Q1-a)

1. **What is predictive analytics?**

Predictive analytics uses statistical techniques, machine learning, and data mining to analyze historical data and predict future outcomes. It helps in making informed decisions and forecasting trends in various domains like marketing, healthcare, and finance.

**b) What are the applications of Association Rule Mining?** Association Rule Mining is used in:

* 1. Market Basket Analysis to understand product associations.
  2. Fraud detection by identifying suspicious patterns.
  3. Cross-selling by recommending related products.
  4. Web usage mining to analyze user behavior.

**c) What is WEKA?**

WEKA (Waikato Environment for Knowledge Analysis) is a data mining software developed at the University of Waikato. It provides tools for machine learning, data preprocessing, classification, regression, and visualization.

**d) Write any two advantages of Big Data.**

* 1. Enables real-time decision-making by analyzing vast data quickly.
  2. Helps businesses gain insights into customer behavior, improving strategies.

**e) Define Regression Analysis.**

Regression analysis estimates relationships between dependent and independent variables. It is used to predict outcomes and identify trends, widely applied in economics, marketing, and risk management.

**f) Write four applications of Data Science.**

* 1. Predictive analytics for business forecasting.
  2. Image and speech recognition in AI systems.
  3. Fraud detection in banking.
  4. Personalized recommendations in e-commerce.

**g) Write any two needs of Machine Learning.**

* 1. Automates repetitive tasks with improved efficiency.
  2. Enhances predictions and classifications using data-driven models.

**h) Write two disadvantages of the EM algorithm.**

* 1. May converge to a local optimum rather than a global one.
  2. Requires significant computational time for large datasets.

**i) What is R?**

R is a programming language used for statistical computing and graphics. It provides various packages and libraries for data analysis, visualization, and machine learning.

**j) Enlist any four tools used for Big Data.**

* 1. Hadoop
  2. Spark
  3. Hive
  4. Apache Kafka

**Q-1-b)**

1. **What is population?**

In big data, "population" refers to the entire set of data points or individuals under study, from which a sample may be drawn for analysis. It represents the total dataset that could be analysed.

1. **Define sample.**

A "sample" in big data refers to a subset of data taken from the larger population. It's used to analyze and make inferences about the entire population without having to process the complete dataset.

1. **What is analytics?**

Analytics is the process of examining data to uncover patterns, insights, and trends. It involves using statistical methods, algorithms, and tools to interpret and make decisions based on the data.

1. **Types of analytics, explain.**

* **Descriptive Analytics:** Focuses on summarizing past data to understand what has happened. Example: reports, dashboards.
* **Diagnostic Analytics:** Explains why something happened by examining historical data. Example: root cause analysis.
* **Predictive Analytics:** Uses data, statistical models, and machine learning to predict future outcomes. Example: forecasting trends.
* **Prescriptive Analytics:** Recommends actions to achieve desired outcomes based on predictions and optimization models. Example: decision-making tools.

1. **Explain the application of big data.**

Big data has various applications across industries:

* **Healthcare:** Improves patient care by analyzing medical records, predicting outbreaks, and personalizing treatments.
* **Finance:** Detects fraud, manages risk, and enhances customer services through real-time data analysis.
* **Retail:** Personalizes marketing, optimizes supply chains, and predicts customer behavior.
* **Manufacturing:** Enhances quality control, predicts equipment failures, and streamlines production processes.
* **Transportation:** Optimizes routes, reduces fuel consumption, and improves traffic management.

1. **Explain digital data and its types.**

Digital data refers to information that is stored and processed in a computer-readable format. It can be categorized into the following types:

* **Structured Data:** Organized and easily searchable in predefined formats, such as databases or spreadsheets (e.g., names, dates).
* **Unstructured Data:** Lacks a specific structure, making it harder to search and analyze, such as text documents, videos, or social media posts.
* **Semi-Structured Data:** Contains both structured and unstructured elements, like JSON or XML files, where the data is organized but not in a traditional database format.

1. **What is predictive analytics?**

Predictive analytics is the use of historical data, statistical algorithms, and machine learning techniques to predict future events or trends. It helps organizations forecast outcomes, identify risks, and make informed decisions by analyzing patterns in past data.

1. **Advantages of big data.**

The advantages of big data include:

* **Better Decision-Making:** Provides data-driven insights to improve business strategies.
* **Enhanced Efficiency:** Optimizes processes, reduces waste, and increases productivity.
* **Personalization:** Enables customized products, services, and marketing.
* **Real-Time Insights:** Helps respond quickly to trends and issues as they arise.
* **Competitive Advantage:** Provides deeper market insights and helps businesses stay ahead of competitors.

