



P.E.S. College of Engineering, Mandya - 571 401

(An Autonomous Institution affiliated to VTU, Belagavi)

Sixth Semester, B.E. - Computer Science and Engineering

Semester End Examination; July / Aug. - 2022

Data Analytics

Time: 3 hrs

Max. Marks: 100

Course Outcomes

The Students will be able to:

CO1: Analyze data sets using Descriptive Statistics.

CO2: Apply data pre-processing methods on raw data set.

CO3: Apply unsupervised algorithms for the give problem.

CO4: Apply supervised algorithms for the give problem.

CO5: Design and Implement real time applications in data analytics

Note: I) PART - A is compulsory. Two marks for each question.

II) PART - B: Answer any **Two** sub questions (from a, b, c) for a Maximum of **18 marks** from each unit.

Q. No.	Questions	Marks	BLs	COs	POs																					
I : PART - A		10																								
I a.	List the short taxonomy of Data Analytics.	2	L2	CO1	1,2																					
b.	What is a PCA?	2	L2	CO2	1,2,3																					
c.	What is Silhouette internal Index?	2	L2	CO3	2,3																					
d.	Write the predictive performance measure for a binary classification task.	2	L2	CO4	2,3																					
e.	What is Ensemble learning?	2	L2	CO5	2,3																					
II : PART - B		90																								
UNIT - I		18																								
1 a.	Explain CRISP–DM methodology in detail.	9	L3	CO1	1,2																					
b.	Calculate the correlation between the “Age and “Glucose level” attributes using the given Table–1b, and discuss the type of correlation between them.	9	L3	CO1	1, 2																					
<table border="1" style="margin: auto; border-collapse: collapse;"> <thead> <tr> <th style="width: 20%;">Subject</th> <th style="width: 20%;">Age</th> <th style="width: 20%;">Glucose level</th> </tr> </thead> <tbody> <tr><td style="text-align: center;">1</td><td style="text-align: center;">43</td><td style="text-align: center;">99</td></tr> <tr><td style="text-align: center;">2</td><td style="text-align: center;">21</td><td style="text-align: center;">65</td></tr> <tr><td style="text-align: center;">3</td><td style="text-align: center;">25</td><td style="text-align: center;">79</td></tr> <tr><td style="text-align: center;">4</td><td style="text-align: center;">42</td><td style="text-align: center;">75</td></tr> <tr><td style="text-align: center;">5</td><td style="text-align: center;">57</td><td style="text-align: center;">87</td></tr> <tr><td style="text-align: center;">6</td><td style="text-align: center;">59</td><td style="text-align: center;">81</td></tr> </tbody> </table>						Subject	Age	Glucose level	1	43	99	2	21	65	3	25	79	4	42	75	5	57	87	6	59	81
Subject	Age	Glucose level																								
1	43	99																								
2	21	65																								
3	25	79																								
4	42	75																								
5	57	87																								
6	59	81																								
Table–1b																										
c.	Compute the location and the dispersion statistics for the attribute “age” where its values are 24, 72, 18, 59, 47, 11, 61 and 33.	9	L3	CO1	1,2																					

UNIT - II

18

- 2 a. Explain the main problems that affect the data quality. 9 L2 CO2 1, 2,3
- b. Define discretization. Explain the steps involved in discretization. Apply the same to convert 9 quantitative values 2, 3, 5, 7, 10, 15, 16, 19, and 20 into three bins, whose nominal values are A, B and C, using association by width and association by frequency. 9 L2 CO2 1,2,3
L3
- c. Define Dimensionality reduction. List the different ways and explain any one in detail. 9 L2 CO2 1, 2,3

UNIT - III

18

- 3 a. Define clustering. Explain the main types of clusters. 9 L2 CO3 2,3
- b. Apply K-means algorithm for the dataset $K = \{2, 3, 4, 10, 11, 12, 20, 25, 30\}$ to form two cluster. 9 L2 CO3 1
- c. Find the frequent itemset and generate association rules for the following given transaction dataset. Assume that minimum support threshold = 2 and with the association confident threshold 50%

Transaction ID	Items
T ₁	b, d, c, a
T ₂	e, d, c
T ₃	a, b
T ₄	a, c, d
T ₅	f, g, d, b

9 L3 CO3 2,3

UNIT - IV

18

- 4 a. Explain how the predictive performance for regression is measured. 9 L2 CO4 2,3
- b. Write a pseudocode for an K-NN algorithm and, list the advantages and disadvantages of K-NN. 9 L2 CO4 2,3
- c. Explain how naive bayes algorithm used for classification. 9 L2 CO4 2,3

UNIT - V

18

- 5 a. Write a Hunt decision tree induction algorithm and list its advantages and disadvantages. 9 L2 CO5 2,3
- b. Explain the main hyper parameters of ANN. 9 L2 CO5 2,3
- c. Explain the different phases of text mining task. 9 L2 CO5 2,3

* * * *