National University of Computer and Emerging Sciences, Chiniot-Faisalabad Campus

Course Name:	Data Mining	Course Code:	CS4038
Degree Program:	BS(CS)	Semester:	Spring-2024
Exam Duration:	60 Minutes	Total Marks:	60
Paper Date:	Tuesday, February 27, 2024	No of Page(s):	8
Sections:	ALL		
Exam Term & Type:	1st Sessional I Closed Book	Required Answer Book:	No
Course Instructor	Ms. Ayesha Liaqat		

Solution

Student Name:_____ Roll No._____ Section:_____ Invigilator's Signature

Q.Part#	Q.1	Q.2	Q.3	Q.4	Q.5	Q.6	Q.7	Q.8	Q.9	Q.10	Q.11	Q.12	Q.13	Q.14	Q.15	Total Obtained	Total Marks	Examiner Signature and Date
Marks																	60	

Instruction/Notes:	Attempt all questions. Programmable calculators are not allowed.
Vetted By	Vetter Signature:
Declaration by course instructor	The question paper has an 100% dissimilarity as compared to the question papers of the same subject from the last two years.

Question # 1. Data types, exploration and understanding (5+6+12+4+4+4=35)

Answer the following questions precisely in the given space ONLY.

a. Identify which one of the following categories the following data mining task belongs to

i. Classification
ii. Association analysis
iii. Clustering
iv. Regression
v. Anomaly detection

(i) Segmenting geographical regions based on climate data for urban planning.

Answer: Clustering

(ii) Predicting student performance based on factors such as study hours, attendance, and socioeconomic status.

Answer: Regression

(iii) Flagging unusual network traffic patterns to detect potential cyber-attacks or security breaches.

Answer: Anomaly Detection

(iv) Recognizing handwritten digits in postal codes for automated sorting in mail processing centers.

Answer: Classification

(v) Analyzing clickstream data to understand user navigation patterns and optimize website layout for improved user experience.

Answer: Association Analysis

b. Consider the data sample given below. Identify the type of each attribute as binary, discrete, or continuous. Also classify them as qualitative (nominal, ordinal or Binary) or quantitative (interval or ratio).

Empid	Name	Gender	Age	Designation	Salary
101	John	M	40	Director	130,000
102	James	M	28	Officer	60,000
104	Robert	M	35	Manager	75,000
107	Alex	F	28	Assistant	35,000
109	David	M	28	Officer	60,000
110	William	M	29	Officer	60,000
111	Michael	M	30	Assistant	35,000
120	Daniel	M	40	Assistant	40,000

Attribute name	Data type
Empid	Discrete, Qualitative (Nominal)
Name	Discrete, Qualitative (Nominal)
Gender	Discrete, Qualitative (Binary)
Designation	Discrete, Qualitative (Ordinal)
Age	Discrete, Quantitative (Ratio)
Salary	Discrete, Quantitative (Ratio)

c. Given the above data, standardize the "*Salary*" attribute using "Z-score" normalization method. No credit will be given for a direct answer, show all the necessary steps.

Z-score Formula:

$$v' = \frac{v - \mu}{\sigma}$$

$$\mu = \frac{\sum x}{n} = \frac{130000 + 60000 + 75000 + 35000 + 60000 + 60000 + 35000 + 40000}{8} = 61875$$

$$\sigma = \sqrt{\frac{1}{n-1}\sum (x - \bar{x})^2} = 31045.76116$$

Salary (v)	Normalized Salary (v')
130,000	$\frac{130,000 - 61875}{31045.76116} = 2.1943$
60,000	$\frac{60000 - 61875}{31045.76116} = -0.0603$
75,000	$\frac{75000 - 61875}{31045.76116} = 0.4227$
35,000	$\frac{35000 - 61875}{31045.76116} = -0.8656$
60,000	$\frac{60000 - 61875}{31045.76116} = -0.0603$
60,000	$\frac{60000 - 61875}{31045.76116} = -0.0603$
35,000	$\frac{35000 - 61875}{31045.76116} = -0.8656$
40,000	$\frac{40000 - 61875}{31045.76116} = -0.7046$

Is there any outlier present in the data? Mention datapoints which are outliers.

There is no outlier present in the data.

a. Is the complete elimination of noise from a dataset possible and desirable? Justify your answer.

<< Any answer with valid reasoning is accepted >>.

b. If you discover outliers in a dataset during exploratory data analysis, how would you decide whether to remove them or keep them in the analysis?

<< Any answer with valid reasoning is accepted >>.

c. Consider two variables, X and Y, with a covariance of 50 and a correlation coefficient of 0.8. What does this information tell us about the relationship between X and Y?

