GURU NANAK INSTITUTE OF TECHNOLOGY

An Autonomous Institute under MAKAUT

2022

BASIC DATA SCIENCE

MCA20-E304F

TIME ALLOTTED: 3Hours **FULL MARKS:70**

The figures in the margin indicate full marks.

Candidates are required to give their answers in their own words as far as practicable

GROUP – A

(Multiple Choice Type Questions)

Answer any *ten* from the following, choosing the correct alternative for each question: **10×1=10**

		Marks	CO No
1. i)	Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases? a) Central Limit Theorem b) Central Mean Theorem c) Centroid Limit Theorem d) All of the mentioned	1	CO2
ii)	_____ graph displays information as a series of data points connected by straight-line segments. a) Bar b) Scatter c) Histogram d) Line	1	CO1
iii)	The expected value or _____ of a random variable is the center of its distribution. a) mode b) median c) mean d) bayesian inference	1	CO2
iv)	What are the drawbacks of the linear model? a) The assumption of linearity of the errors b) It can't be used for count outcomes or binary outcomes c) There are overfitting problems that it can't solve d) All of the above	1	CO4
v)	In statistical data , P-Value > 0.05 means a) weak evidence against the null hypothesis b) strong evidence against the null hypothesis c) marginal value d) None of the above	1	CO3

vi) Which of the following is not a supervised learning — 1 — CO5
a) PCA
b) Linear regression
c) Decision tree
d) All of the above

vii) The square root of the variance is called the _____ deviation. — 1 — CO2
a) empirical
b) mean
c) continuous
d) standard

viii) Which of the following is the probability calculus of beliefs, given that beliefs follow certain rules? — 1 — CO3
a) Bayesian probability
b) Frequency probability
c) Frequency inference
d) Bayesian inference

ix) Data has been collected on visitors' viewing habits on a bank's website. Which technique is used to identify pages commonly viewed during the same visit to the website? — 1 — CO4
a) Clustering
b) Classification
c) Association Rules
d) Regression

x) The Least Square Method uses ___. — 1 — CO4
a) Linear polynomial
b) Linear regression
c) Linear sequence
d) None of the mentioned above

xi) Clustering belongs to ___ data analysis. — 1 — CO1
a) Supervised
b) Unsupervised
c) Both A and B
d) None of the mentioned above

GROUP – B
(Short Answer Type Questions)

(Answer any *three* of the following) **3 x 5 = 15**

		Marks	CO No
2.	Explain overfitting and underfitting in detail with an example.	5	CO3
3.	What are descriptive and inferential statistics with examples?	5	CO2
4. a.	What do you mean by data science?	3	CO1
b.	State some disadvantages of data science.	2	CO1

		Marks	CO No
5. a.	What are population and sample explain with an example.	3	CO2
b.	What is the goal of A/B Testing?	2	CO2
6.	Explain general techniques for handling large volumes of data.	5	CO4

GROUP – C
(Long Answer Type Questions)

(Answer any *three* of the following) **3 x 15 = 45**

		Marks	CO No
7. a.	Why are Feature extraction and engineering so important in machine learning?	7	CO4
b.	Is median or mode better for outliers? Which measure of central tendency is not affected by outliers?	8	CO4
8. a.	Discuss Conditional probability with an example in detail.	8	CO2
b.	Write a function to create a matrix given its shape and a function for generating its elements. Then use the function to generate to 5 x 5 identity matrix.	7	CO2
9. a.	Discuss Bias-Variance Tradeoff in detail.	8	CO5
b.	Discuss the need for fitting the model in multiple regression.	7	CO5
10. a.	Explain the k-means clustering algorithm. What are its drawbacks?	7	CO5
b.	What is a decision tree? Explain how the decision tree is constructed using the ID3 algorithm.	8	CO5
11.	Write a short note: (Any three)	3×5=15	
a.	Exploratory Data Analysis (EDA)	5	CO1
b.	Z-score standardization	5	CO2
c.	Data visualization	5	CO2
d.	Linear regression	5	CO4
e.	Random forest	5	CO4