Birla Institute of Technology and Science, Pilani
Second Semester 2022-23
CS F415 Data Mining
Mid-Semester Exam

Date: 13.03.2023 **Max. Time: 90 minutes** **Maximum Marks: 50 [25%]**

Note: **Answers to all questions must be analytical, precise and complete.**

1. Write short answers for the following: [3*5=15]
   a. Differentiate bagging and boosting ensemble learning techniques.
   b. Explain the terms overfitting and underfitting
   c. Explain the independence assumption in the context of Naïve Bayes classifier
   d. Give a brief justification why accuracy alone is not a sufficient metric for evaluating classifier performance.
   e. What do you understand by "pruning" in decision trees and why is it required?

2. Solve the following: [4*2=8]
   a. For the given vectors **x** and **y**, calculate: (i) Jaccard coefficient, and (ii) Cosine similarity.
      x = 1 1 0 1 0 0
      y = 0 1 0 0 1 1
   b. Use equi-depth binning (with 3 bins) followed by smoothing by bin boundaries to perform smoothing for this data: 14, 28, 19, 5, 31, 31, 24, 25, 26, 28, 29, 32.

3. Given below is a dataset which tells us whether Ram played football or not, given a certain set of weather conditions (outlook, temperature, humidity and wind). Identify the best split attribute based on Information gain. Show all calculations. [10]

Outlook	Temperature	Humidity	Wind	Played football(yes/no)
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

4. For the dataset given above, use Naïve Bayes classifier to predict whether Ram will play football if the outlook is overcast, the temperature is cool, the humidity is high and the wind is weak. [11]

5. The dataset given below contains two attributes A and B followed by a class label. Use the k-nearest neighbor classifier to find the class of the last tuple according to its (a) 1-nearest neighbor, and (b) 3-nearest neighbors. Show all steps involved. [3+3=6]

A	B	Class
25	130	X
35	155	Y
30	130	X
40	165	Y
35	140	X
55	190	Y
40	145	?

****