

Reg No.: _____

Name: _____

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY
B.Tech Degree S5 (S,FE) Examinations June 2026 (2019 Scheme)



Course Code: ADT301
Course Name: FOUNDATIONS OF DATA SCIENCE

Max. Marks: 100

Duration: 3 Hours

PART A

(Answer all questions; each question carries 3 marks)

Marks

- | | | |
|----|---|---|
| 1 | Illustrate a real-world application of data science, such as in disease tracking, and its impact. | 3 |
| 2 | Explain the role of data science in modern business decision-making and provide an example. | 3 |
| 3 | Describe the process of handling missing values during data cleaning. | 3 |
| 4 | Discuss data integration and its role in data preprocessing. | 3 |
| 5 | Define lazy learning and explain how K-Nearest Neighbour (K-NN) works as a lazy learner. | 3 |
| 6 | Describe a multilayer feed-forward neural network and its role in classification tasks | 3 |
| 7 | Differentiate between support and confidence in the context of association rule mining. | 3 |
| 8 | Describe how distance measures are used in clustering algorithms and provide an example. | 3 |
| 9 | Explain the concepts of sensitivity and specificity and their importance in model evaluation. | 3 |
| 10 | What is bagging, and how does it improve model performance? | 3 |

PART B

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

Module -1

- | | | |
|----|--|---|
| 11 | a) List and briefly describe key steps in the data science process | 7 |
| | b) Explain a major ethical or privacy concern in data science and why it matters. | 7 |
| 12 | a) Describe a major research challenge in data science and its implications for future developments. | 7 |

- b) Explain any three tools or platforms commonly used in data science and their main functions. 7

Module -2

- 13 a) Given a dataset containing the annual income (in thousands) of households in a specific region: 32, 45, 33, 47, 52, 51, 70, 73, 120, 130, 128, 125, 118, 130, 140. Use binning methods to smooth these data. Illustrate the steps. 6
- b) Discuss data quality and its impact on data mining results. Explain the types of data quality issues that can arise and describe how data preprocessing techniques address these issues. 8
- 14 a) What is data integration, and why is it crucial in data preprocessing? Explain how data integration is performed and discuss the challenges that can arise when integrating data from multiple sources. 6
- b) A researcher is analysing the exam scores of students in a class. The scores (out of 100) are as follows: 55, 88, 47, 90, 75, 61, 82, 95, 68, 73. Apply the following methods to normalize the data. Normalize the data to a range of [0, 1]. 8
- (i) min-max normalization. (ii) z-score normalization.

Module -3

- 15 a) Describe the process of decision tree induction in detail. Explain how decision trees are constructed, including the criteria used for splitting nodes. 8
- b) Outline the Naive Bayesian classification using the below dataset classify given data: X : (Color= Green, Legs=2, Height=Tall, Smelly=No, Species= ?). 6

No	Color	Legs	Height	Smelly	Species
1	White	3	Short	Yes	M
2	Green	2	Tall	No	M
3	Green	3	Short	Yes	M
4	White	3	Short	Yes	M
5	Green	2	Short	No	H
6	White	2	Tall	No	H
7	White	2	Tall	No	H
8	White	2	Short	Yes	H

- 16 a) What is a Support Vector Machine (SVM) and how does it work? How SVM finds the optimal hyperplane for classification. Discuss one advantage and one limitation of using SVMs. 8

- b) What is case-based reasoning in classification? Explain how it works and describe a scenario in which it is useful. 6

Module -4

- 17 a) What is DBSCAN ? Explain how it identifies clusters based on density, and discuss its advantages. 7
- b) Describe the process of generating association rules from frequent itemsets. Define key concepts and explain how these metrics help evaluate the strength of association rules. 7
- 18 a) Discuss the role of distance measures in clustering algorithms. Explain at least two common distance measures. 4
- b) Consider the following dataset, find all frequent item sets and generate all strong association rules for them. Let minimum support count be 3 and minimum confidence be 60%. 10

1	Milk	Egg	Bread	Butter
2	Milk	Butter	Egg	Ketch up
3	Bread	Butter	Ketch up	
4	Milk	Bread	Butter	
5	Bread	Butter	Cooki es	
6	Milk	Bread	Butter	Cooki es
7	Milk	Cooki es		
8	Milk	Bread	Butter	
9	Bread	Butter	Egg	Cooki es
10	Milk	Butter	Bread	
11	Milk	Bread	Butter	
12	Milk	Bread	Cooki es	Ketch up

Module -5

- 19 a) Explain the concept of a confusion matrix and its components. How is it used to evaluate the performance of a classification model? 5
- b) Assume the following: A health organization developed a machine learning model to screen for a specific disease in patients. The model was tested on 200 patients. Out of all patients who truly have the disease, 50 were correctly identified by the model, but 30 were missed and classified as not having the disease. Among the patients who do not have the disease, 20 were incorrectly labeled as positive by the model, while 100 were correctly identified as negative. Construct the confusion matrix for the search and calculate the precision, recall, specificity and accuracy scores for the search. 9
- 20 a) Describe K-fold cross-validation, outlining the steps involved. What are the advantages and drawbacks of using K-fold cross-validation compared to hold-out validation? 7
- b) What is an ROC curve, and how is the AUC used to assess model performance? Discuss the advantages of ROC-AUC over accuracy for model comparison. 7
