Avinashilingam Institute for Home Science and Higher Education for Women
Deemed to be University Estd. u/s 3 of UGC Act 1956, Category A by MHRD (now MoE)
Re-accredited with A++ Grade by NAAC. CGPA 3.65/4, Category I by UGC
Coimbatore - 641043, Tamil Nadu, India

Bachelor's Degree Examination - November 2025
V Semester

Class : III UG 2022 Batch Repeater **Time : 3 Hours**
Major : Computer Science **Max. Marks : 100**

21BCSC21 Fundamentals of Data Science

Course Outcomes:

CO1. Students will be able to apply the basic Data Science knowledge on the day to day problems they encounter.
CO2. Students will realize that there are various phases that contribute to the completion of a Data Science Project and can select among the various modeling techniques.
CO3. Students will be able to apply Regression techniques for modeling a data science project.
CO4. Students will be able to apply the Clustering and Association rule mining for modeling a data science project.
CO5. Students can reproduce the knowledge gained and come out with a sample, case study which they come across in their daily life and implement, document and present the same using the R Tool

Part A **10 x 1 = 10**
Choose the Correct Answer

1. In data cleaning, removing duplicate records is known as: CO1 K1
   a. Sampling b. Deduplication c. Normalization d. Aggregation

2. Which visualization in R is most suitable for representing the distribution of a single numeric variable? CO1 K1
   a. Bar Plot b. Histogram c. Line Chart d. Pie Chart

3. Which metric is most suitable for evaluating ranking models? CO2 K1
   a. F1-Score b. ROC curve
   c. Mean Average Precision d. Confusion Matrix

4. Quantifying modelsoundness helps in: CO2 K1
   a. Measuring the training speed
   b. Checking if the model is logically consistent and reliable
   c. Improving code quality
   d. Collecting more data

5. Making predictions using a trained regression model in R is done using: CO3 K1
   a. predict() b. lm_predict() c. predict_lm() d. regression()

6. Which of the following is NOT an assumption of linear regression? CO3 K1
   a. Linearity b. Independence of errors
   c. Homoscedasticity d. Categorical output variable

7. Hierarchical clustering produces: CO4 K1
   a. Only flat clusters b. A tree-like structure called dendrogram
   c. Only one cluster d. Random partitions

8. Which metric evaluates the strength of an association rule beyond random chance? CO4 K1
   a. Confidence b. Support c. Lift d. Entropy

9. Which format is commonly generated by knitr for reports? CO5 K1
   a. .csv b. .Rmd or .HTML c. .exe d. .py

10. Which R package is commonly used to produce milestone documentation? CO5 K1
    a. ggplot2 b. shiny c. knitr d. dplyr

Part B

5 x 6 = 30

Answer ALL questions

Each answer should not exceed 400 words or two pages

11.a. Explain the different stages of the Data Science process with a neat diagram.	CO1 K2
(or)	
11.b. Discuss three key roles in a data science team and their responsibilities.	CO1 K2
12.a. Illustrate three metrics used for evaluating probability models.	CO2 K3
(or)	
12.b. Describe the process of model validation and its significance.	CO2 K3
13.a. Define linear regression and explain its importance in predictive modeling.	CO3 K2
(or)	
13.b. Differentiate between simple linear regression and multiple linear regression.	CO3 K2
14.a. Discuss two real-world applications of clustering.	CO4 K2
(or)	
14.b. Define cluster analysis and explain its importance in data mining.	CO4 K2
15.a. Discuss best practices for writing comments in R scripts and notebooks.	CO5 K2
(or)	
15.b. Write a note on different formats of reports generated by knitr.	CO5 K2

Part C

5 x 12 = 60

Answer ALL questions

Each answer should not exceed 800 words or four pages

16.a. Demonstrate the process of loading, exploring, and visualizing a dataset in R using code snippets.	CO1 K3
(or)	
16.b. Explain the process of working with relational databases in R.	CO1 K3
17.a. Analyze the importance of ranking models in recommendation systems with a case study.	CO2 K4
(or)	
17.b. Evaluate different techniques for solving scoring problems, explaining their strengths and weaknesses.	CO2 K4
18.a. Explain the predict() function in R with examples of single and multiple prediction scenarios.	CO3 K3
(or)	
18.b. Compare residual plots and diagnostic plots for evaluating model performance.	CO3 K3
19.a. Compare hierarchical clustering and non-hierarchical clustering, highlighting their advantages and limitations.	CO4 K3
(or)	
19.b. Explain the Apriori algorithm with a numerical example of mining frequent item sets.	CO4 K3
20.a. Discuss common pitfalls in documentation and suggest solutions.	CO5 K4
(or)	
20.b. Explain the importance of visual storytelling in presenting results to stakeholders.	CO5 K4

****