BRAC UNIVERSITY
Department of Computer Science and Engineering
B.Sc. in CS / CSE Program
Mid-Term Exam, Fall 2023

Course: **CSE437 (Data Science: Working with Real World Data) [Sec: 02]**
Full Marks: **25**
Time: **90 minutes**

Note: Course Outcome (CO), Cognitive Level, and Mark of each question are mentioned at the right margin.

---

1. The salary of a software developer depends on experience and skills. The salary of 3 software developers is provided in the following dataset. **Apply** gradient descent to find the optimal values of the learning parameters of a linear machine learning model to predict the salary of any software developer. Consider the sum of squared residual loss function for the assessment of the prediction quality, and learning rate as 0.01, and show only 2 iterations. Encode the categorical variables as High=2 and Low=1. [CO2,C3, Mark: 8 + 2]

Employee	Experience (in years)	Skills	Salary (in thousand taka)
S1	3	High	120
S2	4.5	Low	90
S3	2	High	80

Afterward, **predict** the salary of a highly skilled software developer with 4 years of experience.

2. When you're looking for a community to live in, there are a lot of factors you might consider, e.g., home affordability, proximity to work, and quality of the nearby natural environment. Livability is a measurement of how attractive a neighborhood, city, and/or region is for you based on a variety of factors. [CO2,C2, Mark: 8 + 2]
Form 2 clusters using the K-Means clustering algorithm to group similar livable cities together; considering the *Safeness score, Environment, and Affordability* features of the following dataset. Use cosine similarity function to measure the similarity among the cities. Encode the categorical variables as Excellent=3, Good=2, Poor=1 and Expensive = 2, Affordable=1. **Show** only 1 iteration.

No.	***City***	***Safeness score***	***Environment***	***Affordability***
1	***Tokyo***	88	Excellent	Expensive
2	***Seoul***	86	Good	Affordable
3	***Singapore***	84	Good	Expensive
4	***Delhi***	50	Unhealthy	Affordable
5	***New York***	70	Good	Expensive
6	***Karachi***	40	Unhealthy	Affordable
7	***Amsterdam***	90	Excellent	Affordable

Afterward, **determine** the cluster of Dhaka city from the formed cluster based on the following feature values:

City	*Safeness score*	*Environment*	*Affordability*	*Cluster?*
Dhaka	52	Unhealthy	Expensive	?

3. What does the GINI define? What is the significance of the best splitter in decision tree construction? Why regression tree is called a non-linear model? What is a stump in Ababoost? How to assign the weight of the vote of each stump of the Adaboost classifier? [CO1,C1, Mark: 5]