EAST WEST UNIVERSITY
Department of Computer Science and Engineering
B.Sc. in Computer Science and Engineering Program
Mid Term I Examination, Fall 2020 Semester

Course:	**CSE 447 Data Mining, Section-1**
Instructor:	**Md Mostofa Kamal Rasel, Assistant Professor, Department of CSE**
Full Marks:	**30 (20 will be counted for final grading)**
Time:	**1 Hour and 20 Minutes**

Note: There are **6** (**SIX**) questions, answer ALL of them. Mark of each question are mentioned at the right margin.

---

1. What is data mining? What are the other terminologies referring to data mining? [Mark: 2]

2. A retailer company uses different type of data formats, such as, flat text, csv, MS word, RDBMS, etc. for keeping the records of their products, sales, customers, employees, displayed items, etc. Suppose that the company wants to use a unified system to store and manage their data. They also want to utilize this unified system to learn new knowledge by exploiting the stored data. Therefore, they hire a team of experts to build their desired system. Suppose that you are a member of that team who has expertises on data mining. Given the above scenario of the company, answer the following questions. [Mark: 2x3]

i. **Which** data mining steps will you suggest to extract knowledge?
ii. **Which** technologies are you going to use for extracting knowledge?
iii. **Explain** the major issues that you are going to focus on?

3. Suppose that you are trying to understand the data of a company mentioned above (*see question 2*). Based on your observation on the data, answer the following questions. [Mark: 2x2]

i. **Determine** all data types that might be needed to handle for the data mining system. Explain each data type with at least one example.
ii. Suppose that there are some binary data attributes that should be categorized as either symmetric or asymmetric binary variables. **Describe** the property based on which a binary variable is categorized. **Give** an example for each of symmetric and asymmetric binary variables.

4. Suppose that we have datasets X, T and Z in the Appendix (*see Page 3*). Based on these datasets answer the following questions. [Mark: 2+2+3+2]

i. **Calculate** the mean and median of the dataset X?
ii. **Evaluate** the standard deviation of the dataset X?
iii. Let dataset T represents the symptoms and test results of three patients. Let the values Y and P be 1 and the value N be 0 in dataset T. **Which two patients** have the most probability of having similar disease.
iv. Let dataset Z represents the frequency of representative topics in two different documents (*doc1* and *doc2*). **Determine** the similarity between the two documents.

5. Suppose that the customer income data is not available in the datasets of the above -mentioned retailer company (*see question 2*). **Suggest** some techniques with examples to handle these unavailable data. [Mark: 2]

6. **Evaluate** the output result for the following questions: [Mark: 2+2+2+1]

i. Consider Table S in the Appendix (*see Page 3*), which presents a simplified example of stock prices observed at five time points for Jamuna Electronics and Walton Digi-tech, a high-tech company. If the stocks are affected by the same industry trends, will their prices rise or fall together?.

ii. Use these methods to normalize the following group of data: 120, 130, 140, 90, 100
(a) min-max normalization by setting min = 0 and max = 1
(b) z-score normalization using the mean absolute deviation instead of standard deviation
(c) normalization by decimal scaling

Appendix

X = {7, 12, 5, 8, 5, 9, 13, 12, 19, 7, 12, 12, 13, 3, 4, 5, 13, 8, 7, 6}

T =

Name	**Gender**	**Fever**	**BodyAche**	**Test-1**	**Test-2**	**Test-3**	**Test-4**
Raihan	M	Y	Y	P	N	N	P
Kabir	M	Y	N	N	N	N	P
Samira	F	Y	Y	P	N	N	N

Z =

Doc	Mango	Lichi	Jackfruit	Rain	Guava	Summer	Cold	Hot	Winter
doc1	3	7	0	2	1	1	0	3	0
doc2	1	2	1	1	1	2	2	0	3

Stock prices of Walton Digi-tech and Jamuna Electronics at Dhaka Stock Exchange

S =

Time Point	**Walton Digi-tech**	**Jamuna Electronics**
T1	40	28
T2	25	21
T3	27	19
T4	14	10
T5	12	12