

2056

B.E. (Information Technology)

Fourth Semester

ITMS-301: Data Analytics with Python

Time allowed: 3 Hours

Max. Marks: 50

NOTE: Attempt five questions in all, including Question No. 1 which is compulsory and selecting two questions from each Unit. Any missing or misprinted data maybe assumed suitably.

x-x-x**1. Attempt the following:-**

- (a) Explain the purpose of pandas in data analysis. How does it differ from NumPy and what are some key DataFrame operations?
- (b) Differentiate between Type I (α) and Type II (β) errors in hypothesis testing.
- (c) A binary classification model produced the following confusion matrix:

	Predicted Positive	Predicted Negative
Actual Positive	50	10
Actual Negative	5	35

Using the above confusion matrix, calculate: Accuracy and Precision

- (d) State the primary difference between K-Means and Hierarchical clustering methodologies.
- (e) Briefly contrast Linear Regression with Logistic Regression in terms of their target variables. (5x2)

UNIT - I

- 2. (a) Explain the concept of Probability Distributions. How is the normal distribution applied in analysing the distribution of sample means?
- (b) A sample of 100 students has a mean score of 75 and a known population variance of 25. Calculate the 95% confidence interval for the population mean. (Assume Z-value for 95% is 1.96). (2x5)
- 3. (a) Describe the step-by-step process of conducting a Two-Sample T-test. Provide a real-world engineering scenario where it would be applicable.
- (b) Explain the concept of One-Way ANOVA. Why is Post Hoc Analysis necessary after a significant ANOVA result? (2x5)

P.T.O.

(2)

4. (a) Write a Python script/snippet demonstrating how to import a dataset, calculate the mean, and find the variance using standard libraries.
- (b) What is a Randomized Block Design? Explain its purpose and how it is utilized within a Two-Way ANOVA framework. (2x5)

UNIT - II

5. (a) Detail the mathematical formulation of a Multiple Regression Model. How does Residual Analysis assist in validating the assumptions of this model?
- (b) A bank wants to predict whether a customer will default on a loan based on:
- income,
 - credit score,
 - previous loan history.

If you have to choose between Logistic Regression and Linear regression to solve this problem, which one you will choose and why? Also, Explain the role of Maximum Likelihood Estimation in estimating the parameters of Logistic Regression. (2x5)

6. (a) Explain how an ROC curve is constructed. How is the Area Under the Curve used to evaluate the performance of a classification model?
- (b) A survey was conducted among 100 students to examine whether there is an association between Gender and preference for Online Learning Mode.

The observed data are shown below:

Gender	Prefer Online Learning	Prefer Offline Learning	Total
Male	30	20	50
Female	15	35	50
Total	45	55	100

Using the Chi-Square Test of Independence, test whether Gender and Learning Preference are independent at 5% level of significance.

(Assume the critical value of χ^2 at 5% significance level and 1 degree of freedom is 3.84.) (2x5)

(3)

7. (a) Perform one iteration of the K-Means clustering algorithm for the following data points:

Point Coordinates

A	(2, 4)
B	(3, 3)
C	(8, 7)
D	(9, 8)

Assume:

- Number of clusters $K=2$
- Initial centroids are:
 - $C1=(2,4)$
 - $C2=(8,7)$

What are the common challenges faced when selecting the initial 'K' value? (5)

- (b) Discuss Classification and Regression Trees (CART). (1)

Consider the following training dataset used for predicting whether a student will pass an exam.

Student	Study Hours	Attendance	Assignment Submitted	Result
S1	High	Good	Yes	Pass
S2	High	Poor	Yes	Pass
S3	Medium	Good	Yes	Pass
S4	Low	Good	No	Fail
S5	Low	Poor	No	Fail
S6	Medium	Poor	Yes	Pass
S7	Medium	Good	No	Pass
S8	Low	Poor	Yes	Fail
S9	High	Good	No	Pass
S10	Low	Good	Yes	Fail

Calculate the entropy of the complete dataset. Calculate the Information Gain for each attribute. (4)

x-x-x