriori principle to reduce computational overhead.

5. The FP-growth algorithm constructs a compact data structure called the FP-tree to represent the dataset. It recursively partitions the dataset based on item frequencies and mines frequent patterns directly from the FP-tree without generating candidate itemsets, leading to improved efficiency.

6. Frequent pattern mining often serves as a precursor to mining association rules. Once frequent itemsets are identified, association rules can be generated by considering subsets of these itemsets and calculating their confidence values.

7. Frequent pattern mining finds applications in various domains such as market basket analysis, recommendation systems, web usage mining, bioinformatics, and network traffic analysis. It enables businesses to understand customer behavior, identify cross-selling opportunities, and optimize processes based on discovered patterns.

8. Despite its utility, frequent pattern mining faces challenges such as scalability issues with large datasets, the curse of dimensionality in high-dimensional data, and the need to handle noise and sparsity effectively. Advanced techniques,

parallelization, and optimization strategies are employed to address these challenges.

**6. Describe Bayesian Belief Network with an example**

A Bayesian Belief Network (BBN) is a graphical probabilistic model that represents probabilistic relationships among a set of variables using a directed acyclic graph (DAG). In a BBN, nodes in the graph represent random variables, and edges represent dependencies between variables, indicating causal or influential relationships. Each node is associated with a conditional probability distribution that quantifies the probability of the node given its parents in the graph. BBNs facilitate reasoning under uncertainty by allowing inference on the probability distribution of unobserved variables given evidence on observed variables. They enable decision-making, prediction, and risk assessment by providing a formal framework for representing and reasoning about uncertain knowledge and dependencies among variables. BBNs find applications in various domains, including healthcare, finance, engineering, and environmental science, where modeling complex systems with uncertainty is crucial for informed decision-making and analysis.

Let's consider a medical diagnosis scenario as an example of a Bayesian Belief Network (BBN). Suppose we have three variables: "Symptoms" (S), "Disease" (D), and "Test Result" (T).

- **Symptoms (S)**: Represents whether a patient exhibits certain symptoms associated with a disease, such as fever, cough, and fatigue.
- **Disease (D)**: Represents the presence or absence of a particular disease, such as influenza.
- **Test Result (T)**: Represents the outcome of a diagnostic test for the disease, such as a PCR test for influenza.

The BBN would depict causal relationships between these variables. For instance:

- Symptoms (S) influence the likelihood of having the Disease (D).
- The Disease (D) influences the likelihood of observing certain Symptoms (S).
- The Test Result (T) is influenced by the presence of the Disease (D).
- The observed Symptoms (S) may also influence the decision to conduct a Test (T).

**7. a) Briefly explain classification problems and general approaches to solve them.**

**b) Explain the merits and de-merits of the lazy learning method.**

**a**) Classification problems involve predicting a categorical label or class for a given input based on its features. These problems are prevalent in various domains, such as finance, healthcare, and image recognition. To solve classification problems, several general approaches are commonly employed. Firstly, supervised learning algorithms, such as decision trees, logistic regression, support vector machines (SVM), and neural networks, are widely used. These algorithms learn from labeled training data to build a model that maps input features to class labels. Additionally, ensemble methods like random forests and gradient boosting combine multiple models to improve predictive accuracy. Moreover, probabilistic classifiers, including Naive Bayes and Bayesian networks, estimate the probability of each class label given the input features, enabling uncertainty quantification. Furthermore, deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), excel in capturing complex patterns and dependencies in high-dimensional data, making them particularly effective for image, text, and sequential data classification tasks. Overall, a diverse range of algorithms and techniques exists to tackle classification problems, each with its strengths and suitability for different types of data and applications.

**b)** Lazy learning, also known as instance-based learning or memory-based learning, is a machine learning approach where the system memorizes the training instances and makes predictions based on their similarity to new, unseen instances. This method offers several merits. Firstly, lazy learning models are flexible and adaptable to complex, nonlinear patterns in the data, as they do not impose strong assumptions about the underlying data distribution. Secondly, they are computationally efficient during training, as they do not require an explicit model-building phase. Moreover, lazy learning methods are robust to noisy data and can handle dynamic or evolving datasets effectively, as they continuously update their knowledge based on new instances. Additionally, lazy learning inherently supports incremental learning, allowing the system to incrementally incorporate new data without retraining the entire model.

