KATHMANDU UNIVERSITY
End Semester Examination
March, 2025

Marks Scored:

Level : B.Tech.
Year : II

Course : AICC 202
Semester : I

Exam Roll No. :

Time: 30 mins.

F. M. : 10

Registration No.:

Date : March-21-025

SECTION "A"
[20Q. × 0.5 = 10 marks]

Choose and encircle the most appropriate answer.

1. What is the primary goal of Data Science?
   a. Data entry and storage
   b. Collecting efficient strategies
   c. Practicing knowledge and insights
   d. Extracting knowledge and insights

2. Which of the following is **NOT** a step in the data science process?
   a. Data Collection b. Data Shopping c. Data Storage d. Model Evaluation

3. In data science, the word "distribution" is mostly related to _____?
   a. Error results b. Probability c. Boolean value d. None of above

4. If we have dataframe df with feature salary. Then df_sub = df[df['salary'] > 120000] gives?
   a. Less than 120 K
   b. Boolean output
   c. Higher than 120 K
   d. None of these

5. Which is not representing matplotlib library?
   a. distplot b. regplot c. jointplot d. Tickers

6. Elo ranking method_____ ?
   a. analysing sequences of binary comparisons
   b. updating results computing results on a random basis
   c. performs statistical operations for updating kernel
   d. uses automatic learning process.

7. What is the purpose of exploratory data analysis (EDA)?
   a. To clean the data
   b. To summarize and visualize the data
   c. To deploy the model
   d. To write reports

8. How many pairwise comparisons would there be for an ANOVA with four groups?
   a. 4 b. 6 c. 8 d. 12

9. The question related to "What might happen next" mostly represents ____?
   a. Descriptive analysis
   b. Diagnostic analytics
   c. Predictive analytics
   d. Prescriptive analytics

10. What is the term for a variable that is used to predict another variable?
    a. Independent variable
    b. dependent variable
    c. Target variable
    d. Output variable

11. Which of the following algorithm is **NOT** included in Reinforcement Learning category?
    a. Q-Learning Model
    b. R-Learning Model
    c. TD Learning Model
    d. Gaussian Mixture Model

12. What is the relationship between Data preparation stage and Model Evaluation & Tuning stage?
    a. Modelling
    b. Re-iterate till satisfactory model performance
    c. Update the mathematical model regularly
    d. Data preparation

13. Which metric is commonly used to evaluate classification models?
    a. Accuracy
    b. Mean Squared error
    c. Temperature
    d. R-Squared

14. Training objective (cost function) is a proxy for _____?
    a. Diagnosing bias vs. variance
    b. Organize ML goals
    c. Lower level tasks and debugging
    d. For real world objective

15. A Type II error occurs in classification results when_____?
    a. A true positive is classified as negative
    b. A true negative is classified as positive
    c. A false positive is classified as negative
    d. A classifier correctly predict all instances

16. Artifacts are _____?
    a. Data loss
    b. Server Crashes
    c. Systematic error during processing
    d. Data cleansing

17. Which of the following is **NOT** an application of Spark?
    a. Interactive Analytics
    b. Streaming Processing
    c. Extract, Transform and Load
    d. Text generation

18. Data quality refers to checking data with _____?
    a. Missing values b. Invalid values c. Misleading values d. All of above

19. Which of the following are characteristics of Handoop?
    a. Node and cluster
    b. Distributed storage and processing
    c. Both a and b
    d. None of above

20. Select most suitable answer for this statement "A data scientist in retail might make a good data scientist in IT"?
    a. True
    b. False
    c. Depends upon a data scientist behavior
    d. Not Sure

KATHMANDU UNIVERSITY
End Semester Examination
March, 2025

21 March-025

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

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

SECTION "B"
[6Q. × 4 = 24 marks]

Attempt *ANY SIX* questions.

1. What do you mean by Computational Science, Data Science and Real Science? Write down similarities and difference between Data science vs. Computer Science.

2. How many descriptive statistics parameter you familiar with? Explain with suitable analysis. (you should describe at least five df.methods())

3. What are the significance of p-value? Interpret your use case example to differentiate two groups in your training dataset.

4. What are data visualization perspective? Describe 1-D, 2-D, and 3-D data visualization techniques using suitable examples.

5. Why evaluation metrics are important in machine learning? Describe any five evaluation metrics using real case scenario.

6. Explain data imputation? How you handle missing data? Give example with codes?

7. What are Hadoop, Hive and Spark? What are the similarities and difference between them?

SECTION "C"
[2Q. × 8 = 16 marks]

Attempt *ANY TWO* questions.

8. What is web-scrapping for big data? Develop a case study using data using web-scrapping as a data collection techniques. Your method should use descriptive statistics in the discussion.

9. What is classification model and evaluation metrics? Develop any classification model including problem definition, algorithm development, python code and discussion for future directions?

10. Write short note on: [4Q × 2=8]
    a. Data science lifecycle
    b. Data Types and its application in ML.
    c. BMI distribution
    d. Characteristics of good data scientists.