KATHMANDU UNIVERSITY
End Semester Examination [C]
December, 2024

Marks Scored:

Level : B.Tech.
Year : II

Course : AICC 202
Semester : I

Exam Roll No. :

Time: 30 mins.

F. M. : 10

Registration No.:

Date : 13-Dec.

SECTION "A"

[20Q. × 0.5 = 10 marks]

Choose and encircle in the most appropriate option from each set of choices.

1. Which of the following statements about the Data Science Hierarchy is FALSE?
   a. Business Intelligence involves deriving insights from historical data.
   b. Data Analytics primarily focuses on cleaning and structuring raw data.
   c. Data Engineering is critical for creating data pipelines and managing large-scale storage.
   d. Data Science involves using predictive modeling to derive insights.

2. Which of the following accurately represents the progression of the Data Science Lifecycle in terms of data preparation and model deployment?
   a. Data Collection → Data Cleaning → Model Evaluation → Data Visualization → Model Deployment
   b. Model Training → Data Visualization → Data Cleaning → Model Deployment → Model Evaluation
   c. Data Wrangling → Model Training → Model Evaluation → Data Collection → Model Deployment
   d. Data Collection → Data Wrangling → Exploratory Data Analysis (EDA) → Model Development → Model Deployment

3. In the context of Data Science Applications, which technique would be most appropriate for grouping customers into different segments based on their purchasing behavior?
   a. Classification b. Regression c. Clustering d. Data Engineering

4. What is the primary goal of Data Discovery in the Data Science process?
   a. To visualize data using graphs and charts
   b. To uncover patterns and insights from raw data before analysis
   c. To clean and remove errors from data
   d. To enrich the data by adding external information

5. Why is Data Cleaning considered a critical step in the Data Science workflow?
   a. It helps to encrypt sensitive data for security purposes.
   b. It organizes data into a structured format for visualization.
   c. It transforms categorical data into numerical data.
   d. It reduces data redundancy and ensures data accuracy for analysis.

6. Feature Selection is used in Data Science to _____:
   a. Add new features to the dataset to improve model performance
   b. Validate the accuracy of each feature before publishing
   c. Remove irrelevant or redundant features to simplify the model
   d. Transform data into a more structured format for reporting

7. What problem is illustrated by the example where Red beats Green, Green beats Blue, and Blue beats Red in Arrow's Impossibility Theorem?
   a. The issue of incomplete data
   b. The violation of transitivity in ranking preferences
   c. The impact of scoring on regression analysis
   d. The failure of binary comparisons in predictive modeling

8. The Body Mass Index (BMI) is designed to indicate whether an individual's weight is _____:
   a. Proportional to their body fat percentage
   b. Correlated to their muscle mass and bone density
   c. Appropriate given their height in meters
   d. Directly related to their physical strength

9. What is the primary reason for using normalization techniques, such as z-scores, in data analysis?
   a. To simplify the computation of binary comparisons
   b. To ensure that variables are comparable in magnitude
   c. To increase the number of outliers in the data
   d. To generate scores with extreme variability

10. According to Tufte's Visualization Aesthetic, what should be maximized to improve the effectiveness of a data visualization?
   a. The amount of chartjunk used
   b. The use of three-dimensional graphics
   c. The data-ink ratio
   d. The number of colors to emphasize data points

11. What is the "Lie Factor" in data visualization, and how should it ideally be maintained?
   a. It measures the distortion in data representation, and it should ideally be close to 1.
   b. It quantifies the aesthetic appeal of a chart, and it should be maximized.
   c. It evaluates the amount of data-ink used, and it should be minimized.
   d. It measures the number of extraneous elements in a graph, and it should be as high as possible.

12. When designing bar graphs, why is it important to start the axes at zero?
   a. To simplify the computation of data statistics
   b. To ensure the data distribution appears more balanced
   c. To maintain graphical integrity and avoid misleading representations
   d. To make the chart visually appealing to viewers

13. What is one of the key advantages of using Box and Whisker Plots for visualizing data distributions?
   a. They provide a 3D view of the data distribution.
   b. They display the median, quartiles, and potential outliers in a concise format.
   c. They use color scales to represent density.
   d. They are designed to show categorical data relationships.

14. What is a major drawback of using stacked area plots for visualizing data trends?
   a. They require a log scale to be effective.
   b. They make it difficult to interpret trends in the middle areas.
   c. They cannot display categorical data.
   d. They only work with two-dimensional data sets.

15. How did John Snow's data map contribute to understanding the 1854 Cholera epidemic?
   a. It visualized temperature variations over time.
   b. It identified a correlation between population density and cholera cases.
   c. It revealed that a contaminated water pump was the source of the outbreak.
   d. It showed the effectiveness of different medical treatments.

16. In the context of Predictive Analysis, why is it important to account for uncertainty in statistical models?
    a. Because predictions are guarantees of future outcomes.
    b. Because statistical models are always perfectly accurate.
    c. Because real-world data is often incomplete and predictions are based on probabilities rather than certainties.
    d. Because predictive models do not rely on historical data.

