UNIVERSITI
TEKNOLOGI
PETRONAS

FINAL EXAMINATION
MAY 2024 SEMESTER

COURSE : TFB2063 - DATA SCIENCE

DATE : 5 AUGUST 2024 (MONDAY)

TIME : 9:00 AM - 12:00 NOON (3 HOURS)

INSTRUCTIONS TO CANDIDATES

1. Answer **ALL** questions in the Answer Booklet.
2. Begin **EACH** answer on a new page in the Answer Booklet.
3. Indicate clearly answers that are cancelled, if any.
4. Where applicable, show clearly steps taken in arriving at the solutions and indicate **ALL** assumptions, if any.
5. **DO NOT** open this Question Booklet until instructed.

Note :

i. There are **NINE (9)** pages in this Question Booklet including the cover page

ii. **DOUBLE-SIDED** Question Booklet.

1. a. Discuss any **THREE (3)** processes involved in the data science methodology.

[6 marks]

b. Explain any **THREE (3)** other characteristics of big data besides volume, velocity, and variety.

[6 marks]

c. Differentiate between types of data and data types

[4 marks]

d. Discuss the potential impacts of data quality on ethical decision making in data analysis

[4 marks]

2. a. **TABLE Q2** shows the data frame called a `dfPlayers` for the Players Profile.

TABLE Q2: Players Profile

Player Name	Age	Club	Height (cm)	Weight (lbs)	Foot	Joined
Pierre-Emerick	29	Arsenal	189	176	Right	Jan 31, 2018
Alexandre	27	Arsenal	179	161	Right	Jul 5, 2017
Bernd Leno		Arsenal	192	183	Right	Jul 1, 2018
Mkhitaryan	29	Arsenal	156	161	Both	Jul 5, 2016
Granit Xhaka	26	Arsenal	185	500	Left	Jul 1, 2016
Shkodran		Arsenal	182	181	Right	Aug 30, 2016
Jack Grealish	23	Aston Villa	179	148	Right	Mar 1, 2008
John McGinn	24	Aston Villa	179	154	Left	Aug 8, 2018
Anwar El Ghazi	50	Aston Villa	188	150	Right	Jan 31, 2017
Conor	27	Aston Villa	159	350	Left	Jan 26, 2017
James Chester	30	Aston Villa	155	174	Right	Aug 12, 2016
James Chester	30	Aston Villa	155	174	Right	Aug 12, 2016
James Chester	30	Aston Villa	155	174	Right	Aug 12, 2016
Jonathan Kodjia	46	Aston Villa	182	192	Right	Aug 30, 2016
Callum Wilson	26	Aston Villa	162	148	Right	Jul 4, 2014

Using the `dfPlayers` data frame, write R code for the following:

i. Remove any duplicated entries.

[3 marks]

ii. Identify and handle any outliers in the "Age" and "Weight" columns and impute using the median value of each column.

[NOTE: Assume the players age ranges between [18 and 38] and the weight range between [135 and 185 lbs] **].**

[4 marks]

iii. Identify any missing values in the " Age" column using `is.na()` function and impute using the mean value.

[3 marks]

iv. Remove the "Foot" column from the `dfPlayers` data frame

[3 marks]

v. Use `rbind()` function to add the following player to the `dfPlayers` data frame

Salah	32	Liverpool	175	170	Both	Jul 5, 2016

[3 marks]

b. Using the `dfPlayers` data frame in **part (a),** determine the output of the following code snippet:

i.
```
dfPlayers [c(1,3), ]
```

[2 marks]

ii.
```
tail(dfPlayers,3)
```

[2 marks]

3. a. Assume that the model produces the classification results as shown in **FIGURE Q3a**. Determine the model's accuracy.

```
> table(tree.pred, iris$Species)

tree.pred     setosa versicolor virginica
  setosa          50          0         0
  versicolor       0         47         1
  virginica        0          3        49
```

FIGURE Q3a: Classification Prediction Results

[4 marks]

b. You are to develop a classification model, using a training dataset that contains a target variable with the values of 1s and 0s. **FIGURE Q3b** shows the counts of 1s and 0s in the target variable.

