|  |  |  |
| --- | --- | --- |
|  | **BAHRIA UNIVERSITY, (Karachi Campus)**  *Department of Software Engineering*  **Assignment 1 - Spring 2023** |  |



COURSE TITLE: **DATA MINING**

COURSE CODE: **CSC-452**

Class: **BSE-6 (A/B)**  Shift: **Morning**

Course Instructor: **Dr. Sorath Hansrajani**  Time Allowed: **1 week**

Release Date: **14-03-2023**

Due Date: **20-03-2023**  Max. Marks: **20 Marks**

**Name: M Muaz Shahzad Student ID: 02-131202-081 Section: BSE 6-B**

**Note:** Each question carry equal marks. **[CLO1: 20 Marks]**

**Question 01**

1. What is Data Mining? Describe the steps involved in data mining when viewed as a process of knowledge discovery.

**Data Mining:** Data mining is the process of extracting useful patterns and insights from large datasets using statistical and computational techniques. It is a form of knowledge discovery that involves identifying meaningful relationships, correlations, and patterns in the data that can be used to inform business decisions, scientific research, and other applications.

**Steps Of Data Mining**

The steps involved in data mining when viewed as a process of knowledge discovery include:

**Data Cleaning and Preprocessing:** This involves cleaning the data to remove any errors, inconsistencies, or missing values, and transforming the data into a format that can be analyzed using data mining techniques.

**Data Integration:** This involves combining data from multiple sources into a single dataset that can be used for analysis. This step may involve merging datasets, resolving conflicts between different data sources, and ensuring that the data is compatible with the data mining algorithms being used.

**Data Selection:** This involves selecting a subset of the data that is relevant to the analysis and removing any irrelevant or redundant data. This step may involve filtering the data based on certain criteria or using sampling techniques to select a representative subset of the data.

**Data Transformation:** This involves transforming the data into a format that can be used by the data mining algorithms. This may involve normalizing or scaling the data to ensure that it is consistent across different features or variables or transforming the data into a different format altogether.

**Data Mining:** This is the core step of the process, where various data mining algorithms are applied to the dataset to extract patterns and insights. The choice of algorithm depends on the nature of the problem and the type of data being analyzed.

**Pattern Evaluation:** This involves evaluating the patterns and insights that have been identified using the data mining algorithms. This step may involve testing the patterns against new data to ensure that they are accurate and reliable.

**Knowledge Representation:** This involves presenting the patterns and insights in a format that can be easily understood and used by business analysts, researchers, or other stakeholders. This may involve using visualization techniques or other forms of data representation to communicate the insights effectively.

Data mining is a complex and iterative process that requires careful planning, rigorous analysis, and ongoing evaluation to ensure that the insights are accurate, reliable, and useful for the intended application.

1. How is a data warehouse different from a database? How are they similar?

A data warehouse is a type of database that is designed for data analysis and decision-making. While both data warehouses and databases are used to store data, there are several key differences between them:

**Purpose**

1. The primary purpose of a database is to manage and organize data, while the primary purpose of a data warehouse is to support business intelligence and decision-making.
2. Data warehouses are optimized for query and analysis, while databases are optimized for transaction processing.

**Data Structure**

1. Databases typically store data in a normalized format, with data stored in multiple tables that are linked together by primary and foreign keys.
2. In contrast, data warehouses typically store data in a denormalized format, with data organized into tables that are optimized for query performance.

**Data Integration**

1. Data warehouses are often used to integrate data from multiple sources, such as transactional databases, legacy systems, and external sources.
2. Databases are typically designed to manage data within a single application or system.

**Historical Data**

1. Data warehouses often store historical data, allowing users to analyze trends and patterns over time.
2. Databases typically only store current or recent data.

**Query and Analysis**

1. Data warehouses are optimized for complex queries and data analysis.
2. Databases are optimized for simple queries and data retrieval.
3. Explain the difference and similarity between classification and regression.

