UNIVERSITI TEKNOLOGI PETRONAS

FINAL EXAMINATION
JANUARY 2025 SEMESTER

COURSE : FEM2063/FFM2063 - DATA ANALYTICS

DATE : 8 APRIL 2025 (TUESDAY)

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 **ELEVEN (11)** pages in this Question Booklet including the cover page and the appendix.

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

1. a. A grocery store chain wants to use data analytics to improve sales, customer satisfaction, and inventory management. The store collects data on daily sales, customer preferences, stock levels, and seasonal trends. The store considers using the four types of Big Data analytics: Descriptive, Diagnostic, Predictive, and Prescriptive Analytics.

   i. For each type of analytics, provide **ONE (1)** example to improve the store's business.

   [4 marks]

   ii. Identify any Big Data Analytics type that would be helpful in increasing the daily sales. Justify your answer.

   [2 marks]

   iii. Describe **ONE (1)** data analytics challenge when applying predictive analytics for store sales.

   [2 marks]

   b. **TABLE Q1** provides data on daily temperature, humidity levels, and public satisfaction with the weather in Kuala Lumpur.

TABLE Q1: Weather in Kuala Lumpur

Day	Temp. (C)	Humidity (%)	Weather Condition	Satisfaction Level	Rain Alert
Monday	33	70	Sunny	Satisfied	No
Tuesday	29	85	Rainy	Unsatisfied	Yes
Wednesday	31	75	Cloudy	Neutral	No
Thursday	34	60	Sunny	Very Satisfied	No
Friday	28	90	Thunder storm	Very Unsatisfied	Yes
Saturday	30	80	Cloudy	Neutral	No

i. Identify the data type for each variable in **TABLE Q1**. Justify your answer.

[10 marks]

ii. One of the 5V's of Big Data plays a crucial role in weather forecasting. Explain which "V" is the most important for real-time weather monitoring in Kuala Lumpur. Justify your answer.

[2 marks]

2. Research aims to develop multiple linear regression (MLR) models to predict the delivered thermal energy consumption of school buildings in the Federation of Kuala Lumpur. The study compares annual analytical (Y_{ana}) and empirical (Y_{emp}) thermal energy consumption (kWh/year) to assess their alignment with the building needs. **TABLE Q2** shows the results of the MLR models.

TABLE Q2: Analytical vs Empirical Delivered Thermal Energy

Explanatory variable	Coefficients Y_{ana}	Coefficients Y_{emp}
Constant	- 115,650.1	787,268.9
X1 = Number of heating degree in °C	83.73	247.85
X2 = Total number of users	147.445	N/A
X3 = Year of construction of building	N/A	- 1520.26
X4 = Building shape factor	- 38,876.24	N/A
X5 = Windows area of the envelope	N/A	- 348.28
X6 = Walls area of the envelope	N/A	N/A
X7 = Transmission heat transfer	N/A	N/A
X8 = Total heat transfer	15.29	62.74
X9 = Degree of efficiency	N/A	- 568,583.85
X10 = Daily hours of operation	16,029.65	N/A
X11 = Unit cost for heating	- 103,179.4	N/A

N/A: not applicable.

a. Construct **NULL** hypotheses for the Y_{ana} and Y_{emp}.

[2 marks]

b. Construct regression models for dependent variables Y_{ana} and Y_{emp} based on the coefficients provided for each model as given in **TABLE Q2**.

[4 marks]

c. Regression models in **part (b)** were developed after all the insignificant variables were removed. Provide **TWO (2)** common methods used in selecting the significant variables. Explain your answers.

[2 marks]

d. From ANOVA analysis it was found that R^2 for Y_{ana} is 0.70 while R^2 for Y_{emp} is 0.897. Explain the impact of the respective R^2 to each regression model.

[2 marks]

e. Explain any **TWO (2)** possible factors that may have affected the R^2 value given in **part (d).**

[2 marks]

f. During the development of the linear model for Y_{ana}, the parameter X_2 (total number of users) was removed unintentionally. Discuss how this removal may impact values of R^2, standard error and F-value. Justify all your answers.

[4 marks]

g. The variables X_9, X_{10}, and X_{11} are the largest contributions towards analytical and empirical delivered thermal energy. Explain how each of these parameters impacts the regression models.

[4 marks]

3. A manufacturing plant monitors its machines using three key operational parameters; X_1 (Motor Temperature in °C), X_2 (Vibration Level in mm/s), and X_3 (Oil Pressure in kPa). The plant categorizes machines into three health conditions (Y); 1 = Healthy (Normal Operation), 2 = Warning (Needs Inspection), and 3 = Faulty (Requires Maintenance). The historical data from different machines is recorded as given in **TABLE Q3**.

