



SRM VALLIAMMAI ENGINEERING COLLEGE

(An Autonomous Institution)

SRM Nagar, Kattankulathur – 603 203



DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

QUESTION BANK



VII SEMESTER

OAD101 FOUNDATION OF DATA SCIENCE AND ANALYTICS

Regulation – 2023

Academic Year 2026 – 2027 (ODD SEMESTER)

Prepared by

P.Jebilla, Assistant Professor (O.G)/AI&DS



SRM VALLIAMMAI ENGINEERING COLLEGE

(An Autonomous Institution)

SRM Nagar, Kattankulathur – 603 203



DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

QUESTION BANK

SUBJECT: OAD101- FOUNDATION OF DATA SCIENCE

SEM / YEAR: VII SEMESTER/ FINAL YEAR

UNIT 1 INTRODUCTION			
Data Science: Benefits and uses – facets of data - Data Science Process: Overview – Defining research goals – Retrieving data – data preparation - Exploratory Data analysis – build the model – presenting findings and building applications - Data Mining - Data Warehousing –Basic statistical descriptions of Data.			
PART – A			
Q.No	Question	BT Level	Competence
1	Mention how Data Science helps in making data-driven decisions.	BTL 2	Understand
2	State the characteristics (5Vs) of Big Data.	BTL 2	Understand
3	What is machine Learning?	BTL 1	Remember
4	Define Data Mining?	BTL 1	Remember
5	List the major sources from which big data is generated.	BTL 1	Remember
6	Mention the categories of data.	BTL 2	Understand
7	List some of the application domains of data science.	BTL 1	Remember
8	What is structured data? Give some examples.	BTL 1	Remember
9	Define the term 'retrieving data' in the context of Data Science.	BTL 1	Remember
10	Why machine-generated data is crucial in modern analytics?	BTL 2	Understand
11	Why the data is to be cleaned?	BTL 2	Understand
12	List the phases involve in the data science process.	BTL 1	Remember
13	Define data cleaning.	BTL 1	Understand
14	Define Project Charter and list its objectives.	BTL 1	Remember
15	List the components of a Project Charter.	BTL 1	Remember

16	List any four data visualization techniques used in Data Science.	BTL 1	Understand
17	Mention any four challenges in collecting real-time data	BTL 2	Understand
18	Define data warehouse, DataMart and data lake.	BTL 2	Understand
19	State any four factors influencing model selection	BTL 2	Understand
20	What is a dummy variable?	BTL 1	Remember
21	What do you meant by exploratory data analysis?	BTL 1	Remember
22	List out the methods for combining data from different table.	BTL 1	Remember
23	State two applications of predictive modeling.	BTL 2	Understand
24	On what factors the modeling technique is being selected.	BTL 2	Understand

PART – B

Q.No	Question	BT Level	Competence
1	Analyze the applications of Data Science in healthcare, finance, manufacturing, education and smart cities with suitable examples.(16)	BTL 4	Analyze
2	Explain the overview of the data science process.(16)	BTL 3	Apply
3	Apply the concepts of Project Charter to formulate a project plan for a Data Science application by identifying the problem statement, objectives, scope, deliverables, and constraints.	BTL 3	Apply
4	Classify the different categories of data used in Data Mining with suitable examples.(16)	BTL 3	Apply
5	Analyze how setting a clear research aim influences the success and direction of a data science project.(16)	BTL 4	Analyze
6	Explain various techniques used for retrieving data from multiple sources in a Data Science project.(16)	BTL 4	Analyze
7	Design a detailed workflow for preparing raw datasets for analysis by incorporating data collection, cleaning, transformation, integration, and feature engineering. Justify each stage.(16)	BTL 6	Create
8	Analyze various data cleansing techniques and recommend appropriate methods for improving data quality in different scenarios.(16)	BTL 4	Analyze
9	Apply ETL concepts to design a workflow for integrating data collected from multiple heterogeneous sources.	BTL 3	Apply
10	Develop a Data Science solution for analyzing smart city data collected from multiple sensors.	BTL 6	Create
11	Explain the steps involved in model building with suitable diagrams.	BTL 4	Analyze
12	Analyze different facets of data and explain their significance in Data Science.	BTL 4	Analyze
13	Analyze the role of Exploratory Data Analysis (EDA) in improving data understanding and model performance.	BTL 4	Analyze
14	Explain briefly about Data science and its life cycle.	BTL 4	Analyze
15	Analyze how Data Science differs from Big Data in terms of purpose, processing techniques, and outcomes.	BTL 4	Analyze
16	Evaluate Cloud Computing and Big Data technologies by comparing their features, advantages, limitations, and applications in Data Science.	BTL 5	Evaluate

17	Analyze the impact of Big Data technologies on modern Data Science applications with suitable case studies.	BTL 4	Analyze
----	---	-------	---------

UNIT 2 DATA MANIPULATION

Jupyter Notebook - IPython Magic Commands - NumPy Arrays-Universal Functions –Aggregations – Computation on Arrays – Fancy Indexing – Sorting arrays – Structured data –Data manipulation with Pandas – Data Indexing and Selection – Handling missing data –Hierarchical indexing – Combining datasets – Aggregation and Grouping – String operations – Working with time series – High performance

PART – A

Q.No	Question	BT Level	Competence
1	What is the difference between Python Shell and Jupyter Notebook in terms of their interface and functionality?	BTL 1	Remember
2	Explain the purpose of IPython magic commands and provide an example	BTL 1	Remember
3	What is the purpose of the %timeit magic command in IPython?	BTL 2	Understand
4	Explain the process of sorting arrays in NumPy	BTL 2	Understand
5	Define NumPy arrays and explain their significance in data manipulation tasks	BTL 1	Remember
6	Differentiate np.zeros() and np.ones().	BTL 1	Remember
7	State the role of Pandas in data manipulation and analysis.	BTL 1	Remember
8	What is a universal function (ufunc) in NumPy?	BTL 1	Remember
9	What is broadcasting in NumPy? State one advantage.	BTL 1	Remember
10	List any four NumPy aggregation functions.	BTL 1	Understand
11	What is the difference between np.sum() and np.nansum()?	BTL 2	Understand
12	How does fancy indexing differ from basic indexing in NumPy arrays?	BTL 2	Understand
13	Define a structured array in NumPy.	BTL 1	Remember
14	Define a Pandas Series object with a syntax example.	BTL 1	Remember
15	What is a Pandas DataFrame? How is it constructed from a dictionary?	BTL 1	Remember
16	Differentiate between loc and iloc.	BTL 2	Understand
17	How is missing data represented in Pandas?	BTL 1	Remember
18	Define Hierarchical Indexing (MultiIndex) in Pandas.	BTL 1	Remember
19	What is the purpose of the pd.concat() function? How does it differ from pd.merge()?	BTL 2	Understand
20	List any three vectorized string methods available in Pandas for string operations.	BTL 1	Remember
21	What is a Pandas Timestamp? How is a date range created using pd.date_range()?	BTL 1	Remember
22	What is vectorization in Pandas? Why is it faster than traditional loops?	BTL 2	Understand
23	What is the purpose of the groupby() function in Pandas? Give an example.	BTL 2	Understand

24	What are indexers in Pandas?	BTL 2	Understand
----	------------------------------	-------	------------

PART – B			
Q.No	Question	BT Level	Competence
1	Evaluate various NumPy array creation techniques and justify their suitability for different data analysis applications with suitable examples.	BTL 5	Evaluate
2	Design a complete data preprocessing workflow using NumPy and Pandas for a real-world dataset, including handling missing values, indexing, grouping, and aggregation.	BTL 6	Create
3	Apply NumPy Universal Functions (ufuncs) to perform mathematical and statistical operations on arrays. Illustrate with suitable examples.	BTL 3	Apply
4	Analyze the role of IPython Magic Commands in improving code execution, debugging, and performance in Jupyter Notebook with suitable examples.	BTL 4	Analyze
5	Analyze aggregation and grouping operations on a real dataset	BTL 4	Analyze
6	Apply NumPy broadcasting rules to perform arithmetic operations on multidimensional arrays and interpret the output.	BTL 3	Apply
7	Analyze indexing, slicing, fancy indexing, and sorting techniques in NumPy with suitable examples.	BTL 4	Analyze
8	Apply structured arrays in NumPy to create and manipulate compound datasets for real-world applications.	BTL 3	Apply
9	Analyze techniques for handling missing data in Pandas and justify the selection of appropriate methods for different scenarios.	BTL 4	Analyze
10	Explain high-performance computation techniques in NumPy and Pandas using vectorization, pd.eval(), and query() with examples.	BTL 4	Analyze
11	Analyze the efficiency of NumPy arrays and Python lists with respect to execution time, memory usage, and computational performance.	BTL 4	Analyze
12	Explain the Pandas Series and DataFrame objects with suitable examples.	BTL 4	Analyze
13	Apply data indexing and selection techniques using .loc[], .iloc[], Boolean indexing, and conditional filtering on a sample dataset.	BTL 3	Apply
14	Analyze commonly used vectorized string operations in Pandas and demonstrate their application using a sample dataset.	BTL 4	Analyze

15	Analyze Pandas time series features such as Timestamps, Time Deltas, Date Ranges, Resampling, and Shifting with suitable examples.	BTL 4	Analyze
16	Evaluate concat(), merge(), and join() methods in Pandas for integrating datasets and justify the most appropriate method for different scenarios.	BTL 5	Evaluate
17	Design a complete data-analysis workflow using NumPy and Pandas that loads a dataset, handles missing values, performs hierarchical grouping and aggregation, and derives meaningful insights; justify each design decision.	BTL 6	Create

UNIT 3 - MACHINE LEARNING

The modeling process - Types of machine learning - Supervised learning - Unsupervised learning - Semi-supervised learning- Classification, regression - Clustering – Outliers and Outlier Analysis .

PART – A

Q.No	Question	BT Level	Competence
1.	Define Machine Learning.	BTL 1	Remember
2	What is the modeling process in Data Science?	BTL 1	Remember
3	List the steps involved in building a machine learning model.	BTL 1	Remember
4	What is model selection? Why is it important?	BTL 2	Understand
5	What is meant by training and testing a model?	BTL 2	Understand
6	Define overfitting and underfitting in a machine learning model.	BTL 2	Understand
7	List the different types of machine learning	BTL 1	Remember
8	What is supervised learning? Give an example.	BTL 2	Understand
9	What is unsupervised learning? Give an example.	BTL 2	Understand
10	Differentiate between supervised learning and unsupervised learning.	BTL 2	Understand
11	Why is semi-supervised learning useful when labeled data is scarce?	BTL 2	Understand
12	What is classification in machine learning? Give an example.	BTL 2	Understand
13	Define regression. How does it differ from classification.	BTL 2	Understand
14	List any four supervised learning algorithms	BTL 1	Remember
15	Differentiate between classification and regression.	BTL 2	Understand
16	State any two performance measures used to evaluate classification models.	BTL 1	Remember
17	State any two performance measures used to evaluate regression models.	BTL 1	Remember
18	What is clustering? Give an example application	BTL 2	Understand
19	What is K-Means clustering	BTL 1	Remember
20	Why is feature selection important in the modeling process?	BTL 2	Understand
21	Define an outlier in a dataset	BTL 1	Remember
22	Define Outlier Analysis and mention its significance.	BTL 2	Understand
23	List any two techniques used for detecting outliers	BTL 1	Remember

24	Explain how hierarchical clustering differs from K-means clustering.	BTL 2	Understand
PART – B			
Q.No	Question	BT Level	Competence
1	Explain the steps involved in the data science modeling process with a neat diagram.	BTL 4	Analyze
2	Evaluate various supervised learning techniques and recommend suitable applications for each with justification.	BTL 5	Evaluate
3	Analyze unsupervised learning techniques and explain their applications in clustering and pattern discovery.	BTL 4	Analyze
4	Analyze semi-supervised learning and explain how it combines labeled and unlabeled data to improve learning performance with a suitable example.	BTL 4	Analyze
5	Analyze supervised, unsupervised, and semi-supervised learning techniques by comparing their characteristics, advantages, limitations, and applications.	BTL 4	Analyze
6	Design a classification model for email spam detection and justify the selection of an appropriate classification technique.	BTL 6	Create
7	Apply regression techniques to solve a predictive analytics problem and compare linear regression with logistic regression using suitable examples.	BTL 3	Apply
8	Apply the complete machine learning modeling process to develop a predictive solution for a business problem and justify each stage involved.	BTL 3	Apply
9	Analyze the different types of machine learning and discuss their advantages, limitations, and real-world applications.	BTL 4	Analyze
10	Apply the Naïve Bayes classifier to develop a text classification model and explain the steps involved with a suitable example.	BTL 3	Apply
11	Apply Decision Tree and Random Forest classifiers to solve a suitable classification problem and explain their working principles with examples.	BTL 3	Apply
12	Apply K-Nearest Neighbor (KNN) and Logistic Regression algorithms to solve suitable classification problems and illustrate their implementation with examples.	BTL 3	Apply
13	Analyze different outlier detection techniques and recommend suitable methods for various types of datasets with appropriate justification.	BTL 4	Analyze
14	Discuss the impact of outliers on machine learning models and explain the techniques to handle them.	BTL 4	Analyze
15	Analyze the statistical, distance-based, and density-based outlier methods used for outlier detection.	BTL 4	Analyze
16	Evaluate the performance of K-means, Hierarchical, and DBSCAN clustering algorithms for a given dataset and justify which one is most suitable	BTL 5	Evaluate
17	Design a complete machine learning pipeline for a real-world application (such as customer segmentation or fraud detection) incorporating data preprocessing, model development, classification/clustering, outlier analysis, model evaluation, and deployment. Justify each design decision.	BTL 6	Create

UNIT 4 - DATA VISUALIZATION

Matplotlib – Simple line plots – Simple scatter plots – visualizing errors – density and contour plots – Histograms – legends – colors – subplots – text and annotation – customization – three dimensional plotting - Geographic Data with Basemap - Visualization with Seaborn

PART – A

Q.No	Question	BT Level	Competence
1	What is visualization?	BTL 1	Remember
2	What is Matplotlib used for?	BTL 2	Understand
3	Why do we import Matplotlib in Python?	BTL 1	Remember
4	Write a short note on Pyplot.	BTL 1	Remember
5	State any four applications of Basemap in geographic data visualization.	BTL 2	Understand
6	What is the use of Basemap in Python?	BTL 2	Understand
7	Name any four plots available in Seaborn.	BTL 1	Remember
8	List any four types of error bars.	BTL 2	Understand
9	What is a contour plot?	BTL 1	Remember
10	What is Seaborn?	BTL 1	Remember
11	Define Histogram. State its purpose.	BTL 2	Understand
12	What is the use of Python Seaborn library?	BTL 2	Understand
13	What do you understand by annotating text?	BTL 1	Remember
14	What is the purpose of format strings in Matplotlib?	BTL 1	Remember
15	Write a note on three-dimensional plotting.	BTL 1	Remember
16	What are error bars used for?	BTL 2	Understand
17	State two applications of histograms?	BTL 2	Understand
18	Write the syntax for creating a simple line plot using Matplotlib.	BTL 1	Remember
19	What is the difference between Matplotlib and Pyplot?.	BTL 2	Understand
20	Write the syntax for creating a simple line plot?	BTL 1	Remember
21	What are scatter plots?	BTL 1	Remember
22	List any four visualization techniques used in Exploratory Data Analysis (EDA).	BTL 1	Remember
23	What is the use of density plot?	BTL 2	Understand
24	What is the significance of density and contour plots in EDA?	BTL 2	Understand

PART – B

Q.No	Question	BT Level	Competence
1	Explain the architecture of Matplotlib and describe the role of the Pyplot module in creating visualizations with suitable examples.	BTL 4	Analyze
2	Illustrate various plotting techniques in Matplotlib, including line plots, scatter plots, histograms, and bar charts, with suitable examples.	BTL 3	Apply
3	Analyze various visualization techniques suitable for categorical and numerical datasets and recommend appropriate charts for different analytical scenarios..	BTL 4	Analyze

4	Apply Seaborn to visualize statistical relationships using suitable plots and explain the generated outputs with Python code examples.	BTL 3	Apply
5	Design an effective visualization for a given dataset using line plots, scatter plots, and annotations, and justify your choice of visualization techniques.	BTL 6	Create
6	Analyze Seaborn distribution plots such as histogram, KDE plot, box plot, and violin plot with suitable examples.	BTL 4	Analyze
7	Analyze Matplotlib and Seaborn by comparing their features, advantages, limitations, and suitable application scenarios.	BTL 4	Analyze
8	Explain various customization techniques in Matplotlib, including colors, legends, labels, annotations, and subplots, with examples.	BTL 4	Analyze
9	Evaluate the role of Matplotlib in scientific data visualization and justify its importance in Python-based data analysis.	BTL 5	Evaluate
10	Analyze different error visualization techniques in Matplotlib and explain their significance in representing uncertainty using suitable examples.	BTL 4	Analyze
11	Apply scatter plots and error bar plots in Matplotlib to visualize a given dataset and interpret the generated results.	BTL 3	Apply
12	Apply Basemap to visualize geographic information such as maps, country boundaries, and location-based datasets using suitable Python programs.	BTL 3	Apply
13	Apply customization features in Matplotlib, including legends, colors, annotations, and axis formatting, to improve visualization quality with suitable examples.	BTL 3	Apply
14	Evaluate the effectiveness of Matplotlib and Seaborn in creating statistical and scientific visualizations. Justify the selection of the appropriate library for different visualization tasks, including contour plots, error bars, histograms, and distribution plots, with suitable examples.	BTL 5	Evaluate
15	Analyze the role of density plots, histograms, subplots, and Seaborn visualizations in Exploratory Data Analysis (EDA).	BTL 4	Analyze
16	Explain three-dimensional plotting with example programs. (16)	BTL 4	Analyze
17	Design a Python-based visualization solution using Seaborn to generate pair plots for a real-world dataset and interpret the relationships revealed by the visualization.	BTL 6	Create

UNIT 5 HANDLING LARGE DATA

Problems - techniques for handling large volumes of data - programming tips for dealing with large data sets- Case studies: Predicting malicious URLs, Building a recommender system - Tools and techniques needed - Research question - Data preparation - Model building – Presentation and automation.

