SRM VALLIAMMAI ENGINEERING COLLEGE
(An Autonomous Institution)
SRM Nagar, Kattankulathur – 603 203

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

QUESTION BANK

II SEMESTER

CP3263 **-** ADVANCED DATA SCIENCE

Regulation – 2023

Academic Year 2024 – 2025

Prepared by

Mrs.V.Prema(Sr.G)

SRM VALLIAMMAI ENGINEERING COLLEGE
(An Autonomous Institution)
SRM Nagar, Kattankulathur-
603203

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

QUESTION BANK

SUBJECT : CP3263 **- DATA SCIENCE**
SEM/YEAR : II/ I

UNIT - I: INTRODUCTION TO DATA SCIENCE			
Data science process - roles, stages in data science project - working with data from files - working with relational databases - exploring data - managing data - cleaning and sampling for modeling and validation - introduction to NoSQL.			
PART – A			
Q. No	Question	BT Level	Competence
1	What is the primary goal of Data Science?	BTL1	Remember
2	**Differentiate** Business Intelligence (BI) and Data Science.	BTL1	Remember
3	**Compare** Data Science and Statistics.	BTL2	Understand
4	**Define** Data Science.	BTL1	Remember
5	**List** out the areas in which Data Science can be applied.	BTL2	Understand
6	What is NoSQL and how is it different from SQL databases?	BTL1	Remember
7	**Compare** Big Data with Data Science.	BTL1	Remember
8	**State**. the concept of data validation	BTL2	Understand
9	**List** out Mention two common file formats used in data science.	BTL4	Analyze
10	Can Data Science Predict the Stock Market? **Examine**.	BTL2	Understand
11	**Discuss** exploratory data analysis (EDA)	BTL3	Apply
12	**Give** Drew Conway's Venn diagram of Data Science.	BTL6	Create
13	**Specify** the life cycle of Data Science.	BTL3	Apply
14	**Illustrate** the use of Data Science with an example.	BTL5	Evaluate
15	**Show** the ways in which decision making and predictions are made in Data Science.	BTL4	Analyze
16	**Differentiate** between supervised and unsupervised learning.	BTL4	Analyze
17	**Analyze Data Science ethics**.	BTL5	Evaluate
18	**Analyze** the roles of Data Science.	BTL3	Apply
19	Why is model validation important?	BTL1	Remember
20	**Develop** a general algorithm for Data Science process.	BTL6	Create
21	List the different sources of data in a data science project.	BTL 3	Apply

22	What is the importance of statistical analysis in data science?	BTL-4	Analyze
23	How does data visualization support decision-making?	BTL 6	Create
24	What is the role of machine learning in data science?	BTL5	Evaluate
PART – B			
1	Explain the complete workflow of a data science project.	BTL-3	Analyze
2	**Describe** life cycle of Data Science with neat diagram.	BTL-3	Analyze
3	Discuss different types of databases used in data science.	BTL-3	Analyze
4	i. **Discuss** nature of data. ii. Give detail description of applications of data.	BTL-2	Understand
5	Explain the various data-cleaning techniques with examples.	BTL-2	Understand
6	Discuss the process of exploratory data analysis and its importance.	BTL-3	Analyze
7	How do you handle high-dimensional data? Discuss feature selection techniques.	BTL-3	Analyze
8	Explain the role of big data in modern data science applications.	BTL-4	Apply
9	How do you evaluate the performance of a machine learning model?	BTL-3	Analyze
10	Extrapolate big data analytics and **Develop** a summary of various applications in the real world scenario.	BTL-4	Apply
11	**Describe** the roles and stages in data science project.	BTL-3	Analyze
12	Discuss a case study where data science was used for business decision-making.	BTL-3	Analyze
13	**Explain relational databases iin detail**	BTL-4	Apply
14	What is data sampling? Explain various sampling techniques.	BTL-4	Apply
15	Explain different methods to deal with missing and corrupted data.	BTL-4	Apply
16	How does machine learning contribute to predictive analytics?	BTL-3	Analyze
17	Compare and contrast SQL and NoSQL databases.	BTL-3	Analyze

UNIT II MODELING METHODS

Choosing and evaluating models - mapping problems to machine learning, evaluating clustering models, validating models - cluster analysis - K-means algorithm, Naïve Bayes - Memorization Methods - Linear and logistic regression - unsupervised methods.