The covariance of 50 indicates that there is a positive linear relationship between X and Y.
whereas
the correlation of 0.8 indicates the strength and direction of this relationship. The correlation of 0.8 represents a strong positive linear relationship between 2 variables.

Question # 2. Data Pre-processing: smoothing noise (15)

Consider the following numeric data. Partition it into bins using equal-width and equal-frequency binning. Bins=3
Smooth the noise by applying bin-by-boundaries and analyze which binning technique worked best for this data.
Show all the necessary steps, no credit will be given for a direct answer.

5 8 10 15 50 72 92 104 215

Your Solution:

Equal-width Binning: $W = \frac{B-A}{K}$ $\quad B=215, A=5, K=3$

$$W = \frac{215-5}{3} = \frac{210}{3} = 70$$

Bin Intervals = $A+w, \; A+2w, \; A+3w, \ldots$

$5+70, \; 5+2(70), \; 5+3(70), \ldots$

$= 75, 145, 215$

Bin 1 = 5, 8, 10, 15, 50, 72

Bin 2 = 92, 104

Bin 3 = 215

Smoothing by Bin-boundaries:

Bin 1 = 5, 5, 5, 5, 72, 72

Bin 2 = 92, 104

Bin 3 = 215

Equal Frequency Binning:

Frequency of each bin = $N/k = 9/3 = 3$

Bin 1 = 5, 8, 10

Bin 2 = 15, 50, 72

Bin 3 = 92, 104, 215

Smoothing by bin-boundaries:

Bin 1 = 5, 10, 10

Bin 2 = 15, 72, 72

Bin 3 = 92, 92, 215

Analysis and Interpretation

Equal-width binning has worked well for this data as it clearly identifies/ Separates extreme values/ outliers in one bin as ~~given~~ in Bin3. The data points which are close to each other are grouped into one bin.

Unlike Equal Frequency binning has combined very distant values in one group.

Question # 3. Chi-square test of independence (10)

Let's consider a hypothetical case where we test the effectiveness of a drug for a certain medical condition. Suppose we have 125 patients under study and 65 of them were treated with the drug. The remaining 60 patients were kept as control samples. The health condition of all patients was checked after a week. The following table shows if their condition improved or not. Just by looking at it, can you tell if the drug had a positive effect on the patients with 5% of significance level. How would you perform the Chi-Square test? Show all the necessary steps to get any credit. Chi-square distribution table is also given below for your reference.

	Responded	Not Responded	*Total*
Treated	45	20	*65*
Not Treated	27	33	*60*
Total	*72*	*53*	*125*

df	P = 0.05	P = 0.01	P = 0.001
1	3.84	6.64	10.83
2	5.99	9.21	13.82
3	7.82	11.35	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52
6	12.59	16.81	22.46
7	14.07	18.48	24.32
8	15.51	20.09	26.13
9	16.92	21.67	27.88
10	18.31	23.21	29.59
11	19.68	24.73	31.26
12	21.03	26.22	32.91
13	22.36	27.69	34.53
14	23.69	29.14	36.12
15	25.00	30.58	37.70
16	26.30	32.00	39.25
17	27.59	33.41	40.79
18	28.87	34.81	42.31
19	30.14	36.19	43.82
20	31.41	37.57	45.32
21	32.67	38.93	46.80
22	33.92	40.29	48.27

Your Solution:

(1) Hypothesis formulation +2

H_o = Drug treatment and health condition of patients are Independent.

H_a = Drug treatment and health condition of patients are dependent.

(2) Calculate Expected values. +2

$e_{11} = \frac{72 \times 65}{125} = 37.44$ $\quad e_{21} = \frac{72 \times 60}{125} = 34.56$

$e_{12} = \frac{53 \times 65}{125} = 27.56$ $\quad e_{22} = \frac{53 \times 60}{125} = 25.44$

	Responded	Not Responded	Total
Treated	45 (37.44)	20 (27.56)	65
Not treated	27 (34.56	33 (25.44)	60
Total	72	53	125

(3) Calculate Chi-Square value. +3

$$\chi^2 = \frac{(45-37.44)^2}{37.44} + \frac{(20-27.56)^2}{27.56} + \frac{(27-34.56)^2}{34.56} + \frac{(33-25.44)^2}{25.44}$$

$$= 1.5265 + 2.0737 + 1.6537 + 2.2466$$

$$= 7.500$$

(4) Calculate Critical value at $\alpha = 5\% = 0.05$ and +2

$$DF = (r-1)(c-1)$$
$$= (2-1)(2-1) = (1)(1) = 1$$

So at $\alpha = 0.05$ and $DF = 1$

$$C.V = 3.84$$

(5) Conclusion +1

Since $\chi^2 > C.V$ $(7.500 > 3.84)$ so we reject the null hypothesis and conclude that Drug treatment and health condition are dependent on each other at 5% significance level.