United International University (UIU)

Dept. of Computer Science & Engineering (CSE)

Midterm: Fall - 2023

Course: CSE 4891 || Data Mining

Marks: 30, Time: 1 hour 45 Minutes

Figures in the right-hand margin indicate full marks.

Any examinee found adopting unfair means will be expelled from the trimester/ program as per UIU disciplinary rules.

---

1. (a) From the observed data, is Gender independent of weight at 5% significance level? Find Chi-Square value and use the Table 2 to find the answer. **[3]**

	About right	Overweight	Underweight	Row Totals
Female	560	163	37	760
Male	295	72	73	440
Column Totals	855	235	110	1200

Table 1: Observed Data

	P				
DF	0.10	0.05	0.01	0.005	0.001
1	2.706	3.841	6.635	7.879	10.828
2	4.605	5.991	9.210	10.597	13.816
3	6.251	7.815	11.345	12.838	16.266
4	7.779	9.488	13.277	14.860	18.467
5	9.236	11.070	15.086	16.750	20.515
6	10.645	12.592	16.812	18.548	22.458

Table 2: Chi Square Distribution

(b) What does cosine similarity of 1 mean? Find the distance/similarity between two documents (Table 3) in terms of Manhattan distance and cosine similarity: **[3]**

Document	Hello	UIU	CSE	Data	Knowledge
Document A	4	1	12	7	9
Document B	11	9	6	4	8

Table 3: Frequency of different words in two documents

2. (a) Find out covariance between two variables ***time*** and ***distance*** from the following sample data: (8, 2), (4, 2), (2, 4), (1, 5), (5, 2). **[2]**
What does a correlation value of 0 mean? Explain briefly.

(b) Why is dimensionality reduction necessary? Explain briefly **[2]**

(c) The boxplot in Figure 1 represents the runs scored by ***Tamim*** and ***Shakib***, find median, interquartile range, maximum and minimum values for both distribution. **[2]**

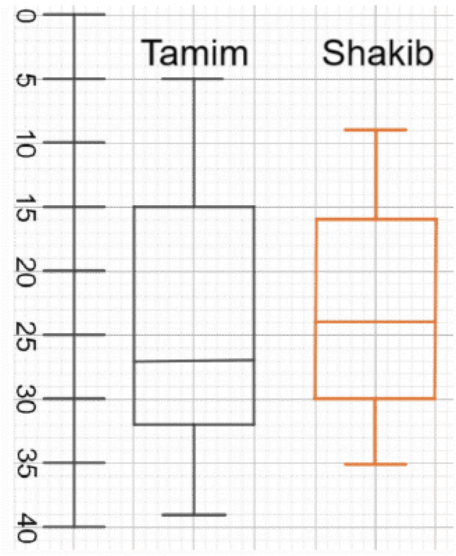

Figure 1: Boxplots

.

3. (a) Apply Apriori algorithm to find frequent item sets from the transactions described in Table 5. **[5]**

Transaction ID	Items
1	A, B, C
2	B, C, D
3	D, E
4	A, B, D
5	A, B, C, E
6	A, B, C, D

Table 5: Transactions

(b) Look at the data in Table 6. Now apply k-nearest neighbor with k=3 and classify a new data point with Age=20, Weight=35. **[3]**

(c) "KNN is a lazy learner" - explain this statement. Explain a disadvantage of being a lazy learner. **[2]**

Age	Weight	Class
40	20	Poor
50	50	Rich
60	90	Rich
10	25	Poor
70	70	Rich
60	10	Poor
25	80	Rich

Table 6: For KNN

Preparation	Captain	Income	Win
Excellent	Bad	High	Yes
Fair	Bad	Low	No
Fair	Bad	High	Yes
Poor	Good	High	No
Excellent	Bad	Low	Yes
Fair	Good	Low	Yes
Poor	Bad	High	No
Poor	Good	Low	No
Fair	Bad	High	Yes

Table 7: For Decision Tree

4. (a) In Table 7, previous results of a BD Cricket Team are provided where ***Preparation, Captain,*** and ***Income*** are input variables and the outcome is ***Win***. Build a decision tree based on the given data by calculating ***Information Gain*** in each step. **[5]**

(b) Given the following confusion matrix in Table 8, find Accuracy, Precision, Recall, and F1-Score. Is measuring only accuracy enough? Why/Why not? **[3]**

	Actually Positive (1)	***Actually Negative (0)***
Predicted Positive (1)	50	10
Predicted Negative (0)	5	100