Reg No.:_______________ Name:_______________

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

Fifth Semester B.Tech Degree (S, FE) Examination June 2024 (2019 Scheme)

Course Code: ADT 301

Course Name: FOUNDATIONS OF DATA SCIENCE

Max. Marks: 100 Duration: 3 Hours

PART A

(Answer all questions; each question carries 3 marks)

			Marks
1		What is data science? Why data science is required?	3
2		Compare structured and unstructured data with an example?	3
3		List various data reduction strategies?	3
4		What are the problems to be considered during data integration task?	3
5		Is regression a supervised learning technique? Justify your answer.	3
6		Illustrate the strength and weakness of KNN classifiers?	3
7		Infer the conditions to be satisfied for an association rule to be strong? Illustrate with an example.	3
8		How can you infer Euclidean distance between two points in a cluster?	3
9		Compare and contrast precision. recall and F-measure?	3
10		Explain ensemble learning? list out different types	3

PART B

(Answer one full question from each module, each question carries 14 marks)

Module -1

11	a)	Demonstrate the different stages in the data science process?	7
	b)	Identify the different domains where data science plays an active role?	7
12	a)	List and briefly explain various tools and skills required for data science?	7
	b)	Summarise ethical aspects of data science?	7

Module -2

13	a)	Briefly explain the pre-processing techniques available in data mining?	8
	b)	What is data visualization and explain different techniques used for visualizing data?	6
14	a)	Explain the terms data reduction and data transformation with an example?	7

b) Given the following data for the attribute age: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. Use binning methods to smooth these data with a bin depth of 3. Illustrate your steps 7

Module -3

15 a) What is meant by decision tree induction? Explain the working of the decision tree algorithm with information gain? 7

b) Why naive Bayesian classification is called "naive"? Briefly outline the major ideas of naive Bayesian classification 7

16 a) Briefly describe the classification processes using 7

(i) Bayesian belief network (ii) Rule based classification

b) Describe the concept of Support Vector Machine Classification for linear and non-linear data? 7

Module -4

17 a) Consider the following transactions and find frequent itemsets and generate association rules for them. Let minimum support count be 2 and minimum confidence is 60%. 10

TID	ITEMSETS
T1	A, B
T2	B, D
T3	B, C
T4	A, B, D
T5	A, C
T6	B, C
T7	A, C
T8	A, B, C, E
T9	A, B, C

b) Explain the working of k-means algorithm with the help of an example? 4

18 a) What is the Apriori algorithm used for? Give the steps used in the Apriori algorithm to find the most frequent itemsets 7

b) Differentiate between Agglomerative and Divisive Hierarchical Clustering? 7

Module -5

19 a) Explain the different methods for improving the model performance? 7

b) Suppose 10000 patients get tested for flu; out of them, 9000 are actually healthy and 1000 are actually sick. For the sick people, a test was positive for 620 and 7 negative for 380. For the healthy people, the same test was positive for 180 and negative for 8820. Construct a confusion matrix for the data and compute the precision and recall for the data?

20 a) Why Bootstrap sampling is considered as the building block for many modern machine learning algorithms? 7

b) Explain the general process of k-fold cross validation? 7

***