1. **Difference between unstructured and structured data.**

* **Structured Data:** Organized and easily searchable in predefined formats, such as tables, rows, and columns. Example: Customer names, transaction records.
* **Unstructured Data:** Lacks a predefined structure, making it harder to search and analyze. Example: Emails, social media posts, multimedia files.

1. **Explain the 5 V's of big data.**

The 5 V's of big data refer to key characteristics that define and differentiate big data:

* **Volume:** The vast amount of data generated and collected from various sources.
* **Velocity:** The speed at which data is generated and processed.
* **Variety:** The different types and formats of data, including structured, semi-structured, and unstructured data.
* **Veracity:** The accuracy and reliability of data.
* **Value:** The usefulness and insights gained from analyzing the data.

1. **Explain the stages of the data analytics life cycle.**

The data analytics life cycle typically includes the following phases:

* **Data Collection:** Gathering raw data from various sources.
* **Data Preparation:** Cleaning and transforming data to ensure quality and consistency.
* **Data Exploration:** Analyzing data to understand patterns, trends, and relationships.
* **Data Modeling:** Applying statistical or machine learning models to make predictions or identify patterns.
* **Data Evaluation:** Assessing the model's performance and accuracy using metrics and validation techniques.
* **Data Visualization:** Creating charts, graphs, and other visual representations to communicate findings.
* **Data Deployment:** Implementing the model or findings in real-world scenarios.
* **Monitoring and Maintenance:** Continuously tracking the performance of the deployed model and making necessary adjustments.

1. **Explain different types of population.**

In statistics and data analysis, different types of populations can be classified based on their characteristics:

* **Finite Population:** A population with a limited number of data points or individuals. Example: The number of students in a specific school.
* **Infinite Population:** A population with an unlimited or uncountable number of data points. Example: The number of possible outcomes from rolling a fair die repeatedly.
* **General Population:** The entire set of individuals or data points of interest in a broad context. Example: All consumers in a country.
* **Target Population:** The specific group within the general population that a study or analysis is focused on. Example: All adults aged 18-35 in a specific city for a marketing study.

1. **What is big data?**

Big data refers to extremely large and complex datasets that are beyond the capabilities of traditional data processing tools to capture, store, manage, and analyze. It typically involves data characterized by the 5 V's: volume, velocity, variety, veracity, and value.

1. **What is data manipulation?**

Data manipulation is the process of adjusting, organizing, or transforming data to make it more suitable for analysis. It involves tasks such as cleaning, filtering, aggregating, and restructuring data to ensure accuracy and improve its usability for generating insights or making decisions.

1. **What is data science?**

Data science is an interdisciplinary field that combines statistical analysis, data processing, and machine learning to extract insights and knowledge from data. It involves collecting, cleaning, analyzing, and interpreting large datasets to support decision-making and solve complex problems across various domains.

1. **Enlist the stages of data science.**

* **Data Collection**
* **Data Preparation (Cleaning & Wrangling)**
* **Data Exploration**
* **Data Modeling**
* **Data Evaluation**
* **Data Visualization**
* **Data Deployment**
* **Monitoring and Maintenance**

1. Explain the process of data analysis.

* Define Objectives: Identify the goals and questions you want to answer with the data.
* Data Collection: Gather relevant data from various sources.
* Data Cleaning: Prepare the data by handling missing values, correcting errors, and removing duplicates to ensure quality and consistency.
* Data Exploration: Use statistical methods and visualizations to understand the data's structure, patterns, and relationships.
* Data Transformation: Convert and format the data as needed for analysis, which may involve aggregating, normalizing, or creating new variables.
* Data Modeling: Apply statistical or machine learning models to analyze the data and test hypotheses or make predictions.
* Interpret Results: Analyze the outcomes of the models to draw conclusions and answer the initial questions.
* Communicate Findings: Present the results through reports, dashboards, or visualizations to stakeholders in a clear and actionable manner.
* Make Decisions: Use the insights gained from the analysis to inform decisions or strategies.

**Q-2-a)**

**a) Explain the advantages and disadvantages of the Apriori algorithm.**

**Advantages:** Efficient in finding frequent item sets and generating association rules in datasets.

**Disadvantages:** Computationally expensive for large datasets as it scans repeatedly.

**b) Explain four data types in R.**

* 1. Numeric: Represents numbers.
  2. Integer: Whole numbers.
  3. Character: Strings of text.
  4. Logical: True or False values.