17. In Prescriptive Analysis, which technique is used to evaluate different decisions and determine the most advantageous outcome?
    a. Descriptive statistics
    b. Correlation analysis
    c. Game theory
    d. Frequency distribution

18. The term "Veracity" in the context of the Three V's of Big Data refers to _____:
    a. The speed at which data is generated and processed
    b. The variety of data formats and types
    c. The volume of data being analyzed
    d. The accuracy and trustworthiness of the data

19. Consider the following Python code snippet for a simple word count MapReduce program:
    a. The `map()` function groups the words by frequency before emitting the key-value pairs.
    b. The `reduce()` function splits the input text into words and assigns a count of 1 to each word.
    c. The `map()` function uses statistical probabilities to assign weight
    d. The `reduce()` function sums up the values associated with each unique word to get the total count.

20. Why is cultivating soft skills important for data scientists in the current operational environment?
    a. They enable data scientists to communicate insights effectively and work collaboratively with diverse teams.
    b. Soft skills help in writing more efficient algorithms.
    c. Soft skills replace the need for technical knowledge in data science.
    d. They are only important for data scientists working in academia.

KATHMANDU UNIVERSITY
End Semester Examination [C]
December, 2024

Level : B.Tech.
Year : II
Time : 2 hrs. 30mins.

13-Dec.

Course : AICC 202
Semester : I
F. M. : 40

SECTION "B"
[6 Q. × 4 = 24 marks]

Attempt *ANY SIX* questions.

1. With a clear diagram, outline the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology. Describe each step and its purpose and provide an example of how a data science project would use CRISP-DM to extract insights from customer purchasing data. [1+3]

2. Explain the importance of data cleaning as a part of the data wrangling process. Describe three common data cleaning techniques and provide practical examples of how each technique is applied to a dataset. [1+3]

3. Explain the concept of data normalization in data analysis. Compare min-max normalization and z-score normalization and describe scenarios where each method would be most appropriate. Provide an example calculation for min-max normalization on a sample data set. [1+2+1]

4. Describe how Anscombe's Quartet illustrates the limitations of relying solely on summary statistics for data analysis. Provide examples of how visualizations can reveal differences between datasets that appear identical based on statistical measures like mean, variance, and correlation. [2 + 2]

5. Explain the difference between Descriptive Analysis and Diagnostic Analytics. Discuss the purpose of each and provide a real-world example of how they can be applied in a business context, such as a retail company analyzing sales data. [2+2]

6. Suppose you are processing a massive dataset of web server logs using MapReduce to determine the most frequently accessed URLs. Describe how you would design the map() and reduce() functions for this task, and explain the importance of the "combiner" function in optimizing performance. [2+2]

7. Discuss the importance of cultivating soft skills for Next-Generation Data Scientists. How do these skills complement technical abilities, and why are they essential for problem-solving and effective communication? [1+1+2]

P.T.O.

SECTION "C"

[2Q. × 8 = 16 marks]

Attempt *ANY TWO* questions.

8. Analyze the given dataset and identify at least 8 common data quality issues (such as missing values, duplicate entries, inconsistent formatting, or outliers). Outline the data wrangling and cleaning tasks you would perform to correct these issues and ensure the final dataset is accurate and ready for analysis. [Note: No coding is required.]

ID	Name	Age	Height(cm)	Weight(kg)	Country	Email
101	Michael	29	182	78	USA	michael@example.com
102	Sarah	-10	165	65	usa	sarah.example.com
103	Raj	45	one hundred seventy	90	India	raj@example.com
104	Anna		170	seventy-five	UK	anna@domain.com
105	John	thirty		82	united kingdom	john@example.com
106	Chloe	27	160	-60	France	chloe@.com
107		33	175		FRANCE	

9. You are preparing a presentation to show the trends in annual sales for three different products over the past five years. You have been given the raw sales data in a table format. Outline the steps you would take to transform this data into an effective and clear visualization. Consider the following:
   a. The type of chart(s) you would use and why. [4]
   b. How you would apply Tufte's principles, such as maximizing the data-ink ratio and minimizing chartjunk. [3]
   c. Any specific techniques you would use to ensure the visualization is easily interpretable by your audience. [1]

10. You have been provided with a dataset containing information about customer purchases, including columns for "Customer ID," "Date of Purchase," "Product Category," "Quantity Purchased," and "Total Revenue." Using principles of data analysis:
   a. Describe how you would summarize the data to provide insights into purchasing patterns. Discuss which measures of central tendency (mean, median, or mode) and dispersion (range, variance, or standard deviation) you would use, and explain why. [3]
   b. Identify at least two potential reasons why certain product categories might have lower sales compared to others. Explain how you would use the data to investigate these reasons. [2]
   c. Recommend two types of visualizations (e.g., bar chart, line plot, histogram) that would help present your findings clearly to stakeholders. Justify your choices based on the type of data and the insights you aim to convey. [3]