	PART – A		
Q. No	Questions	BT Level	Competence
1	What is clustering in machine learning?	BTL1	Remember
2	Write the K-means algorithm in simple terms.	BTL1	Remember
3	Define Naïve Bayes, and where is it used?	BTL2	Understand
4	Differentiate between linear and logistic regression.	BTL1	Remember
5	What are supervised and unsupervised learning methods?	BTL2	Understand
6	How do you choose the best model for a dataset?	BTL1	Remember
7	List the memorization methods in machine learning?	BTL1	Remember
8	Why is model evaluation necessary?	BTL2	Understand
9	Difference overfitting and underfitting.	BTL4	Analyze
10	Define loss function in machine learning?	BTL2	Understand
11	What are the key metrics used for evaluating classification models?	BTL3	Apply
12	Define bias-variance tradeoff.	BTL6	Create
13	What is cross-validation?	BTL3	Apply
14	List the significance of the training and testing dataset split.	BTL5	Evaluate
15	What are precision and recall in classification models?	BTL4	Analyze
16	What is the role of feature scaling?	BTL4	Analyze
17	Define decision trees and their applications.	BTL5	Evaluate
18	What is meant by ensemble learning?	BTL3	Apply
19	Analyse the role of hyperparameter tuning in model optimization.	BTL1	Remember
20	What are kernel methods in SVM?	BTL6	Create
21	Write principal component analysis (PCA).	BTL 3	Apply
22	What are the advantage of principal component analysis (PCA).	BTL-4	Analyze
23	Differentiate between batch and online learning.	BTL 6	Create
24	What is the significance of model interpretability?	BTL5	Evaluate
	PART – B		
1	Explain the different types of regression models in machine learning.	BTL1	Remember
2	Discuss the role of clustering algorithms in unsupervised learning.	BTL1	Remember
3	Compare Naïve Bayes and Decision Tree classifiers.	BTL1	Remember
4	Explain in detail how K-means clustering works.	BTL2	Understand
5	How do you evaluate a machine learning model's performance?	BTL2	Understand
6	Explain logistic regression and its application in binary	BTL3	Apply

	classification.		
7	Describe different techniques for feature selection.	BTL4	Analyze
8	Discuss overfitting and how to prevent it.	BTL4	Analyze
9	What is ensemble learning? Explain different ensemble techniques.	BTL5	Evaluate
10	Discuss different loss functions used in deep learning.	BTL6	Create
11	Explain the steps involved in choosing an appropriate machine learning model.	BTL4	Analyze
12	How is hyper parameter tuning performed in machine learning models?	BTL1	Remember
13	Discuss the application of support vector machines (SVM).	BTL3	Apply
14	Explain principal component analysis (PCA) with an example.	BTL2	Understand
15	Compare decision trees and random forests.	BTL4	Analyze
16	How do you optimize a model using cross-validation?	BTL4	Analyze
17	Discuss a real-world case study where machine learning models were applied.	BTL5	Evaluate

UNIT - III: INTRODUCTION TO R			
Reading and getting data into R - ordered and unordered factors - arrays and matrices - lists and data frames - reading data from files - probability distributions - statistical models in R - manipulating objects - data distribution.			
Q. No	Questions	BT Level	Competence
PART – A			
1	What are lists in R?	BTL1	Remember
2	How do you read data from files in R?	BTL1	Remember
3	List the significance of probability distributions in R.	BTL2	Understand
4	What is a data frame in R?	BTL1	Remember
5	How are statistical models implemented in R?	BTL2	Understand
6	Define ordered and unordered factors in R.	BTL1	Remember
7	What is an array in R?	BTL1	Remember
8	How do you manipulate objects in R?	BTL2	Understand
9	List the concept of matrices in R.	BTL4	Analyze
10	What are different data types available in R?	BTL2	Understand
11	Write the different ways to install and load packages in R?	BTL3	Apply
12	How does R handle missing data?	BTL6	Create
13	What is the difference between a vector and a list in R?	BTL3	Apply

14	Illustrate merge two data frames in R.	BTL5	Evaluate
15	Point out the use of apply functions in R?	BTL4	Analyze
16	How do you perform correlation analysis in R?	BTL4	Analyze
17	What are logical vectors in R?	BTL5	Evaluate
18	Write the use of tapply in R?	BTL3	Apply
19	List the role of statistical functions in R.	BTL1	Remember
20	Can you create a histogram in R?	BTL6	Create
21	What is the difference between lapply and sapply?	BTL 3	Apply
22	Analysis the subset data in R?	BTL-4	Analyze
23	What is the significance of ggplot2 in R?	BTL 6	Create
24	How do you export data from R?	BTL5	Evaluate