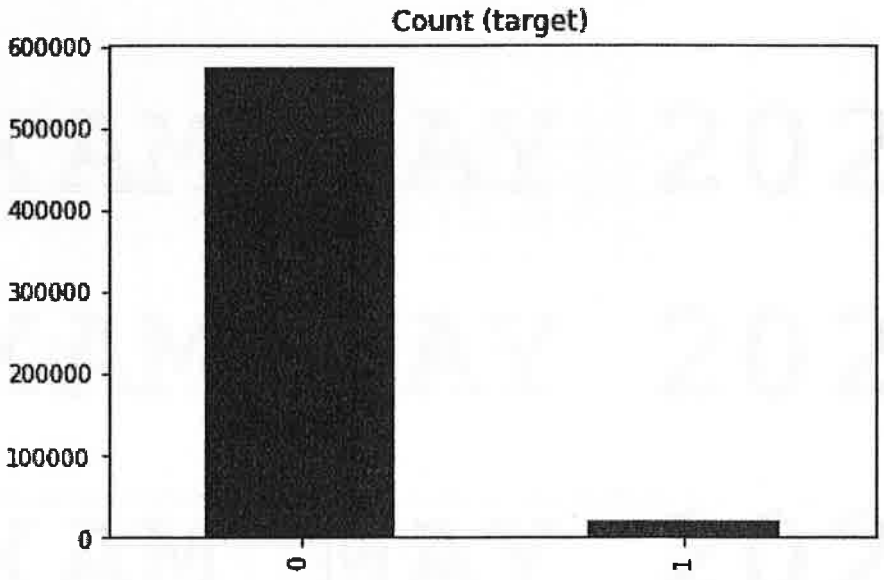

FIGURE Q3b: The Count of 1s and 0s in The Target Variable

i. Identify the issue with the dataset and explain how it affects your classification modeling.

[3 marks]

ii. Analyze the issue in **part (b)(i)** and explain **TWO (2)** techniques to resolve it.

[4 marks]

c. Given the correlation heatmap shown in **FIGURE Q3c** and TAX as the target variable. Identify **THREE (3)** independent variables to be selected as features for modeling. Justify each selection.

