Assignment 1

(COMP3605 - Introduction to Data Analytics, 2023-2024)

Date Available: Sunday, October 1, 2023
Due Date: 11:50 PM, Sunday, October 15, 2023
Total Mark: 100 marks

Answer ALL Questions

INSTRUCTIONS
1. Type or write your answers neatly.
2. Show all working of your answers.
3. Your solutions must be your own. You must not share your working or solutions with your peers.
4. You are not permitted to copy, summarize, or paraphrase the work of others in your solutions.
5. Submit your answers in a single zipped file named A1_ID.zip to the email comp3605@gmail.com, where ID is replaced with your student ID. The file A1_ID.zip contains
- a single PDF file containing all of your typed, handwritten, and screenshots answers.
- a signed and dated UWI Plagiarism Declaration indicating that the work submitted is your own.

Question 1
You are given a training data set *D* shown in the table below for a binary classification problem. The class label attribute **Concern** has two different values $\{C_1, C_2\}$ = {satisfied, dissatisfied}. Possible values of **Weight**, **Height**, **Food Habit**, and **Fitness Program** are {light, average, heavy}, {short, average, tall}, {no restriction, no red meat, vegetarian} and {Yes, No}, respectively.

The Class-Labeled Training Data Set *D*

No.	Weight	Height	Food Habit	Fitness Program	Concern
1	average	short	vegetarian	Yes	satisfied
2	average	short	no restriction	No	dissatisfied
3	heavy	average	no red meat	No	dissatisfied
4	heavy	tall	vegetarian	No	satisfied
5	heavy	average	vegetarian	No	satisfied
6	light	short	no restriction	Yes	satisfied
7	light	average	no restriction	No	dissatisfied
8	average	tall	no restriction	Yes	satisfied

a. Use a multiway split to compute the information gain for the attribute **Weight**. [10 marks]
b. Compute the gain ratio for the attribute **Height** using *Gain*(**Height**) = 0.2657. [10 marks]
c. Use a binary split to compute the Gini index for the attribute **Food Habit** and the splitting subset {no restriction, no red meat}. [10 marks]
[**Total mark**: **30**]

PLEASE TURN TO THE NEXT PAGE

Question 2
You are given a training dataset *D* shown in the table below for a binary classification problem.

The class-labeled training dataset *D*

No.	Symptom	Body Temperature	Vaccinated	Lockdown Status	Going out
1	Cough	Normal	false	Full	Yes
2	Cough	Mild	true	Free	Yes
3	Cough	Normal	true	Free	Yes
4	Sore throat	Normal	true	Free	Yes
5	Sore throat	High	false	Free	No
6	Fever	Mild	true	Partial	No
7	Fever	Mild	true	Free	Yes
8	Fever	High	false	Free	No
9	Fever	High	true	Free	Yes
10	Cough	High	false	Full	No
11	Cough	High	false	Partial	No
12	Sore throat	Mild	true	Partial	Yes
13	Sore throat	High	true	Free	No
14	Cough	Normal	true	Full	No
15	Sore throat	High	true	Free	Yes

The attributes are **Symptom**, **Body Temperature**, **Vaccinated**, **Lockdown Status**, and **Going out**. The **Going out** attribute is the class label attribute (i.e., target feature) that has two different values {Yes, No}. Let C_1 = Yes and C_2 = No.
Given the test instance $\boldsymbol{X}$ = (**Symptom** = Sore throat, **Body Temperature** = Normal, **Vaccinated** = false, **Lockdown Status** = Partial). What the class label will a naive Bayesian classifier predict for the given test instance $\boldsymbol{X}$?

[Total mark: **40]**

Question 3 Consider a training data set *D* that contains p = 60 positive examples and n = 100 negative examples. Suppose that we are given the following two candidate rules.
Rule r_1: covers p_1 = 50 positive examples and n_1 = 5 negative examples,
Rule r_2: covers p_2 = 2 positive examples and n_2 = 0 negative example.
Which rule is better according to
a. the *accuracy* metric? [5 marks]
b. the *coverage* metric? [5 marks]
c. the *FOIL_Gain* metric? Assume that the initial rule R_0: {} → + covers p_0 = 60 positive examples and n_0 = 100 negative examples. [6 marks]
d. Laplace metric? [7 marks]
e. the likelihood ratio statistic *R*? [7 marks]
[**Total mark**: **30**]

End of Assignment 1