PART – B

1	Discuss different data structures available in R.	BTL 1	Remember
2	Explain data manipulation techniques in R.	BTL 4	Analyze
3	How do you implement probability distributions in R?	BTL 1	Remember
4	Discuss different statistical models that can be implemented in R.	BTL 1	Remember
5	Explain the significance of data visualization in R.	BTL 2	Understand
6	How does R handle large datasets efficiently?	BTL 5	Evaluate
7	Discuss the various control structures in R.	BTL 2	Understand
8	Explain different ways of reading and writing data in R.	BTL 6	Create
9	Compare and contrast base R plotting and ggplot2.	BTL 1	Remember
10	Discuss how regression analysis can be performed in R.	BTL 4	Analyze
11	Explain time series analysis in R.	BTL 2	Remember
12	Discuss different statistical hypothesis testing methods in R.	BTL 4	Analyze
13	How do you implement machine learning algorithms in R?	BTL 3	Apply
14	Explain data wrangling techniques in R.	BTL 3	Apply
15	Discuss the concept of functional programming in R.	BTL 4	Analyze
16	Explain different clustering methods implemented in R.	BTL 3	Apply
17	How does R integrate with other programming languages?	BTL 3	Apply

UNIT - IV: MAP REDUCE

Introduction - distributed file system - algorithms using map reduce, Matrix-Vector Multiplication by Map Reduce - Hadoop - Understanding the Map Reduce architecture - Writing Hadoop Map Reduce Programs - Loading data into HDFS - Executing the Map phase - Shuffling and sorting - Reducing phase execution.

Q.No	Questions	BT Level	Competence
	PART – A		
1	What is Hadoop?	BTL1	Remember
2	Draw the MapReduce architecture.	BTL1	Remember
3	How does MapReduce work in Hadoop?	BTL2	Understand
4	What is HDFS?	BTL1	Remember
5	List the role of the mapper in MapReduce.	BTL2	Understand
6	What is shuffling and sorting in MapReduce?	BTL1	Remember
7	How does the reducer function work?	BTL1	Remember
8	Describe the significance of the partitioner in MapReduce?	BTL2	Understand
9	What is a combiner in Hadoop?	BTL4	Analyze
10	Define job scheduling in Hadoop.	BTL2	Understand
11	What is speculative execution in Hadoop?	BTL3	Apply
12	Illustrate the significance of YARN in Hadoop.	BTL6	Create
13	What are the different file formats supported by Hadoop?	BTL3	Apply
14	Point out the key benefits of MapReduce?	BTL5	Evaluate
15	Write the input splits in Hadoop?	BTL4	Analyze
16	Illustrate the word count example in MapReduce.	BTL4	Analyze
17	Name the input splits in Hadoop?	BTL5	Evaluate
18	What is a sequence file in Hadoop?	BTL3	Apply
19	How Map Reduce can be used for data aggregation.	BTL1	Remember
20	What are the advantages of distributed computing?	BTL6	Create
21	Define matrix-vector multiplication in MapReduce?	BTL-2	Understand
22	Write the difference between Hadoop 1.x and Hadoop 2.x.	BTL-1	Remember
23	What is the role of NameNode and DataNode?	BTL-1	Remember
24	How does replication work in HDFS?	BTL-1	Remember