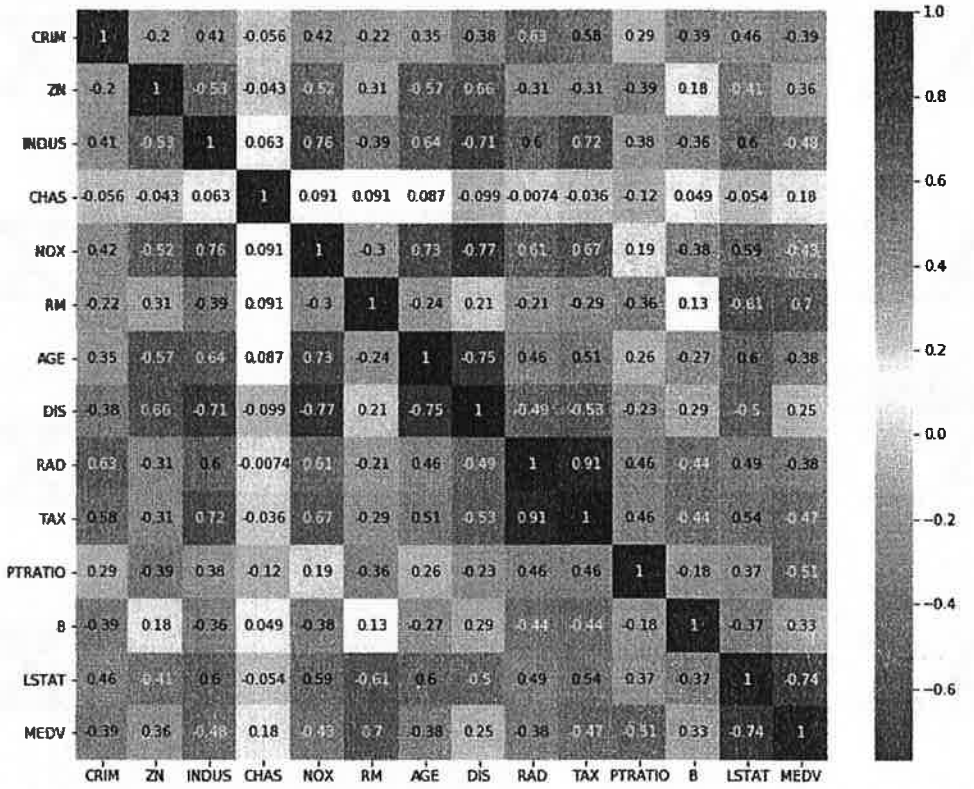

FIGURE Q3c: Correlation Heatmap

[6 marks]

d. List **THREE (3)** benefits of performing feature selection before modelling.

[3 marks]

4. a. **TABLE Q4** shows a dataset consisting of 7 transactions.

TABLE Q4: The dataset.

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Create a k-means data clustering model using the dataset in **TABLE Q4.** Use the Euclidean distance function, K=2, and suppose individuals 1 and 4 are selected as the initial means.

[8 marks]

b. Given the decision tree model's structure shown in **FIGURE Q4.** Explain the model in terms of splitting rules, terminal nodes and how the classes are determined based on the splitting rules.

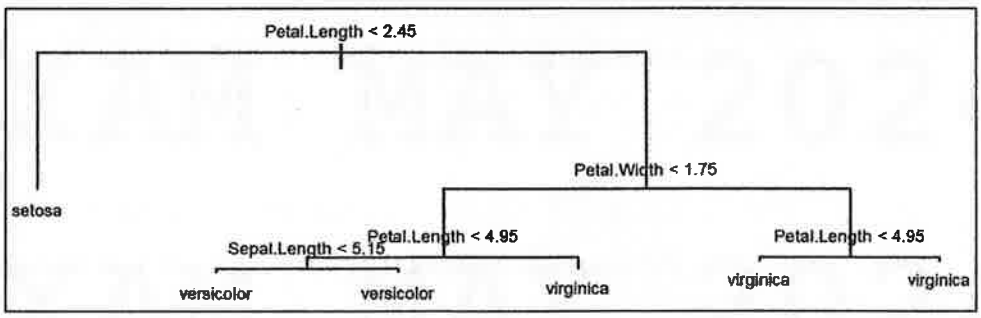

FIGURE Q4. Decision Tree Model's Structure

[8 marks]

c. Define any **FOUR (4)** preprocessing steps that can be applied to the data for predictive modelling.

[4 marks]

5. a. Assume that a simple linear regression prediction model named as `lm_happiness` is developed to predict a person's happiness given his/her monthly income. Using R, the properties of the model are shown in **FIGURE Q5**.

```
Call:
lm(formula = happiness ~ income, data = income.data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.02479 -0.48526  0.04078  0.45898  2.37805

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.20427    0.08884   2.299   0.0219 *
income       0.71383    0.01854  38.505   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7181 on 496 degrees of freedom
Multiple R-squared:  0.7493,  Adjusted R-squared:  0.7488
F-statistic:  1483 on 1 and 496 DF,  p-value: < 2.2e-16
```

i. Using the simple linear regression model, predict the person's happiness given the income of RM 15k.

[4 marks]

ii. Using the `lm_happiness` model, write the R code snippet to implement the prediction of the income in **part (a)(i).**

[4 marks]

b. Assume that we have the following dataset with **one** dependent variable ***Y*** and **two** independent variables X_1 and X_2.

TABLE Q5: The dataset.

y	X_1	X_2
140	60	22
155	62	25
159	67	24
179	70	20
192	71	15
200	72	14
212	75	14
215	78	11

i. Create a multiple linear regression model on the provided dataset.

[4 marks]

ii. Discuss **ONE (1)** key difference between the multiple and simple linear regression.

[2 marks]

c. List any **THREE (3)** of the evaluation metrics used to measure the performance of the designed models.

[3 marks]

d. Outline the consequences of not having proper data collection.

[3 marks]

- END OF PAPER -