UNIVERSITI TEKNOLOGI PETRONAS

FINAL EXAMINATION
MAY 2023 SEMESTER

COURSE : **FEM3043 - BIG DATA ANALYTICS**

DATE : **16 AUGUST 2023 (WEDNESDAY)**

TIME : **2:30 PM - 5:30 PM (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 **SIX (6)** pages in this Question Booklet including the cover page .

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

1. a. Explain the need for HADOOP with **ONE (1)** example.

[3 marks]

b. Draw the Hadoop HDFS, MapReduce, and YARN architectures.

[9 marks]

c. Differentiate between logical and physical data models with an illustration of each data model.

[8 marks]

2. a. As a data scientist tasked with developing pipeline corrosion predictions for Rahmat Oil Sdn Bhd, it is essential to propose a standardized data science methodology for predictive analytical projects. This methodology should provide a clear project execution plan that the team can understand. The following development phases should be included in your proposal to outline the process of creating pipeline corrosion predictions:

i. Business Understanding and Data Understanding [4 marks]

ii. Data Preparation [3 marks]

iii. Modelling and Training [4 marks]

iv. Evaluation [3 marks]

b. Describe the following terminologies with **ONE (1)** example for each.

i. Supervised Learning [2 marks]

ii. Unsupervised Learning [2 marks]

iii. Reinforcement Learning [2 marks]

3. a. Compare the following analytics with an example based on big data from the agriculture industry.

i. Descriptive analytics versus diagnostic analytics

[5 marks]

ii. Predictive analytics versus prescriptive analytics

[5 marks]

b. As a data scientist, you have realized that 30% of data taken from your company's Electronic Data Warehouse shows signs of data being classified as "corrupt." Recommend **TWO (2)** solutions with justification to ensure that the data has the attribute of high veracity.

[10 marks]

4. In the future, cancer prediction using AI is expected to become more accessible, thereby eliminating the need for hospital visits. Various technologies are currently being utilized and tested in the medical field to facilitate the prediction of cancer. In breast cancer, certain attributes such as clump thickness, uniform cell size, and uniform cell shape are considered valuable, providing insights into whether the cancer is malignant or benign. Similarly, for lung cancer, indicators such as smoking, yellow fingers, anxiety, and peer pressure are considered significant. In prostate cancer, attributes such as radius, texture, perimeter, and area are deemed useful, with the predicted outcome indicating the likelihood of being affected by either type of cancer.

a. Examine the key challenges faced in utilizing big data for cancer predictive analytics, considering the following factors:

i. Management challenges

[5 marks]

ii. Process challenges

[5 marks]

b. Design **TWO (2)** suitable charts for the cancer predictive analytics dashboard:

i. Relationship in data

[5 marks]

ii. Compare data.

[5 marks]

5. The effectiveness of sentiment analysis in data analytics has enabled large organizations, government officials, and governmental bodies to gain insights into how their audiences respond to relevant social media posts, leveraging the vast amount of data generated. However, to uphold privacy and safeguard individual rights, data scientists are ethically obligated to prioritize these concerns.

a. The privacy rights of individuals are put at risk by sentiment analysis, which involves automatically analyzing texts to determine positive or negative feedback. