

THE KAVERY ENGINEERING COLLEGE
(An Autonomous Institution)

B.E/B.TECH DEGREE END SEMESTER THEORY EXAMINATIONS, NOV /DEC 2024

Sixth Semester

Agricultural Engineering

OCS353- Data Science Fundamentals

(Common to Civil Engineering, Chemical Engineering and Mechanical Engineering)

Regulations – 2021

Duration : 3 Hrs

Maximum : 100 Marks

PART – A (ANSWER ALL QUESTIONS) (Marks 10*2=20)

<i>Q.No</i>	<i>Questions</i>	<i>BLT</i>	<i>CO</i>	<i>Marks</i>
1.	Describe the significance of defining research goals in the Data Science process.	RE	1	2
2.	What is the role of data retrieval in the Data Science process?	RE	1	2
3.	What is Jupyter Notebook? State its advantages for data analysis.	UN	2	2
4.	Define aggregation in the context of NumPy. How can you compute the sum of elements along a specific axis of an array?	UN	2	2
5.	Differentiate between classification and regression in supervised learning.	RE	3	2
6.	Define supervised learning and provide one example of a problem that can be solved using supervised learning.	UN	3	2
7.	What is 3D plotting in Matplotlib? How to you create a 3D plot using Matplotlib?	RE	4	2
8.	How do you add a legend to a plot in Matplotlib?	RE	4	2
9.	Why is it important to automate the presentation of results in a data science project?	AN	5	2
10.	What are the key techniques used in building a recommender system?	RE	5	2

PART – B (ANSWER ALL QUESTIONS) (Marks 5*13 = 65)

<i>Q.No</i>	<i>Questions</i>	<i>BLT</i>	<i>CO</i>	<i>Marks</i>
11.	a i Explain the various facets of data in Data Science. How do different types of data influence the process of data analysis?	UN	1	7
	ii Illustrate about data warehousing techniques support data storage, retrieval, and analysis, and how does it differ from traditional databases?	UN	1	6
	Or			
	b i Explain each step involved in Data Science, including defining research goals, retrieving data, data preparation, and building models.	UN	1	7
	ii Explain the common challenges faced during the data preparation phase and explain how techniques like data cleaning and feature selection help improve the quality of the data	UN	1	6
12.	a Discuss the role of IPython Magic Commands in Python programming. Provide detailed examples of commonly used magic commands in IPython.	UN	2	13

Or

	b i	Explain Universal Functions in NumPy. How do they work on NumPy arrays, and what are their advantages over using Python loops for array operations?	UN	2	8
	ii	Provide examples of commonly used ufuncs like np.add, np.multiply, and np.sqrt.	UN	2	5
13.	a	Explain clustering in unsupervised learning. Discuss the differences between partitioning-based clustering algorithms and hierarchical clustering algorithms.	UN	3	13
		Or			
	b	Describe the concept of outlier analysis in machine learning. Discuss different strategies for handling outliers.	UN	3	13
14.	a	Discuss the role of text and annotations in enhancing the clarity of data visualizations with an example.	CR	4	13
		Or			
	b	What are subplots in Matplotlib, and why are they useful in visualizing multiple related plots? Discuss how you would use subplots to display different views of the same dataset.	CR	4	13
15.	a	Explain how data cleaning, transformation, and normalization can enhance the quality of data and lead to better model results.	UN	5	13
		Or			
	b	What tools and techniques are essential for handling large datasets in real-world machine learning applications? Discuss any two tools with an example.	UN	5	13

PART – C (Marks 1*15 = 15)

<i>Q.No</i>		<i>Questions</i>	<i>BLT</i>	<i>CO</i>	<i>Marks</i>
16.	a	Assume you are given a dataset with labeled customer information (age, income, spending score) and the target variable is whether a customer will buy a product (Yes/No). How would you apply supervised learning to build a classification model for this problem? Explain the steps involved, including model selection, training, and evaluation.	AP	3	15
		Or			
	b	Imagine you have a series of data that represents the amount of precipitation each day for a year in a given city. Load the daily rainfall statistics for the city of Chennai in 2024 which is given in a csv file ChennaiRainfall2024.csv using Pandas generate a histogram for rainy days, and find out the days that have high rainfall.	AP	4	15