**Solution:**

**Classification and regression** are two commonly used types of machine learning problems. These are similar in that they both involve predicting an output variable based on input variables. However, the main difference between classification and regression is the type of output variable they predict.

**Classification** involves predicting a categorical variable, or a variable with discrete values that belong to a specific class or category.

Examples of classification tasks include predicting whether an email is spam or not, or predicting whether a customer will churn or not.

**Regression** involves predicting a continuous variable, or a variable with continuous values that can take on any value within a given range.

Examples of regression tasks include predicting a person's height based on their weight or predicting a house's price based on its size and location.

One more difference between classification and regression is the type of algorithms used to solve them. **Classification** algorithms typically use decision trees, random forests, or support vector machines.

**Regression** algorithms typically use linear regression, polynomial regression, or decision trees.

**Question 02**

Define each of the following data mining functionalities: characterization, discrimination, association and correlation analysis, classification, regression, clustering, and outlier analysis. Give examples of each data mining functionality, using a real-life database that you are familiar with.

**Solution:**

**Characterization:** Characterization is the process of summarizing and describing the general properties of a dataset. This functionality involves identifying patterns and trends in the data to gain a better understanding of its overall characteristics.

**Example:** An example of characterization in a real-life database would be analyzing sales data to determine which products are the most popular, which customers purchase the most frequently, and which regions generate the most revenue.

**Discrimination:** Discrimination is the process of identifying patterns that distinguish different groups within a dataset. This functionality involves comparing different groups to identify significant differences in their characteristics.

**Example:** An example of discrimination in a real-life database would be analyzing customer data to determine if there are any demographic or behavioral differences between customers who make purchases regularly versus those who only make purchases occasionally.

**Association and Correlation Analysis:** Association and correlation analysis involve identifying relationships between different variables in a dataset. Association analysis identifies co-occurrence relationships between variables, while correlation analysis identifies how strongly two variables are related to each other.

**Example:** An example of association and correlation analysis in a real-life database would be analyzing retail sales data to determine which products are frequently purchased together (association analysis) and whether there is a correlation between the price of a product and its sales volume (correlation analysis).

**Classification:** Classification is the process of predicting categorical outcomes based on input variables. This functionality involves creating a model that assigns new observations to predefined categories based on their characteristics.

**Example:** An example of classification in a real-life database would be building a model that predicts whether a loan applicant is likely to default or not based on their credit history and other relevant factors.

**Regression:** Regression is the process of predicting continuous outcomes based on input variables. This functionality involves creating a model that estimates the relationship between the input variables and the

continuous output variable.

**Example:** An example of regression in a real-life database would be building a model that predicts a company's revenue based on its advertising spending, product prices, and other relevant factors.

**Clustering:** Clustering is the process of grouping similar observations together based on their characteristics. This functionality involves identifying clusters of observations that are similar to each other and different from other observations in the dataset.

**Example:** An example of clustering in a real-life database would be grouping customers into segments based on their demographic and behavioral characteristics.

**Outlier Analysis:** Outlier analysis involves identifying observations that are significantly different from other observations in the dataset. This functionality involves detecting unusual or anomalous observations that may require further investigation.

**Example:** An example of outlier analysis in a real-life database would be identifying fraudulent transactions in a financial dataset by detecting transactions that are significantly different from normal transaction patterns.

**Question 03**

1. Present an example where data mining is crucial to the success of a business. What data mining functionalities does this business need (e.g., think of the kinds of patterns that could be mined)? Can such patterns be generated alternatively by data query processing or simple statistical analysis?

**Solution:**

Example of a business where data mining is crucial to success is an **E-commerce company**.