However, lazy learning also has its drawbacks. One major disadvantage is its high computational cost during prediction, as it requires comparing the new instance with all training instances to determine the most similar ones. This can become prohibitively expensive for large datasets or in real-time applications where low-latency predictions are required. Additionally, lazy learning models may suffer from overfitting, especially when the training dataset is noisy or contains irrelevant features, as they tend to memorize the training instances without generalizing well

to unseen data. Furthermore, lazy learning methods are sensitive to the curse of dimensionality, as the distance between instances becomes less meaningful in high-dimensional feature spaces, leading to degraded performance. Despite these limitations, lazy learning remains a valuable approach in machine learning, particularly for small to medium-sized datasets and applications where interpretability and adaptability are paramount.

**8. Explain the following.**

**a) Cluster analysis.**

**b) Grid–based methods**

a) Cluster analysis is a data exploration technique aimed at grouping similar data points into clusters, where points within the same cluster share common characteristics. It is an unsupervised learning method commonly used for pattern recognition, data mining, and exploratory data analysis. Cluster analysis helps uncover hidden structures and relationships in datasets by organizing data into meaningful groups without prior knowledge of class labels. Various algorithms, such as K-means, hierarchical clustering, and DBSCAN, are employed to partition data based on similarity measures or density-based criteria. The quality of clusters is evaluated using metrics like silhouette score or cohesion and separation measures. Applications of cluster analysis range from customer segmentation and market research to image segmentation and anomaly detection in diverse fields such as business, biology, and computer vision.

b) Grid-based methods are a class of clustering techniques that partition the data space into a grid structure and then assign data points to grid cells. These methods offer a scalable approach to clustering large datasets by dividing the data space into a finite number of cells, often represented as a grid or lattice. The primary advantage of grid-based methods is their efficiency in handling high-dimensional data and large datasets, as they do not require a full scan of the entire dataset for clustering. Instead, they operate by partitioning the data space into cells and only examine data points within each cell, reducing computational complexity. Examples of grid-based clustering algorithms include STING (Statistical Information Grid), CLIQUE (CLustering In QUEst), and WaveCluster. Grid-based methods find applications in various domains, including spatial data mining, image processing, and anomaly detection, where efficiency and scalability are essential for processing vast amounts of data.

**9. a) How density based method is used for clustering?**

**b) Illustrate K-mean algorithm with an example.**

**a)** Density-based clustering methods aim to discover clusters of arbitrary shapes in datasets by identifying regions of high density separated by regions of low density. These methods operate based on the notion that clusters are areas

where data points are densely concentrated, and they can handle clusters of varying shapes and sizes. One popular density-based clustering algorithm is DBSCAN (Density-Based Spatial Clustering of Applications with Noise).

In DBSCAN, clusters are formed by identifying core points, which are data points surrounded by a specified number of neighbors within a defined distance (epsilon). Additionally, there are two types of points in DBSCAN: core points and border points. Core points have at least the minimum number of neighbors within the epsilon distance, while border points have fewer neighbors but are reachable from a core point within the same cluster. Points that are not core points or border points are considered noise or outliers.

The process of clustering using DBSCAN involves the following steps:

1. Select Parameters: Specify the distance threshold (epsilon) and the minimum number of points required to form a cluster (minPts).
2. Identify Core Points: Determine which data points have at least minPts neighbors within the epsilon distance.
3. Expand Clusters: Form clusters by expanding around core points and including their neighboring points as part of the same cluster.
4. Assign Border Points: Assign border points to clusters if they are reachable from a core point.
5. Label Noise: Identify and label remaining points as noise or outliers if they do not belong to any cluster.

b) The K-means algorithm is a popular clustering technique used to partition a dataset into K clusters. Here's an illustration of the K-means algorithm:

1. Initialization: Choose the number of clusters K and randomly initialize K cluster centroids (points representing the center of each cluster).
2. Assign Data Points to Nearest Centroid: For each data point in the dataset, calculate the distance to each centroid and assign the point to the nearest centroid. This step creates K clusters.
3. Update Centroids: Recalculate the centroids of the clusters by taking the mean of all data points assigned to each cluster. The centroids represent the new center of each cluster.
4. Repeat: Iteratively repeat steps 2 and 3 until convergence criteria are met. Convergence occurs when the centroids no longer change significantly between iterations or when a predefined number of iterations is reached.
5. Final Clustering: Once convergence is achieved, the algorithm outputs the final clusters, where each data point is assigned to the cluster represented by the nearest centroid.

