INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

End-Autumn Semester Examination 2023-24

Date of Examination: 24.11.2023 Session: FN Duration: 03 hours Full Marks: 100

Subject No. : **CS 61061** Subject: **Data Analytics** Computer Science & Engineering Department

Non-programmable calculators may be allowed.
Statistical tables may be allowed.

Special instructions:

- *Answer to all questions.*
- *All symbols in the question, if not mentioned explicitly bear their usual meanings.*
- *You may make reasonable assumptions, if any.*

---

1. A random sample of 102 MPs (Members of Parliament) was taken regarding the increased budget for the country's defense systems. Each MP was asked the following two questions:

   i. What is your political affiliation?
   ii. Are you in favor of increased arms spending?

   The results are given in Table 1.

		Political affiliation		
		Government	Opposition	None
Opinion	Favor	**16**	**21**	**11**
	No favor	**24**	**17**	**13**

Table 1: Q.1

With reference to the above data, answer the following questions.

(a) Suppose, Chi-square test is advised to analyze the data. Clearly state the hypotheses that can be tested.
(b) Calculate the χ^2-value and test the null hypothesis with 5% level of significance test with a p-value.
(c) What is the degree of correlation of the data?

[2+6+2]

2. WHO (World Health Organization) published the following data (see Table 2).

Country	**Death rate due to alcohol consumption**	**Death rate due to cirrhosis**
France	24.7	46.1
Italy	15.2	23.6
Germany	12.3	23.7
Australia	10.9	7.0
Belgium	10.8	12.3
USA	9.9	14.2

Table 2: Q.2

Given the data in Table 2, answer the following questions.

(a) Calculate the Pearson's correlation coefficient and conclude the coefficient of determination. What does the latter signify?
(b) Take the hypothesis that "Alcohol consumption is injurious to health". What is the Type-I error of your hypothesis testing would be?

[(3+1)+(4+2)]

3. It is generally believed that taller persons make better basketball players because they are better able to put the ball into the basket. Table 3 lists the heights (in inches) of a sample of 25 non-basketball athletes and the number of successful baskets made in a 60s time period.

(a) Perform a correlation analysis relating **Goals** to **Height** to ascertain whether there is a relationship between them.
(b) Estimate the number of goals to be made by an athlete who is 60 inches tall.
(c) How much confidence can be assigned to that estimate?

Table 3: Q.3

	Height	Goals		Height	Goals
1	71	15	**14**	72	16
2	74	19	**15**	71	15
3	70	11	**16**	75	20
4	71	15	**17**	71	15
5	69	12	**18**	75	19
6	73	17	**19**	78	22
7	72	15	**20**	79	23
8	75	19	**21**	70	13
9	72	16	**22**	72	16
10	74	18	**23**	75	20
11	71	13	**24**	76	21
12	72	15	**25**	74	19
13	73	17			

[2 + 2 + 2 + 2]

4. (a) Write down the logistic function for a binary dependent variable with m explanatory variables. Clearly state all the symbols in your expression.
(b) What is the odds? How is it related to the linear regression model?
(c) Obtain the logistic function for the data in Table 4. For each tuple (A, B, C) in Table 4, A denotes the toxic substance present in an individual body, B denotes the number of individuals having toxic A, and C denotes the number of individuals suffering from toxicology problem. **[4+4+2]**

Table 4: Q.4

A	B	C
0.0	50	2
2.1	54	5
5.4	46	5
8.0	51	10
15.0	50	40
19.5	52	42

5. It was decided to analyze the IQ levels of the secondary-level students from two different regions R_1 and R_2 in the country. An independent survey reveals the data as shown in Table 5. The entries in the table are the measurement of IQ levels in the range of 0 to 10 both inclusive. It was needed to analyze if the IQ levels of students in R_1 is better than the same of R_2.

Table 5: Q.5

R_1	R_2
4.5	3.0
6.2	4.5
5.8	3.8
6.0	4.0
7.1	3.7
6.8	3.2
3.8	4.1
7.2	5.5
	6.1
	4.0

Answer to the following questions.

(a) Identify the hypothesis to be tested. Justify your selection.
(b) Calculate the t-statistics based on the sample data in Table 5.
(c) Decide the critical t-value with 5% level of significance test.
(d) Obtain your conclusion.
(e) Measure the confidence interval of the pooled variance.

[1+3+2+1+3

6. A research lab wants to perform clinical trials to study the effectiveness of three drugs manufactured to cure a disease. The data in Table 6 represents the time taken in days to cure the disease for different patients when they consume either drug A, B or C.

Table 6: Q.6

A	B	C
101	90	108
90	105	107
103	84	92
96	83	105
110	92	83
125	100	100
121	88	80
114	77	97

With reference to this study, answer the following questions.

(a) Clearly identify the factors, levels and dependent variables involved in this study.
(b) Calculate the F-statistics.
(c) At the 0.05 level of significance, test whether the mean times for three drugs to cure the disease are equal.

[3+4+3]

7. Past records of a few patients are shown in Table 7. This table includes symptoms (Column 1 to 4) and diagnosis (Column 5). It is proposed to use the Naïve Bayesian classification to test whether a patient with the symptoms has the flu or not.

Table 7: Q.7

Fever	Cold	Running Nose	Headache	Flu
Y	Y	N	mild	N
N	Y	Y	no	Y
Y	Y	N	strong	Y
Y	N	Y	mild	Y
N	N	N	no	N
Y	N	Y	strong	Y
N	N	Y	strong	N
Y	Y	Y	mild	Y

With the above-mentioned problem statement, answer the following questions.

(a) Draw the contingency table which should include all the prior and posterior probabilities.
(b) Calculate the class probabilities **P(Flu = Y|T)** and **P(Flu = N|T)** where **T** denotes a test data as shown below.

Fever	Cold	Running Nose	Headache	Flu
N	Y	N	mild	?

[6+(2+2)]

8. With reference to Table 7, calculate the following estimates for the attribute "Headache".
   (a) Information Gain
   (b) Gain Ratio
   (c) Gini Index

[4+2+4]

9. Answer the following questions with respect to Support Vector Machine (SVM) for classification.
   (a) Write down the expression for hyperplane as a decision boundary to classify linearly separable data. Clearly state all the symbols in your stated expression.
   (b) How you should define the problem when you have to maximize the hyperplane separating vectors pertinent to any one of the two classes and linearly separable.
   (c) Write down the Lagrangian form of solving the problem using Lagrangian multiplier method. Define all symbols in your expression.
   (d) Mention all the constraints which should be solved to find the model parameters.

[2+2+2+4]

10. (a) Compare the strategies for validating a supervised classification model. You can follow the following format for recording your answer.

	Holdout method	**Random sampling**	**Bootstrap method**	**k-fold cross validation**
Concept				
Pros				
Cons				

(b) A confusion matrix is given for a binary class classifier (see Table 9).

		Predicted class	
		Class 1	Class 2
Actual class	Class 1	**412**	**38**
	Class 2	**64**	**396**

Table 9: Q. 10

Use the data in the table to answer the following performance estimations.

i. Observed accuracy
ii. True accuracy (assume $T_\alpha = 1.96$ for $\alpha = 95\%$)
iii. F_1-score

[4+(1+2+3)]

---*---