TABLE Q3: Machines records

Point (p_i)	X_1	X_2	X_3	Y	Distance	
					Euclidean	Manhattan
1	10	2	320	1	$\boldsymbol{d(q, p_1)}$	60
2	20	4	310	2	$\boldsymbol{d(q, p_2)}$	38
3	30	6	300	2	15.03	16
4	40	8	290	2	11.22	16
5	50	10	280	3	20.83	28
6	30	12	270	1	33.91	$\boldsymbol{d(q, p_6)}$
7	70	14	260	3	47.68	$\boldsymbol{d(q, p_7)}$

Using KNN classification, a technician should identify the necessary maintenance action needed for a new machine (q) operating under the following conditions ($q: X_1 = 45\ ^{\circ}\text{C},\ X_2 = 7\ \text{mm/s}, X_3 = 300\ \text{kPa}$).

a. Suggest the initial number of neighbours needed for the classification.

[2 marks]

b. Calculate the unknown distances between the query machine (q) and the neighbour points in **TABLE Q3** (Note: there are two different distance functions used).

[8 marks]

c. Conclude the proper maintenance action needed for the query machine (q) when using $K = 3$, and $K = 5$ neighbours. Justify your answer.

[4 marks]

d. Identify the effect of Oil Pressure variable on the distance function and suggest a proper solution to that effect if any.

[2 marks]

e. Using the Min-Max function, scale each of the neighbour points P_1 and P_5.

[4 marks]

4. a. Explain the use of the python codes ([1] to [5]) given in **TABLE Q4a.**

[5 marks]

TABLE Q4a: Python code

	`from sklearn.ensemble import RandomForestClassifier`
	`from sklearn.datasets import load_iris`
	`from sklearn.model_selection import train_test_split`
	`from sklearn.metrics import accuracy_score`

[1]	`iris = load_iris()` `X, y = iris.data, iris.target`
[2]	`X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)`
[3]	`rf_model = RandomForestClassifier(n_estimators=100, random_state=42)` `rf_model.fit(X_train, y_train)`
[4]	`y_pred = rf_model.predict(X_test)`
[5]	`accuracy = accuracy_score(y_test, y_pred)` `print(f"Random Forest Accuracy: {accuracy}")`

b. **TABLE Q4b** shows the data for playing tennis with Outlook and Humidity as input variables and Play Tennis (Yes or No) as the target variable. Calculate the Entropy value for the following variables, Outlook where Sunny = Yes, and Humidity where Normal = No.

[5 marks]

TABLE Q4b: Data for playing tennis

Day	Outlook	Humidity	Play Tennis
1	Sunny	High	Yes
2	Rain	High	No
3	Sunny	Normal	Yes
4	Sunny	High	Yes
5	Rain	Normal	No
6	Sunny	High	No

c. **TABLE Q4c** shows the PCA eigenvalues results.

TABLE Q4c: PCA eigenvalues results.

Component	Eigenvalue	Percentage of variance explained	Cumulative percentage of variance explained
1	0.3251		
2	0.2279		
3	0.1168		
4	0.0122		
5	0.0086		
6	0.0067		

i. Calculate the missing cells in **TABLE Q4c.**

[6 marks]

ii. Sketch the cumulative percentage variance explained. Justify how many principal components shall be retained.

[4 marks]

5. **TABLE Q5** shows health data of 5 persons. The study aims to cluster the data into two classes: Normal and Overweight class.

TABLE Q5: Health data

Person No.	Weight in kg	Height in cm
P_1	60	157
P_2	65	152
P_3	67	170
P_4	75	175
P_5	82	167

a. Using the K-Means algorithm with Manhattan distance and considering P_1 and P_5 as the initial centroids for Normal and Overweight classes respectively, determine the **TWO (2)** clusters after the first iteration.

[7 marks]

b. Determine the new centroid of each cluster from **part (a)**.

[4 marks]

c. Construct the Manhattan distance matrix for each pair of points $d(p_1, p_2)$ to $d(p_4, p_5)$ in **TABLE Q5.**

[4 marks]

d. Using the constructed **Manhattan** distance matrix in **part (c)** and the Single Linkage proximity function in hierarchical clustering, determine the first **TWO (2)** linkages.

[5 marks]

– END OF PAPER –

APPENDIX

1. Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$$

2. Logistics Regression

$$\hat{P} = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}}$$

3. Discriminant Analysis

$$\pi_k = \frac{n_k}{n}$$

$$\mu_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\sigma^2 = \frac{1}{n-k} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \mu_k)^2$$

$$\delta_k(x)_{discriminant} = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

4. Euclidian Distance

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

5. Manhattan Distance

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$$

6. Single Linkage

$$D(X,Y) = \min_{x \epsilon X, y \epsilon Y} d(x,y)$$

7. Normalization function

$$Scaled\ x = \frac{x - min}{\max - min}$$

8. Entropy

$$Entropy \ = \sum_{i=1}^{c} -P_i * log_2(P_i)$$