**MID SEMESTER LAB EXAM – S1 MTECH RAU**

**MACHINE LEARNING (15m)**

**Date: 06/12/2021 Time: 75 Mins**

Instructions:

1. For questions 5 and 6, where certain theory/comment questions are asked (they are highlighted in bold), write all your answers as comments/text in **one line only** in the “.ipynb” file itself in Google Colab.
2. You have a total of 1 hour and 15 mins to attempt the exam and upload your “.ipynb” files. Other than this, no extra time will be given. Use your time carefully.

Questions:

**1.** In a popularity contest, n contestants are competing against each other, and people are voting for either of them. Voters don’t communicate with each other while casting their votes. Which of the following ensemble method works similar to above-discussed voting procedure? (0.5m)

A. Boosting

B. Bagging

C. A or B

D. None of these

**2.** In what way/ways can you deal with missing or corrupted data in your dataset? (0.5m)

A. Drop the missing rows or columns

B. Assign a value of 0 to missing values

C. Replace missing values with mean/median/mode

D. All of the above

**3.** In K-Fold cross validation technique, if K = N, then it is called as---------. (0.5m)

**4.** -------happens when our ML model can neither model the training data properly nor produce correct results on the test data. (0.5m)

**5A.** Given the following twenty data points, which specify ten possible combinations of A and B do the following: (4m)

a) Perform **K-means clustering** (k = 2 or 3) on the data and **comment** which K value (2 or 3) according to you gives a better result.

b) Predict a classification for a case where A = 1.523 and B =1.706, using the result of k-means clustering. Verify your result visually with the plot above and **comment** your answer.

A B

2.673 2.396

1.289 2.601

0.347 1.203

0.971 1.567

1.486 0.459

1.221 1.106

2.540 0.429

2.159 1.999

0.703 0.186

0.254 2.152

**5B**. Given the same above twenty data points, do the following: (2m)

a) Perform **Agglomerative** **clustering** (k = 2 or 3) on the data and **comment** which K value (2 or 3) according to you gives a better result (plot the dendrogram with p = 3 to verify your answer).

b) Visually predict from the plot, which class will A = 1.523 and B =1.706 belong to. **Compare** your answers with K Means clustering and **comment** which method is better according to you.

**6A.** For this question, use the “**airlines.csv**” file which is attached below. Do the following: (4m)

a) Convert the 'satisfaction' and 'gender' columns into numerical data.

b) Find the most suitable input feature (x) to predict your output (y) in the dataset using a correlation matrix.

c) Perform **logistic regression** on your dataset with a train-test split of 0.35.

d) Find both the test and train accuracy scores for your model. **Explain** if underfitting or overfitting is happening with your data. If both are not happening, explain why.

e) **Comment** your logistic regression equation (by finding out the intercept and coefficients).

f) Plot your logistic regression graph using the test data.

g) Visually predict and **comment** a classification for a case where entertainment = 0.79, using the plot of your logistic regression graph.

**6B.** For the same dataset as above (**airlines.csv**), do the following: (3m)

a) Perform **decision tree** analysis with the same train-test split as above (0.35).

b) Let the maximum depth of your tree be 2. **Explain** why depth is 2 and not greater than 2.

c) Find both the test and train accuracy scores for your model. **Explain** if underfitting or overfitting is happening with your data, if yes, **comment** one way to overcome overfitting (just mention the name of the method). If both are not happening, explain why?

d) Plot your decision tree using the test data

e) Visually predict and **comment** a classification for a case where entertainment = 0.79, using the plot of your decision tree. Is the answer same as logistic regression?

------ The End ------