Bangabandhu Sheikh Mujibur Rahman Science and Technology University
Department of Computer Science and Engineering
4th Year 2nd Semester Mid-Term Examination-2020
Course Title: Machine Learning Course Code: CSE451

Time: 01 Hour **Marks:20**

*Answer any **two (02)** from the following **three (03)** questions.*

1. (a) Decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes. Referring to the example below where we want to classify the instances based on the target feature Y. In the example, we split the instances using three input features A, B, and C. Now, you need to identify which feature split produces more homogeneous sub-nodes using the Information Gain technique. 5

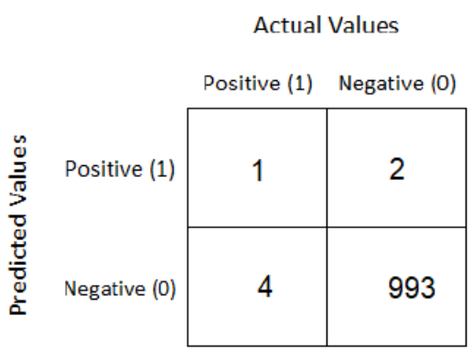
A	**B**	**C**	**Y**
1	1	1	1
1	1	0	1
0	0	1	0
1	0	0	0

(b) Let us consider the following weather data which consists of 1 feature variable (Weather) and 14 instances. You need to calculate the probability of playing cricket when the weather is overcast using Naïve Bayes Classifier. 5

Weather	**Play**	**Weather**	**Play**	**Weather**	**Play**	**Weather**	**Play**
Sunny	Yes	Overcast	Yes	Sunny	Yes	Overcast	No
Sunny	Yes	Rainy	No	Overcast	Yes	Rainy	Yes
Overcast	Yes	Rainy	No	Sunny	Yes		
Rainy	No	Sunny	Yes	Rainy	No		

2. (a) How to choose the value of k in kNN learning? Why is an odd number of k chosen for the binary classification? 3+2

(b) Consider the following confusion matrix for a binary classification model. Find accuracy and F1 score. Explain why accuracy is not enough for this classification? 3+2



3. (a) What are overfitting and underfitting in machine learning models? How these two are related to bias and variance? 2+2

(b) Distinguish between bagging and boosting ensemble methods? Is Random Forests Algorithm based on bagging or boosting? Justify your answer. 3+3