Reg No.:_____ Name:_____

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

B.Tech Degree S5 (S,FE) Examinations June 2026 (2019 Scheme)

Course Code: CDT305
Course Name: DATA ANALYTICS

Max. Marks: 100 Duration: 3 Hours

PART A

(Answer all questions; each question carries 3 marks)

		Marks
1	Explain following attribute types with examples: (i) Ordinal attribute (ii) Nominal attribute (iii) Numeric attribute	3
2	List out various types of relationships that can exist between ordinal and continuous variables?	3
3	What is discretization in data pre-processing, and when is it used?	3
4	Describe how the IQR helps in identifying outliers in a dataset.	3
5	Compare rare pattern and negative pattern with the help of an example.	3
6	How can we use support and confidence measures for the generation of association rules?	3
7	Explain the main assumption behind the Naive Bayes classifier. Why is it called "naive"?	3
8	List and describe three common linkage criteria used in Agglomerative Clustering.	3
9	What is tokenization in text processing?	3
10	Illustrate IR problem with an example.	3

PART B

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

Module -1

11	a)	Describe Online Analytical Processing Server. What are five basic analytical operations that can be performed on an OLAP cube?	6

b) A researcher collected data on the number of hours spent studying per week and the scores obtained on a final exam for a group of students. The data is summarized below: 8

(i)Calculate Spearman's rank correlation coefficient (ρ) for the data.
(ii) Interpret the result

Student	Hours studied	Exam score
1	2	65
2	3	70
3	5	75
4	1	60
5	4	80
6	6	85

12 a) Outline the diagram of the Data Analytics Life Cycle and provide a comprehensive description of each phase, explaining its purpose and key activities. 8

b) The famous passenger liner Titanic hit an iceberg in 1912 and sank. A total of 337 passengers travelled in first class, 285 in second class, and 721 in third class. In addition, there were 885 staff members on board. Not all passengers could be rescued. Only the following were rescued: 135 from the first class, 160 from the second class, 541 from the third class and 674 staff. 6

(a) Determine and interpret the contingency table for the variables "travel class" and "rescue status".

Module -2

13 a) Write about different data cleaning methods? 7

b) The following data represents the test scores of 20 students in a class: 45, 78, 92, 85, 67, 55, 91, 87, 72, 83, 88, 77, 66, 79, 90, 95, 80, 74, 92, 68. Calculate the five-number summary and the interquartile range (IQR). Discuss the importance of these statistics in understanding the data. 7

14 a) Describe the role of data reduction in data pre-processing. What are the main techniques used in data reduction. 5

b) A hiker recorded the distances hiked (in km) and the maximum altitudes (in m) for his last 5 hikes: 9

Hike	Distance(km)	Maximum Altitude(m)
1	2	150
2	4	200
3	3	180
4	5	250
5	1	120

(a) Calculate the interquartile range (IQR) for both variables.
(b) Calculate the absolute median deviation (AMD) for both variables.
(c) Calculate the standard deviation for both variables

Module -3

15 a) Suppose we have the following dataset that has various transactions, and from this dataset, you need to find the frequent item sets and generate the association rules using the Apriori algorithm. Given minimum support=2, minimum confidence=50% 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 different steps of Apriori algorithm with help of a flowchart. 4

16 a) Define frequent patterns in association rule mining. How do techniques like the Apriori algorithm and FP-growth identify frequent patterns? 10

b) Write a note on Query directed mining 4

Module -4

17 a) Explain the K-Means clustering algorithm, describing its key steps and the role of centroids in the clustering process. How does the algorithm determine which data points belong to each cluster, and what are some practical applications where K-Means clustering is effectively used? 10

b) What are attribute selection measures in data mining? Discuss the different criteria or methods used for selecting relevant attributes and explain their importance in the process of feature selection. 4

18 a) Given the following data on a certain set of patients seen by a doctor. Can the doctor conclude that a person having chills, fever, mild headache and without running nose has flu? (Use Naive Bayes classification) 10

Chills	Running nose	Headache	Fever	Has flu
Y	N	Mild	Y	N
Y	Y	No	N	Y
Y	N	Strong	Y	Y
N	Y	Mild	Y	Y
N	N	No	N	N
N	Y	Strong	Y	Y
N	Y	Strong	N	N
Y	Y	Mild	Y	Y

b) Write a note on BIRCH. 4

Module -5

19 a) What are the key components of an inverted index? How do they work together to facilitate efficient search operations? 10

b) Give any four applications of Naïve Bayes Text Classification 4

20 a) Explain the concept of term frequency (TF) and inverse document frequency (IDF) in the context of the Vector Space Model. How are these measures used to compute the relevance of documents to a query? 9

b) What is case-folding? Explain with an example. 5

***