



**UNIVERSITI
TEKNOLOGI
PETRONAS**

EXTENDED ASSIGNMENT JANUARY 2021 SEMESTER

**COURSE : TEB2164 – INTRODUCTION TO DATA
SCIENCE**

DATE : 2nd APRIL 2021

TIME : 15:00 – 14:59

INSTRUCTIONS TO CANDIDATES

1. The Extended Assignment (EA) is an open-book assessment. Students can refer to online resources, learning materials, textbooks, and other reading materials to answer the questions that have been posted in the assessment.
2. Answer **ALL** questions.
3. The duration to complete the EA is **TWENTY-FOUR (24) HOURS**.
4. Students are allowed **ONE (1)** attempt to do the EA successfully where only **ONE (1)** duly completed EA submission is permitted. Multiple submissions are **NOT** allowed.
5. **MAXIMUM** file size for your EA submission to be uploaded to ULearn is **20MB**.
6. Please **upload** your answers in **ONE (1) PDF file**.
7. Please make sure your answer in the PDF file is **clear and readable** and name your file as follows: **"your name_your ID_EA Answer"**
8. Late submission and unclear/unreadable answer will not be accepted.

NOTE: You are required to submit **"CERTIFICATION OF ORIGINALITY"** in the first page of your answer sheet.

1. a. Consider the structure of training data with 32 rows as shown in **TABLE Q1** for a classification problem with four possible classes.

TABLE Q1: Structure of Training Data

ID	<attr1>	<attr2>	<attr3>	<attr4>	Class
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					
25					
26					
27					
28					
29					
30					
31					
32					

Fill up appropriate headers for <attr1>, <attr2>, attr3> and <attr4> according to attributes of your own training data. Headers for ID and Class have been provided. Later, fill up data corresponding to the four attributes and the four-option Class. The type of data for <attr1> is binary, <attr2> is continuous, <attr3> is nominal, and <attr4> is ordinal. The values for each attribute must be diverse. Based on the training data that have been furnished, compute the

Entropy for the overall collection of the training data, the attribute with binary data, the attribute with continuous data using multiway split, the attribute with nominal data using multiway split, and the attribute with ordinal data using multiway split. The split breakdowns for attributes requiring multiway split must be clearly indicated. Lastly, suggest with justification which attribute in the training data is the most heterogenous.

[35 marks]

- b. Suppose that you have been hired by a digital news agency to summarize top-10 daily news on a specific vertical such as computing, medicine, finance, entertainment, or law in Malaysia. As a junior data scientist, suggest a complete text mining process that you will perform to achieve the goal.

[15 marks]

2. **TABLE Q2** displays an unfilled temperature readings summary from ABC weather station in East Borneo comparing October, November, and December from 1990 till 1999 to that of from 2010 till 2019. The table should display the number of months in which the average maximum daily temperature was low ($< 16^{\circ}\text{C}$), medium, or high ($> 26^{\circ}\text{C}$). The investigation aims to discover whether a significant difference between the two rows exists.

TABLE Q2: Temperature Readings Summary

	Low	Medium	High
1990 - 1999			
2010 - 2019			

Firstly, furnish **TABLE Q2** with data in the *Low*, *Medium*, and *High* columns. The data for 1990-1999 must be unique from that of 2010-2019.

Assuming that the readings are independent from month to month, let unknown parameters $p_{d,m}$ be the probability that a month's reading goes to bin $m \in \{Low, Medium, High\}$ in decade $d \in \{1990 - 1999, 2010 - 2019\}$. As a junior data scientist, you have been requested to (i) provide expressions for the maximum likelihood estimates $\hat{p}_{d,m}$, stating what to maximize and over which variables, (ii) establish a null hypothesis H_0 such that the probabilities are identical in both 1990-1999 and 2010-2019 and these probabilities are called q_k to provide the maximum likelihood estimates \hat{q}_k under H_0 , perform a test onto H_0 using the test statistics given as

$$t = \sum_{d,m} \frac{(\hat{p}_{d,m} - \hat{q}_m)^2}{\hat{q}_m}$$

and (iii) considering parametric sampling to compute the distribution for t under H_0 . Additionally, your tasks also include (iv) explaining the relevance of one-sided test vs two-sided test for this investigation, (v) providing pseudocode to compute the p-value for the H_0 test, and finally (vi) explaining an advantage and

a disadvantage of a count-based test as opposed to a linear regression-based test.

[50 marks]

-END OF PAPER-