Reg.No.:												

VIVEKANANDHA COLLEGE OF ENGINEERING FOR WOMEN
[AUTONOMOUS INSTITUTION AFFILIATED TO ANNA UNIVERSITY, CHENNAI]
Elayampalayam – 637 205, Tiruchengode, Namakkal Dt., Tamil Nadu.

Question Paper Code: 6020

M.E. / M.Tech. DEGREE END-SEMESTER EXAMINATIONS – JUNE / JULY 2024
Second Semester
Information Technology
P23ITE08 – DATA SCIENCE
(Regulation 2023)

Time: Three Hours | Maximum: 100 Marks

Answer ALL the questions

Knowledge Levels (KL)	K1 – Remembering	K3 – Applying	K5 - Evaluating
	K2 – Understanding	K4 – Analyzing	K6 - Creating

PART – A

(10 x 2 = 20 Marks)

Q.No.	Questions	Marks	KL	CO
1.	Enlist the Facets of Data Science.	2	K1	CO1
2.	How Cleansing, integrating, and transforming data is performed in data science process.	2	K2	CO1
3.	Describe data modeling process.	2	K1	CO2
4.	What are the problems faced when handling large data?	2	K1	CO2
5.	Briefly elaborate Hadoop Apache a framework requirements which needs to satisfy.	2	K2	CO3
6.	What is the importance of Data Stream?	2	K1	CO3
7.	What objectives are achieved by NoSQL over SQL database? Give suitable example.	2	K1	CO4
8.	Why and when one has to use a graph database?	2	K2	CO4
9.	Why text mining is important and enlist the application of text mining in real-world applications?	2	K2	CO5
10.	How JQuery is initiated to visualize the information out of database?	2	K2	CO5

PART – B

(5 x 13 = 65 Marks)

Q.No.		Questions	Marks	KL	CO
11.	a)	Under big data ecosystem and data science identify the six important components without which we can assume the good system. Analyze and discuss with suitable example.	13	K3	CO1
		(OR)			
	b)	Describe in detail about Exploratory Data Analysis.	13	K2	CO1
12.	a)	Use a technique known as Principal Component Analysis (PCA) to find latent variables in a dataset that describes the quality of wine. Then compare how well a set of latent variables works in predicting the quality of wine against the original observable set.	13	K3	CO2

Fixed acidity	Volatile acidity	Citric acid	Residual sugar	Chlorides	Free sulfur dioxide	Total sulfur dioxide	Density	pH	Sulfates	Alcohol	Quality
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5
7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5

(OR)

	b)	Analyze the problem faced while handling large data. Discuss the techniques to handle large volume of data with suitable example.	13	K3	CO2
13.	a)	Explain in detail about the different components of Hadoop Framework.	13	K2	CO3
		(OR)			
	b)	Describe the concept of filtering streams and explain in detail about how to count distinct elements in a stream.	13	K2	CO3
14.	a)	Analyze the BASE principles of NoSQL databases. Discuss with example NoSQL database types.	13	K3	CO4
		(OR)			
	b)	Design recommendation engine using graph databases. Apply data science process to connected data recommender model.	13	K5	CO4
15.	a)	What is the role of Bag-of-word and Stemming and lemmatization in text mining techniques? The following dataset gives information of ill patients with tumors. Use the ID3 algorithm to build a decision tree for predicting tumor malignancy.	13	K2	CO5

Consider age as a continuous variable and split on a range of age values as <7.5 and >= 7.5.

Age	Vaccination	Tumor size	Tumor Site	Malign ant
5	1	Small	Shoulder	0
9	1	Small	Knee	0
6	0	Small	Marrow	0
6	1	Medium	Chest	0
7	0	Medium	Shoulder	0
8	1	Large	Shoulder	0
5	1	Large	Liver	0
9	0	Small	Liver	1
8	0	Medium	Shoulder	1
8	0	Medium	Shoulder	1
6	0	Small	Marrow	1
7	0	Small	Chest	1

Compute root attribute of the tree? Provide a brief justification for your choice.

i. Compute root attribute of the tree? Provide a brief justification for your choice.

ii. If the root is at level 0, which attribute(s) will be present at level 1. Compute the decision tree till level 1 and draw it. Show the splits of training examples from level 1 in the tree drawn.

(OR)

b) In what ways the results can be communicated to end user by the data scientist. Analyze these way with suitable example. Also, determine the important factors of utilized in data visualization for end user. 13 K3 CO5

PART – C

(1 x 15 = 15 Marks)

Q.No.		Questions	Marks	KL	CO
16.	a)	Consider the cancer dataset to detect the cancer. You've developed a classification model which achieved an accuracy of 93%. Discuss the issues, If you are not satisfied with your model performance? Further, analyse and discuss what you can do about it? Which performance metric will be more suitable for your model and why?	15	K5	CO1
		(OR)			
	b)	Consider some dataset D. This dataset contains numerous variables, some of which are highly correlated and you know these variables. Propose the ways to handle such high dimensional data.	15	K3	CO2