| Name:<br>Enrolment No: | |
|---|---|

**UPES**

**End Semester Examination, December 2024**

**Course: Data Analysis and Simulations** — **Semester : VII**
**Program: BTech. BioTech** — **Duration : 03 Hours**
**Course Code: HSBT4003** — **Max. Marks: 100**

**Instructions:**

**Section A**

**Short answer questions/ MCQ/T&F**

**(20Qx1.5M= 30 Marks)**

| S. No. | | Marks | COs |
|---|---|---|---|
| **Q 1** | **Multiple Choice Questions:** | | |
| **i)** | In data analysis, what is an outlier?<br>A) The most frequently occurring value in a dataset<br>B) A value that significantly deviates from the rest of the dataset<br>C) The middle value when the data is sorted in ascending order<br>D) The sum of all values in a dataset | **1.5** | **CO1** |
| **ii)** | What is the primary goal of exploratory data analysis (EDA)?<br>A) To generate predictive models<br>B) To summarize the dataset characteristics<br>C) To test hypotheses<br>D) To implement machine learning algorithms | **1.5** | **CO2** |
| **iii)** | What is the main goal of hypothesis testing in data analysis?<br>A) To visualize data trends<br>B) To validate assumptions about a population<br>C) To summarize data distribution<br>D) To predict future outcomes | **1.5** | **CO3** |
| **iv)** | Which statistical measure indicates the spread or dispersion of data?<br>A) Mean<br>B) Standard Deviation<br>C) Median<br>D) Mode | **1.5** | **CO1** |
| **v)** | What is the primary purpose of data visualization?<br>A) To store large datasets<br>B) To represent data visually for easier interpretation<br>C) To clean and preprocess data<br>D) To perform statistical tests | **1.5** | **CO1** |
| **vi)** | What is the role of data cleaning in data analysis?<br>A) To identify trends in the data<br>B) To correct or remove inaccurate, incomplete, or irrelevant data<br>C) To format the data for printing<br>D) To encrypt the dataset | **1.5** | **CO1** |

| | | | |
|---|---|---|---|
| **vii)** | In a histogram, what does the height of each bar represent?<br>A) Frequency<br>B) Data range<br>C) Cumulative percentage<br>D) Mode | **1.5** | **CO2** |
| **viii)** | Which of the following is an example of qualitative data?<br>A. Age<br>B. Height<br>C. Gender<br>D. Income | **1.5** | **CO2** |
| **ix)** | Which Excel function is used to count the number of cells that meet a specified condition?<br>A) COUNTIF<br>B) COUNTA<br>C) COUNT<br>D) SUMIF | **1.5** | **CO2** |
| **x)** | In a box plot, what does the line inside the box represent?<br>A) The mean of the data<br>B) The median of the data<br>C) The maximum value of the data<br>D) The standard deviation of the data | **1.5** | **CO3** |
| **xi)** | In reinforcement learning, the agent learns by:<br>A) Imitating labeled data<br>B) Receiving rewards or penalties based on its actions<br>C) Analyzing clusters in the dataset<br>D) Minimizing the loss function | **1.5** | **CO3** |
| **xii)** | In data analysis, what does "data transformation" refer to?<br>A. The process of storing data in a secure format<br>B. The process of cleaning data to remove errors<br>C. The process of changing the format or structure of data for analysis<br>D. The process of normalizing the data for comparison | **1.5** | **CO3** |
| **xiii)** | Which of the following is not a supervised machine learning algorithm?<br>A) K-means<br>B) Naïve Bayes<br>C) SVM for classification problems<br>D) Decision tree | **1.5** | **CO1** |
| **xiv)** | What is the key difference between supervised and unsupervised learning?<br>A) Supervised learning requires labeled data, while unsupervised learning does not.<br>B) Supervised learning predicts labels, while unsupervised learning discovers patterns.<br>C) Supervised learning is used for classification, while unsupervised learning is used for regression.<br>D) Supervised learning is always more accurate than unsupervised learning. | **1.5** | **CO1** |

| | | | |
|---|---|---|---|
| **xv)** | What type of data is considered unstructured?<br>a) Data in relational databases<br>b) Data in spreadsheets<br>c) Data in CSV files<br>d) Text documents and images | **1.5** | **CO3** |
| **xvi)** | Which of the following techniques is commonly used to handle missing data in a dataset?<br>A) Deleting the rows with missing values<br>B) Filling missing values with the mean or median<br>C) Ignoring the missing data<br>D) All of the above | **1.5** | **CO2** |
| **xvii)** | Which type of chart is most suitable for showing the relationship between two continuous variables?<br>A) Bar chart<br>B) Line chart<br>C) Scatter plot<br>D) Pie chart | **1.5** | **CO3** |
| **xviii)** | Which of the following uses data on some object to predict values for another object?<br>a) Predictive<br>b) Exploratory<br>c) Inferential<br>d) None of the mentioned | **1.5** | **CO2** |
| **xix)** | Which of the following techniques is used to identify the relationship between two variables in data analysis?<br>A) Descriptive Statistics<br>B) Regression Analysis<br>C) Time Series Analysis<br>D) Hypothesis Testing | **1.5** | **CO3** |
| **xx)** | Qualitative data includes-<br>A) Numerical and percentile information of a subject<br>B) Every major and minor detail of a subject<br>C) Both of them<br>D) None of them | **1.5** | **CO1** |
| **Section B**<br>**(4Qx5M=20 Marks)** | | | |
| **Q 2** | Define various types of data with their features and examples. | **5** | **CO1** |
| **Q 3** | Explain Prescriptive data analysis with proper example. | **5** | **CO2** |
| **Q 4** | Discuss the best practices for designing effective visualizations. | **5** | **CO2** |
| **Q 5** | Differentiate between Supervised, Unsupervised and Reinforcement Learning. | **5** | **CO3** |

| | | | |
|---|---|---|---|
| **Section C**<br>**(2Qx15M=30 Marks)** | | | |
| **Q 6** | A healthcare provider wants to implement a system that monitors patients' vital signs (e.g., heart rate, oxygen levels, blood pressure) in real-time to detect abnormalities and send alerts to doctors. They have historically labeled data of vital sign patterns leading to emergencies.<br>a) Discuss the challenges of deploying machine learning models in a real-time healthcare monitoring system.<br>b) Explain how supervised and unsupervised learning can be implemented to ensure the accuracy of predictions in real-time. | **15** | **CO3** |
| **Q 7** | Before creating visualizations in Excel, it is essential to clean and prepare the data. Explain the steps involved in data cleaning, and how proper data cleaning can enhance the quality of visualizations? Support with example. | **15** | **CO2** |
| **Section D**<br>**(2Qx10M=20 Marks)** | | | |
| **Q 8** | Describe the process of training a machine learning model with proper steps and example. | **10** | **CO3** |
| **Q 9** | A company wants to analyze customer feedback from surveys. What visualization techniques would you recommend summarizing the qualitative feedback effectively? | **10** | **CO1** |