PART – A

Q.No	Question	BT Level	Competence
1	Define big data and give an example of a large dataset.	BTL 1	Remember
2	List any four problems associated with handling large volumes of data.	BTL 1	Remember
3	What is meant by the term 'large data volume' in the context of data science?	BTL 1	Remember
4	Define chunk processing in handling large datasets.	BTL 1	Remember

5	What is meant by handling large volumes of data?	BTL 1	Remember
6	Mention any two techniques used to handle large data sets.	BTL 1	Remember
7	Define parallel processing in data analysis.	BTL 1	Remember
8	What is streaming data?	BTL 1	Remember
9	What is a malicious URL?	BTL 1	Remember
10	List any four features used in malicious URL prediction.	BTL 1	Remember
11	What is a recommender system?	BTL 1	Remember
12	Define collaborative filtering.	BTL 1	Remember
13	What is content-based filtering?	BTL 1	Remember
14	List any four programming techniques used for handling large datasets.	BTL 1	Remember
15	Why is model evaluation important?	BTL 1	Remember
16	Define presentation and automation in the Data Science life cycle.	BTL 2	Understand
17	Define chunking as a programming technique for large data.	BTL 1	Remember
18	List the steps in data preparation	BTL 2	Understand
19	State any four challenges faced while storing large datasets.	BTL 1	Remember
20	State the role of sampling in handling large datasets.	BTL 2	Understand
21	What is meant by data cleansing?	BTL 1	Remember
22	Compare user-based and item-based collaborative filtering.	BTL 2	Understand
23	What is data preparation?	BTL 1	Remember
24	List two automation tools used in Data Science	BTL 1	Remember

PART – B			
Q.No	Question	BT Level	Competence
1	Analyze various techniques used for handling large volumes of data and justify their suitability for different application scenarios.	BTL 4	Analyze
2	Analyze chunking and sampling techniques for processing large datasets by comparing their advantages, limitations, and application scenarios..	BTL 4	Analyze
3	Apply the Data Science workflow to the case study of predicting malicious URLs, covering problem definition, data collection, feature extraction, model building, and evaluation	BTL 3	Apply
4	Apply the process of building a recommender system by explaining each stage with a suitable real-world example.	BTL 3	Apply
5	Analyze the steps involved in formulating a research question for a data science project.	BTL 4	Analyze
6	Analyze the data preparation phase in handling large datasets and explain its impact on model performance with suitable examples.	BTL 4	Analyze
7	Apply the stages involved in building a machine learning model for a large dataset and explain the purpose of each stage with a suitable example..	BTL 3	Apply
8	Evaluate the role of presentation and automation in deploying a Data Science solution with suitable examples.	BTL 5	Evaluate
9	Apply programming strategies such as chunk processing, vectorization, parallel processing, and memory optimization for handling large datasets with suitable examples.	BTL 3	Apply
10	Analyze collaborative filtering and content-based filtering techniques used in recommender systems.	BTL 4	Analyze

11	Analyze programming techniques used for processing large datasets efficiently and justify their advantages.	BTL 4	Analyze
12	Analyze preprocessing and data preparation techniques used before model building for handling large datasets and explain their impact on prediction accuracy.	BTL 4	Analyze
13	Evaluate collaborative filtering, content-based filtering, and hybrid recommender techniques for e-commerce applications and justify the most appropriate recommendation approach.	BTL 5	Evaluate
14	Analyze the techniques and tools required at different stages of handling large datasets in a Data Science project.	BTL 4	Analyze
15	Evaluate the effectiveness of different techniques for handling large volumes of data in real-time applications	BTL 5	Evaluate
16	Evaluate different machine learning approaches for predicting malicious URLs and justify the most suitable approach based on accuracy, scalability, and performance.	BTL 5	Evaluate
17	Design a complete recommender system architecture for an e-commerce platform, including data collection, preprocessing, model building, recommendation generation, evaluation, and presentation of results.	BTL 6	Design