An E-commerce company collects large amounts of data from various sources, including customer transactions, website clickstream data, social media, and marketing campaigns. This data can be used to uncover patterns and insights that can help improve business operations and increase revenue. Some of the data mining functionalities that would be useful for an e-commerce company include:

**Association rule mining:** This functionality can be used to identify products that are frequently purchased together by customers. By identifying these patterns, the e-commerce company can create product bundles or cross-sell promotions that can increase sales and customer satisfaction.

**Customer segmentation:** This functionality can be used to group customers into different segments based on their purchasing behavior, demographics, or other factors. This can help the e-commerce company tailor their marketing campaigns to specific customer segments and increase the effectiveness of their advertising efforts.

**Recommendation systems:** This functionality can be used to recommend products to customers based on their past purchases, browsing history, and other data. By providing personalized recommendations, the e-commerce company can increase customer satisfaction and retention.

While some of the patterns generated by data mining can be identified through data query processing or simple statistical analysis, these methods may not be as effective or efficient as data mining techniques. Data mining algorithms can handle large volumes of data and can identify complex patterns and relationships that may not be apparent through other methods. Additionally, data mining algorithms can be used to predict future trends and behaviors, which can be valuable for making strategic business decisions.

1. What are the major challenges of mining a huge amount of data (e.g., billions of tuples) in comparison with mining a small amount of data (e.g., data set of a few hundred tuple)?

**Solution:**

Mining a huge amount of data poses several challenges that are not present when mining a small amount of data. Some of the major challenges are:

**Processing Time:** Mining a large dataset takes a lot of processing time and resources. The algorithms used for data mining must handle the large volume of data, which can take a significant amount of time to process. As the dataset grows, the processing time increases exponentially, making it difficult to mine large datasets in a reasonable amount of time.

**Storage:** Storing large amounts of data can be a challenge, as it requires a lot of storage space. Even with modern storage technology, storing and managing large amounts of data can be expensive and time-consuming.

**Quality of Data:** As the dataset grows, the quality of the data can become an issue. Large datasets can contain errors, inconsistencies, and missing data, which can affect the accuracy of the mining results.

**Complexity of Analysis:** Analyzing a large dataset can be complex, as it can contain many variables and factors that interact in complex ways. It can be challenging to identify the most relevant variables and relationships between them, which can affect the accuracy and usefulness of the mining results.

**Bias and Sampling:** When mining a large dataset, it can be difficult to ensure that the data is representative of the entire population. Sampling bias can occur, which can skew the results of the analysis and lead to inaccurate or misleading insights.

Mining a huge amount of data requires significant computing power, storage capacity, and expertise. It is important to carefully consider the challenges and limitations of mining large datasets before embarking on a data mining project.

**Question 04**

Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

1. What is the mean of the data? What is the median?

**Mean:** (13+15+16+16+19+20+20+21+22+22+25+25+25+25+30+33+33+35+35+35+35+36+40+45+46+52+70)/27 = 29.96

**Median =** 25

1. What is the mode of the data? Comment on the data’s modality (i.e., bimodal, trimodal, etc.).

**Mode:** 25, 35

The data appears to be unimodal, with a single peak at 25, 35

1. What is the midrange of the data?

(70+30)/2 = 83/2 = 41.5

**Midrange** = 41.5

1. Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?

First quartile is 25% of data from starting position.

7th position number

Q1 = 20

Third quartile is eliminating the 75% of data from starting position.

Q3 = 35

Hence 21st position

1. Give the five-number summary of the data.

Q1 = 20

Q2 = 25

Q3 = 35

Minimum value = 13

Maximum value = 70

1. Show a boxplot of the data.

A picture containing diagram

Description automatically generated

1. How is a quantile–quantile plot different from a quantile plot?

A graphical tool for comparing the distribution of a sample to a known distribution is the quantile-quantile (QQ) plot. On the other hand, a quantile plot is a graph that compares the sample quantiles to the predicted quantiles of a theoretical distribution. The quantile plot compares the distribution of the sample to a theoretical distribution, whereas the QQ plot is used to determine whether the sample data originates from a known distribution.