B.Sc. in EEE, 7th Semester                October 11, 2023 (**Afternoon**) 21

ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)
ORGANISATION OF ISLAMIC COOPERATION (OIC)
**DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING**

| | |
|---|---|
| Mid-Semester Examination | Winter Semester, A. Y. 2022-2023 |
| Course No.: EEE 4709 | Time: 90 Minutes |
| Course Title: Artificial Intelligence and Machine Learning | Full Marks: 75 |

There are 4 (**four**) questions. Answer all 4 (**four**) questions. The symbols have their usual meanings. Programmable calculators are not allowed. Marks of each question and corresponding COs and POs are written in the brackets.

---

1. A group of researchers at IUT has curated a dataset that contains the trip records of Uber cars, bikes, and CNGs around Dhaka city for the year 2022. The data contains information regarding the pick-up and drop-off dates/times as well as locations, trip distances, weather, fares, and number of passengers. There are a total of 10 million records with 9 different attributes in the data. Answer the following questions based on the representative sample of the data provided below.

Table 1: Sample dataset of Uber trip records

| $Loc_{pick}$ | $Loc_{drop}$ | $time_{pick}$ | $time_{drop}$ | dist | Type | Fare | $n_{pas}$ | Weather |
|---|---|---|---|---|---|---|---|---|
| Uttara | IUT | 8.30 | 9.30 | 10 | car | 650 | 2 | Sunny |
| Uttara | IUT | 8.45 | 9.30 | 11 | car | 640 | 3 | Sunny |
| Uttara | Jamuna | 8.45 | 9.30 | 11 | car | 400 | 3 | Sunny |
| IUT | Nilkhet | 12.45 | 14.51 | 30 | car | 960 | 1 | Sunny |
| IUT | Nilkhet | 12.40 | 13.55 | 30 | car | 1220 | 2 | Cloudy |
| IUT | Banani | 15.45 | 18.15 | 17 | car | 1020 | 3 | Rainy |
| IUT | Uttara | 16.00 | 18.00 | 9.5 | cng | 390 | 1 | Rainy |
| Mirpur | BUET | 7.30 | 11.15 | 12 | cng | 600 | 2 | Sunny |
| Uttara | Airport | 10.00 | 10.30 | 5 | bike | 200 | 1 | Cloudy |

a) You are interested to know which are the most frequent destinations of Uber passengers. Is this a Machine Learning (ML) task? – Justify your answer. How do you think you can determine that from the data? **3 (CO1, PO1)**

b) You want to predict the fare of a trip using the data. In that regard, you decide to design an ML model. Sketch the outline (showing step-by-step process) of the ML pipeline you need to complete the task. **3 (CO2, PO1)**

c) One of the researchers claimed that the 'number of passengers' attribute is not related to the fare of the trip, and you should remove that to avoid overfitting your model. State the different ways you can use to assess the exactitude of the claim. 3 (CO1, PO1)

d) The 'trip distance' column had a total of 350 missing entries. You decided to remove those samples directly from the data before splitting it into training and testing folds. However, one of the researchers warned you that it might cause data leakage problem. Do you agree or disagree with the statement of the researcher? – Justify your answer. 3 (CO1, PO1)

e) One issue you faced after training your model is that it provides different results in different runs. One of the researchers suggested you use a '10-fold stratified cross-validation' method to reduce the variance. Do you approve of the suggestion? – Justify your answer. 3 (CO1, PO1)

f) You are not satisfied with the performance of the linear regression model ($RMSE_{train} = 220, RMSE_{test} = 240$). Therefore, to achieve better results, you decided to use advanced regression techniques like Polynomial or Ridge regression. How likely will these algorithms be able to improve performance? – Explain your answer. 3 (CO2, PO1)

g) One of your classmates suggested that you should 'change the currency from taka to dollar and that it will reduce the RMSE score'. Do you agree with the statement? – Justify your answer. 3 (CO1, PO1)

h) One of the researchers wants to design an AI system that would provide an hourly prediction of hotspots (high, medium, low) for Uber depending on the demand. What changes would you have to bring to the ML pipeline to achieve that? – Design such a system. 4 (CO3, PO3)

2. a) "BFS or DFS are stochastic in nature, unlike adversarial search techniques" – do you agree or disagree with the statement? Justify your answer. 5 (CO1, PO1)

b) Explain how the balance between exploration and exploitation is maintained in the Ant Colony Optimization technique. 5 (CO1, PO1)

c) You have collected a cancer patient's dataset with 20 feature variables. You want to train an ML classifier, but first, you want to perform feature selection. Among Fisher's score, RFE, and GA, which one do you think would be more appropriate and likely to provide better results? – Justify your answer. 5 (CO1, PO1)

3. a) Perform K-means clustering and determine the final clusters of the data points given below and the respective cluster centroids. 8 (CO3, PO3)

   i. Data points - P1(1,1), P2(1,3), P3(2,2), P4(4,4), P5(5,5), P6(6,6), P7(3,11).
   ii. Centroids – C1(2,3), C2(1,6).
   iii. Maximum number of iterations = 3

b) Given another data point P(3,3), which category do you think it belongs to? – Justify your answer. 2 (CO2, PO1)

c) Explain what the shape of silhouette plots indicates. Can the silhouette score be negative? If so, explain how. 5 (CO2, PO1)

4. Table - 2 lists a sample of data from a census. Calculate the following using the data. 20 (CO3, PO3)

   i. Entropy of the dataset.
   ii. Gini-index of the dataset.
   iii. Information gain for each of the attributes and decide which splitting criteria should be used as the root node.

Table 2

| ID | Age | Education | Marital Status | Occupation | Annual Income |
|---|---|---|---|---|---|
| 1 | 39 | bachelors | never married | transport | $25K$–$50K$ |
| 2 | 50 | bachelors | married | professional | $25K$–$50K$ |
| 3 | 18 | high school | never married | agriculture | $\leq 25K$ |
| 4 | 28 | bachelors | married | professional | $25K$–$50K$ |
| 5 | 37 | high school | married | agriculture | $25K$–$50K$ |
| 6 | 24 | high school | never married | armed forces | $\leq 25K$ |
| 7 | 52 | high school | divorced | transport | $25K$–$50K$ |
| 8 | 40 | doctorate | married | professional | $\geq 50K$ |