M 059/033

Enrolment Number										

Total No. of printed pages = 04

Monsoon, 2023

MCA Semester Examinations

Introduction to Data Science

Course Code: MCA23503T

Full Marks – 60 **Time –2 $\frac{1}{2}$ hours**

The figure in the margin indicates full marks for the questions.

Part A

1. **Choose the most appropriate option from the following MCQs** 10×1=10

(a) Which of the following is not a supervised learning?
i) PCA ii) Naive Bayesian
iii) Linear regression iv) Decision tree

(b) Naïve Bayes classifier following assumption is held
i) Conditional independence ii) Conditional dependence
iii) both i) and ii) together iv) None of the mentioned

(c) How do you calculate Confidence of the rule (A ⇒ B) in APriory algorithm?
i) Support(A∪B) / Support (A) ii) Support(A∪B) / Support (B)
iii) Support(A∪B) / Total records iv) None of the mentioned

(d) Examples of Ordinal can be
i) ID Numbers, eye color, zip codes
ii) Rankings, taste of potato chips, grades, height
iii) Calendar dates, temperatures in Celsius or Fahrenheit, phone numbers
iv) Temperature in Kelvin, length, time, counts

(e) Under fitting happens due to
i) Fewer number of features ii) Data has high variance
iii) No use of regularization iv) All of the above

(f) Multiple data sources may be combined is called as
i) Data Reduction ii) Data Cleaning
iii) Data Integration iv) Data Transformation

(g) Which of the following studies the collection, analysis, interpretation or explanation, and presentation of data?
i) Statistics ii) Visualization
iii) Data Mining iv) Clustering

(h) Clustering is a
i) Supervised learning ii) Unsupervised learning
iii) Reinforcement learning iv) all of the mentioned

(i) What is the main goal of binning in data pre-processing?
i) Increasing data dimensionality
ii) Handling missing values
iii) Reducing noise in data
iv) Simplifying and transforming continuous data

(j) Which of the following is NOT a commonly used measure of central tendency for summarizing data?
i) Mean ii) Median
iii) Mode iv) Standard Deviation

Part B
(Answer any four questions) 4×5=20

2. What is the difference between data, information and intelligence? Mention the numeric attribute types with examples.

3. Mention name of three packages for data science in python.

#	Name	Subject	Score
1	Ravi	Physics	86
2	Bimal	Chemistry	76
3	Rohan	Physics	65
4	Ravi	Physics	89
5	Ravi	Maths	82
6	Bimal	Physics	66
7	Bimal	Maths	68
8	Rohan	Maths	69
9	Rohan	Chemistry	77

i) Write a statement in python to display the name of the students who got 80 or more marks in Physics.
ii) Write a python code to display the total scores obtained by each student along with the name.

4. Briefly explain why data pre-processing is necessary? Mention four data pre-processing techniques with examples.

5. Find the mathematical expectation of the random variable X: 2, 3, 4, 2, 2, 4, 5, 6
What is the interquartile range (IQR)? What are noise and outlier in data? How does it relate to the concept of outliers?

6. What are the criteria's commonly used for making splits in a decision tree? Explain the role of confusion matrix to the evaluation of a decision tree classifier.

7. Consider the following dataset

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	-
3	0	1	1	-
4	0	1	1	-

5	0	0	1	+
6	1	0	1	+
7	1	0	1	-
8	1	0	1	-
9	1	1	1	+
10	1	0	1	+

Estimate the conditional probabilities for P(A|+), P(B|+), P(C|+), P(A|-), P(B|-) and P(C|-)

Part C
(Answer any two questions) 2×10=20

8. (a) Mention the steps of hypotheses testing with a flowchart. 4

(b) Consider the following transactional dataset 6

Tid	**Items**
10	A, C, D
20	B, C, D
30	A, B, C, D
40	B, E
50	A, B, D

Where, Tid means transaction id. Find all the frequent patterns using A Priory algorithm. Consider minimum support count σ as 2.

9. (a) Consider the following dataset

X (year's of experience)	*Y* (salary in 1000$)
3	30
8	52
9	54
12	63
3	34
11	59

Find the equation of the line of regression from the data and predict the salary for *X*=7 6

(b) What are some common techniques for handling missing data in a dataset? 4

10. (a) What is the difference between symmetric and asymmetric binary variable?
Consider 2 objects Apple and Banana with four attributes as follows,

Feature of fruit	Sphere	Sweet	Sour	Crunchy
Object i=Apple	Yes	Yes	Yes	Yes
Object j=Banana	No	Yes	No	No

Find the similarity between Apple and Banana. 5

(b) Write the algorithm for *k*-nearest neighbour classification. Given *k* the nearest number of neighbours, and *n*, the number of attributes describing each tuple. 5

Part D
(Short Notes)

2×5=10

11. **(Write any two)**

(a) Page Ranking Algorithm
(b) Simple and multiple linear regression
(c) Bootstrap aggregation and boosting
(d) Feed forward neural network
(e) Dimensionality Reduction

----------X----------