Here's an illustration using a two-dimensional dataset with K=3 clusters:

1. Initialization: Randomly select three points as initial centroids.

2. Assign Data Points to Nearest Centroid: Calculate the distance from each data point to each centroid and assign each point to the nearest centroid, creating three initial clusters.
3. Update Centroids: Recalculate the centroids of the clusters by taking the mean of all data points assigned to each cluster.
4. Repeat: Iteratively repeat steps 2 and 3 until convergence.
5. Final Clustering: Once convergence is reached, the algorithm outputs the final clusters, where each data point is assigned to the cluster represented by the nearest centroid.

**10. Explain the following.**

a) **Spatial data mining.**

b) **Text mining.**

a) Spatial data mining involves the discovery of patterns, relationships, and insights within spatial datasets. Here are five key points about spatial data mining:
Geographic Information Systems (GIS): Spatial data mining utilizes GIS technology to store, analyze, and visualize spatial data, such as maps, satellite imagery, and geographic coordinates.
Spatial Patterns: It identifies spatial patterns and trends in data, such as clustering of events, spatial outliers, hotspots, and spatial autocorrelation, to understand phenomena occurring in geographic space.
Applications: Spatial data mining finds applications in various domains, including urban planning, environmental science, epidemiology, transportation, agriculture, and natural resource management.
Techniques: Spatial data mining employs a range of techniques, including clustering, classification, association analysis, spatial regression, and spatial interpolation, adapted to handle spatially correlated data.
Challenges: Spatial data mining faces challenges such as data heterogeneity, spatial autocorrelation, scale dependency, computational complexity, and uncertainty in spatial data, requiring specialized methods and tools for analysis.

b) Text mining is the process of extracting useful information, patterns, and insights from unstructured textual data. It involves various techniques from natural language processing (NLP), machine learning, and statistics to analyze large volumes of text data. Text mining tasks include text categorization (classification), sentiment analysis, named entity recognition, topic modeling, and information extraction. By transforming unstructured text into structured data, text mining enables organizations to uncover trends, sentiment, and relationships hidden within textual documents. Applications of text mining span across multiple industries, including marketing, customer service, healthcare, finance, and social media analysis. Despite its potential, text mining faces challenges such as dealing with ambiguity, context dependency, language nuances, and the need for domain-

specific knowledge. Advanced text mining techniques, coupled with robust NLP algorithms and scalable computing resources, continue to advance the field, driving innovation in information retrieval, knowledge discovery, and decision-making processes.

**11. Discuss various kinds of patterns to be mined from web/server logs in web usage mining.**

In web usage mining, various kinds of patterns can be mined from web/server logs to understand user behavior and improve website performance. Here are eight types of patterns commonly mined:

1. Clickstream Patterns: Analyze the sequence of pages visited by users during a session, revealing browsing patterns, popular navigation paths, and session durations
2. Page Access Patterns: Identify frequently accessed pages or URLs, indicating popular content, areas of interest, and potential bottlenecks or server load issues.
3. Sessionization Patterns: Group user interactions into sessions based on time gaps or session duration, enabling analysis of session length distributions and user engagement metrics.
4. Referral Patterns: Analyze referral sources (e.g., search engines, social media, external links) that lead users to the website, helping understand traffic sources and effectiveness of marketing campaigns.
5. User Demographics Patterns: Extract demographic information (e.g., location, device type, browser) from user agents or IP addresses, facilitating audience segmentation and targeting.
6. Conversion Patterns: Track user interactions leading to desired outcomes (e.g., sign-ups, purchases, downloads), identifying conversion funnels, drop-off points, and factors influencing conversion rates.
7. Anomaly Detection Patterns: Detect unusual or anomalous user behavior, such as sudden spikes in traffic, abnormal navigation patterns, or suspicious activity indicative of security threats or system errors.
8. Sessionization Patterns: Group user interactions into sessions based on time gaps or session duration, enabling analysis of session length distributions and user engagement metrics.