USN	1	S	I							

MOBILE PHONES PROHIBITED **RCSE31 / CSEP33**

Siddaganga Institute of Technology, Tumakuru – 572 103

(An Autonomous Institution affiliated to VTU, Belagavi, Approved by AICTE, New Delhi)

Fifth Semester B.E. Computer Science & Engg. Examinations Feb. - Mar. 2022

Foundations of Data Science

Time: 3 Hours **Max. Marks: 100**

Note : Answer any five questions choosing one full question from each unit.

Unit - I

1 a) Mention and explain any four data mining tabs. 8

BL:	1	CO:	1	PO:	1	PSO:	3

b) Give the CRISP model for data mining. Explain its stages. 8

BL:	1	CO:	1	PO:	1	PSO:	3

c) Distinguish between Bigdata 1.0 and Bigdata 2.0. 4

BL:	1	CO:	1	PO:	1	PSO:	3

OR

2 a) Describe the data discovery and data preparation performed during the data analytics life cycle for predicting customer churn. 10

BL:	2	CO:	1	PO:	1	PSO:	3

b) With a neat diagram explain how data science is placed in the content of various data related processes. 10

BL:	2	CO:	1	PO:	1	PSO:	3

Unit – II

3 a) Define entropy and information gain. Construct a decision tree for the following data set.

S.No.	Age	Balance	Class
1	20	40K	Yes
2	25	50K	Yes
3	30	30K	Yes
4	35	110K	No
5	40	120K	No
6	40	80K	Yes
7	30	115K	No
8	35	80K	?

Classify the 8th instance based on the decision tree constructed. 10

BL:	1	CO:	2	PO:	3	PSO:	3

b) Differentiate between the following:

i) Predictive and descriptive modeling

ii) Supervised and unsupervised segmentation

iii) Decision tree and probability estimation tree

iv) Induction and deduction. 10

BL:	2	CO:	2	PO:	3	PSO:	3

OR

4 a) Discuss the different measures for selecting the attribute with the best split. Give suitable example. 6

BL:	1	CO:	2	PO:	3	PSO:	3

b) What is probability estimation? Outline steps to construct probability estimation tree with an example. 6

BL:	2	CO:	2	PO:	3	PSO:	3

c) Consider the training examples shown below in Table(1) for a binary classification problem.

Instance	a1	a2	a3	Target class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	F	7.0	+
9	F	T	5.0	-

i) What is the entropy of this collection of training examples with respect to the positive class.

ii) What are the information gains of a1 and a2 relative to these training examples? 8

BL:	3	**CO:**	2	**PO:**	3	**PSO:**	3

Unit – III

5 a) For the following data propose a linear discriminant function and predict the class for 7th tuple.

S.No.	Age	Balance	Class
1	25	40K	Yes
2	20	50K	Yes
3	35	30K	Yes
4	35	110K	No
5	40	120K	No
6	50	80K	No
7	60	115K	?

8

BL:	3	**CO:**	3	**PO:**	2	**PSO:**	3

b) List the objectives of support vector machine? How they are handled? Explain with suitable diagrams. 6

BL:	2	**CO:**	3	**PO:**	2	**PSO:**	3

c) Discuss over fitting in mathematical function. Give suitable example. 6

BL:	2	**CO:**	3	**PO:**	2	**PSO:**	3

OR

6 a) Explain with an example how you can get class probability estimation using logistic regression. 7

BL:	2	**CO:**	3	**PO:**	2	**PSO:**	3

b) Differentiate classification trees with linear classifiers. 6

BL:	2	**CO:**	3	**PO:**	2	**PSO:**	3

c) Illustrate linear discriminant function using Iris data set. 7

BL:	2	**CO:**	3	**PO:**	2	**PSO:**	3

Unit – IV

7 a) What is clustering? Explain the K-means algorithm for clustering. 10

BL:	2	**CO:**	4	**PO:**	3	**PSO:**	3

b) Construct a single link Dendrogram for the Euclidean distance matrix given below. Show all the intermediate distance matrices.

	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0

10

BL:	3	**CO:**	4	**PO:**	3	**PSO:**	3

OR

8 a) Discuss the issues with Nearest-Neighbour methods and explain how to overcome them. 6

BL:	2	**CO:**	4	**PO:**	3	**PSO:**	3

b) Consider the following data set classify David based on similarity moderated voting method.

Customer	Age	Income	Cards	Response
David	37	50K	2	?
John	35	35K	3	Yes
Rachael	22	50K	2	No
Ruth	63	200K	1	No
Jefferson	59	170K	1	No
Norah	25	40K	4	Yes

8

BL:	3	**CO:**	4	**PO:**	3	**PSO:**	3

c) Explain how Hierarchical clustering is carried out. Mention its use. 6

BL:	1	**CO:**	4	**PO:**	3	**PSO:**	3

Unit – V

9 a) Define the following:
i) Support ii) Confidence iii) Lift iv) Leverage v) Soft clustering. 10

BL:	1	**CO:**	5	**PO:**	2	**PSO:**	3

b) Explain the major tent mining approaches based on the kinds of data taken as input? Explain any two. 10

BL:	2	**CO:**	5	**PO:**	2	**PSO:**	3

OR

10 a) Explain how do you measure term frequency and inverse document frequency with suitable example. 10

BL:	2	**CO:**	5	**PO:**	2	**PSO:**	3

b) What is meant by aeration analysis? If the 10000 transactions analyzed, the data shows that 6000 of the customer transactions include computer games, while 7500 include videos and 4000 include both computer games and videos.
For the rule Buys(X, 'Computer Games') => Buys(X, Video) Compute support, confidence, lift, Leverage. 10

BL:	3	**CO:**	5	**PO:**	2	**PSO:**	3