



Code No.: 22ADC09

**CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY (Autonomous)**  
**B.E. / B.Tech (AI&DS) V Sem (Main) Examination December 2024**

**Introduction to Data Science**

**Time: 3 Hours**

**Max Marks: 60**

**Note:** Answer **ALL** questions from **Part-A** at one place in the same order and **Part-B** (**Internal Choice**)

**Part - A**

**(5Q X 2M = 10 Marks)**

		<b>M</b>	<b>CO</b>	<b>BT</b>
1	Mention the role of summary statistics in data exploration.	(2)	1	1
2	Describe the purpose of the LIME technique in model interpretation.	(2)	2	2
3	Distinguish between linear and logistic regression.	(2)	3	2
4	How is R Markdown used in model documentation?	(2)	4	1
5	What is the significance of word clouds in sentiment analysis?	(2)	5	1

**Part - B**

**(5Q X 10M = 50 Marks)**

		<b>M</b>	<b>CO</b>	<b>BT</b>
6	(a) Explain the lifecycle of a data science project with a neat sketch.	(5)	1	2
	(b) Describe typical problems revealed by data summaries.	(5)	1	3
<b>(OR)</b>				
7	(a) Discuss the key roles and responsibilities in a data science project.	(5)	1	2
	(b) Describe the process of loading data into R from various files.	(5)	1	3
8	(a) How do normalization and rescaling impact the modeling process in data science?	(5)	2	3
	(b) Explain the process of creating a sample group column in a dataset for random sampling.	(5)	2	2
<b>(OR)</b>				
9	(a) Explain common methods for handling missing data in a dataset and their impact on analysis.	(5)	2	2
	(b) Discuss the key steps involved in transforming data for analysis in R.	(5)	2	2
10	(a) Demonstrate how to implement linear regression in R to predict continuous outcomes.	(5)	3	2
	(b) Explain the difference between bagging and random forests in ensemble learning.	(5)	3	4
<b>(OR)</b>				
11	(a) Use R to perform a k-means clustering analysis on a sample dataset and interpret the results.	(5)	3	4
	(b) Explain the use of kernel methods for increasing data separation in a classification problem.	(5)	3	2
12	(a) Explain the role of comments and version control in maintaining running documentation.	(5)	4	2

- (b) Describe the role of the `unnest_tokens` function in tidying text data for analysis. (5) 4 2

**(OR)**

- 13 (a) Describe the role of R Markdown in producing milestone documentation and how it supports model deployment. (5) 4 2
- (b) Explain the functionality of the `gutenbergr` package and its `gutenberg_download()` function for downloading and exploring public domain works from Project Gutenberg. (5) 4 2
- 14 (a) Illustrate how sentiment dictionaries identify positive and negative words in a dataset. (5) 5 2
- (b) Explain how `tf-idf` helps in identifying important words in a document. (5) 5 3
- (OR)**
- 15 (a) Describe how tidy data can be applied to perform sentiment analysis (5) 5 3
- (b) Explain how Zipf's law applies to natural language processing and text analysis. (5) 5 3

\*\*\*\*\*