**c) Explain four types of Correlation.**

* 1. **Positive Correlation:** Both variables increase together.
  2. **Negative Correlation:** One variable increases, the other decreases.
  3. **No Correlation:** Variables have no relationship.
  4. **Perfect Correlation:** One variable perfectly predicts the other.

**d) Explain Statistical Inference with a suitable diagram.** Statistical inference draws conclusions about a population based on sample data. Techniques include hypothesis testing and confidence intervals. (You may need to add a simple diagram here showing sample to population inference).

**Q-3)**

**a) Explain 5 V's of Big Data.**

* 1. **Volume:** Large amount of data.
  2. **Velocity:** Speed of data generation.
  3. **Variety:** Different forms of data.
  4. **Veracity:** Accuracy of data.
  5. **Value:** Insights extracted from data.

**b) Explain the phases of the Data Analytics Life Cycle.**

* 1. **Discovery:** Identifying problems.
  2. **Data Preparation:** Cleaning and organizing data.
  3. **Model Planning:** Choosing algorithms.
  4. **Model Building:** Creating predictive models.
  5. **Deployment:** Implementing solutions.

**c) Explain for loop in R programming with syntax and example.**

Syntax:

for (value in sequence) {

statement

}

Example:

for (i in 1:5) {

print(i)

}

**d) Explain the Naive Bayes Algorithm in detail.**

The Naive Bayes Algorithm is a probabilistic classifier based on Bayes' theorem. It assumes all features are independent and is widely used for spam detection, sentiment analysis, and document classification.

**e) Explain Head() and Tail() functions in the dplyr package.** Head() shows the first few rows of a dataset, while Tail() displays the last few rows. These are used to quickly examine data structure.

**Q-4)**

**a) Explain four applications of Big Data.**

* 1. Fraud detection in banking.
  2. Customer behavior analysis in retail.
  3. Smart city development for infrastructure planning.
  4. Personalized healthcare solutions.

**b) What is population?**

Explain different types of population. Population refers to the complete set of elements being studied. Types include:

* 1. Finite Population (e.g., students in a school).
  2. Infinite Population (e.g., stars in the universe).
  3. Real Population (existing entities).
  4. Hypothetical Population (theoretical cases).

**Q5) Short Notes:**

**a) Advantages of SVM algorithm.**

Support Vector Machine (SVM) is effective for high-dimensional spaces, handles non-linear data using kernel tricks, and works well with clear margins of separation.

**b) Statistical Modeling.**

Statistical modeling involves creating mathematical representations of real-world data. It helps in understanding relationships, making predictions, and supporting decision-making.

**c) Data Analysis**. Data analysis involves inspecting, cleaning, and modeling data to extract insights. Techniques include descriptive, diagnostic, predictive, and prescriptive analytics.

**Q-0) Programming:**

**a) Write a R program to sort vector in ascending and descending order.**

# Define a vector

vec <- c(45, 12, 78, 34, 23)

# Sort in ascending order

ascending <- sort(vec)

# Sort in descending order

descending <- sort(vec, decreasing = TRUE)

# Print results

cat("Original Vector: ", vec, "\n")

cat("Ascending Order: ", ascending, "\n")

cat("Descending Order: ", descending, "\n")

**b) Write a c program to create a simple bar plot of five subject marks.**

# Marks for five subjects

subjects <- c("Math", "Science", "English", "History", "Geography")

marks <- c(85, 90, 78, 88, 76)

# Create bar plot

barplot(marks,

names.arg = subjects,

col = "skyblue",

main = "Marks in Five Subjects",

xlab = "Subjects",

ylab = "Marks",

ylim = c(0, 100))

**c) Write a R program to create a multiplication table.**

# Define a number for the multiplication table

num <- 5

# Generate the multiplication table

table <- num \* 1:10

# Print the multiplication table

cat("Multiplication Table for", num, ":\n")

for (i in 1:10) {

cat(num, "x", i, "=", table[i], "\n")

}

**d) Write a R program to find out number is positive and negative.**

# Input a number

num <- as.numeric(readline(prompt = "Enter a number: "))

# Check if the number is positive, negative, or zero

if (num > 0) {

cat("The number", num, "is positive.\n")

} else if (num < 0) {

cat("The number", num, "is negative.\n")

} else {

cat("The number is zero.\n") }