<table>
<tr><td colspan="4" align="center">**PART – B**</td></tr>
<tr><td>1</td><td>Explain the architecture and working of MapReduce.</td><td>BTL1</td><td>Remember</td></tr>
<tr><td>2</td><td>Discuss the components of the Hadoop ecosystem.</td><td>BTL1</td><td>Remember</td></tr>
<tr><td>3</td><td>How is data stored and processed in HDFS?</td><td>BTL1</td><td>Remember</td></tr>
<tr><td>4</td><td>Explain the significance of MapReduce in big data processing.</td><td>BTL2</td><td>Understand</td></tr>
<tr><td>5</td><td>. Discuss different file formats used in Hadoop.</td><td>BTL2</td><td>Understand</td></tr>
<tr><td>6</td><td>Explain different data partitioning techniques in MapReduce.</td><td>BTL3</td><td>Apply</td></tr>
<tr><td>7</td><td>How does MapReduce optimize large-scale computations?</td><td>BTL4</td><td>Analyze</td></tr>
<tr><td>8</td><td>Explain different job scheduling techniques in Hadoop.</td><td>BTL4</td><td>Analyze</td></tr>
<tr><td>9</td><td>Discuss the challenges in implementing MapReduce.</td><td>BTL5</td><td>Evaluate</td></tr>
<tr><td>10</td><td>Explain MapReduce programming with a real-world example.</td><td>BTL6</td><td>Create</td></tr>
<tr><td>11</td><td>Discuss the role of YARN in Hadoop.</td><td>BTL3</td><td>Apply</td></tr>
<tr><td>12</td><td>Compare and contrast batch processing and real-time processing in big data.</td><td>BTL1</td><td>Remember</td></tr>
<tr><td>13</td><td>How does MapReduce perform fault tolerance?</td><td>BTL4</td><td>Analyze</td></tr>
<tr><td>14</td><td>Explain different optimization techniques used in Hadoop.</td><td>BTL2</td><td>Understand</td></tr>
<tr><td>15</td><td>How is security handled in Hadoop?</td><td>BTL 4</td><td>Analyze</td></tr>
<tr><td>16</td><td>Discuss the impact of Hadoop in data-driven decision-making.</td><td>BTL 3</td><td>Apply</td></tr>
<tr><td>17</td><td>Explain how Hadoop integrates with cloud computing platforms.</td><td>BTL 3</td><td>Apply</td></tr>
</table>

UNIT - V: DATA VISUALIZATION

Documentation and deployment - producing effective presentations - Introduction to graphical analysis - plot() function - displaying multivariate data - matrix plots - multiple plots in one window - exporting graph using graphics parameters - Case studies.

PART – A

Q. No	Questions	BT Level	Competence
1	What is the plot() function in R?	BTL1	Remember
2	Define matrix plots	BTL1	Remember
3	Give the advantages of weather forecasting.	BTL2	Understand
4	List the importance of exporting graphs.	BTL1	Remember

5	Illustrate the role of graphical parameters in visualization?	BTL2	Understand
6	What is meant by documentation in data visualization?	BTL1	Remember
7	How can we make effective presentations using graphs?	BTL1	Remember
8	What is graphical analysis?	BTL2	Understand
9	List the different types of charts used in data visualization?	BTL4	Analyze
10	Difference between a histogram and a bar chart?	BTL2	Understand
11	How do you customize a scatter plot in R?	BTL3	Apply
12	What are the major libraries used for data visualization in Python?	BTL6	Create
13	Point out the significance of heatmaps?	BTL3	Apply
14	Write the different types of visual encodings used in charts?	BTL5	Evaluate
15	How do you create an interactive dashboard?	BTL4	Analyze
16	Describe the role of color theory in visualization?	BTL4	Analyze
17	Can Data Science be used in Stock Market Analysis? **Justify**.	BTL5	Evaluate
18	How can we display multivariate data graphically?	BTL3	Apply
19	Write about the data visualization help in storytelling?	BTL1	Remember
20	Define box plots and why are they useful?	BTL6	Create
21	Write different types of line charts.	BTL 3	Apply
22	What is an area chart and when should it be used?	BTL-4	Analyze
23	How does data preprocessing impact visualization?	BTL 6	Create
24	What are some common mistakes in data visualization?	BTL5	Evaluate
PART – B			
1	Discuss the significance of data visualization in analytics.	BTL1	Remember
2	Compare and contrast different types of data visualization techniques	BTL1	Remember
3	Explain various plotting functions in R.	BTL1	Remember
4	Discuss the role of multivariate data visualization in data science.	BTL2	Understand
5	Explain the concept of graphical analysis and its applications.	BTL2	Understand
6	How do we produce effective presentations using visualization?	BTL3	Apply
7	Explain the importance of case studies in data visualization.	BTL4	Analyze
8	How does documentation support data visualization workflows?	BTL4	Analyze
9	Explain how to create a dashboard for business intelligence.	BTL5	Evaluate
10	Compare and contrast static and interactive visualizations.	BTL6	Create
11	Explain the impact of 3D visualizations in analytics.	BTL4	Analyze
12	Discuss the advantages of using Python for data visualization	BTL1	Remember
13	How does storytelling impact decision-making through visualizations?	BTL3	Apply
14	Discuss the best practices for designing effective visualizations.	BTL2	Understand
15	Explain how geospatial visualization is used in real-world applications.	BTL4	Analyze

| 16 | How do interactive graphs enhance user experience? | BTL4 | Analyze |
| 17 | Discuss a case study where data visualization improved business insights. | BTL5 | Evaluate |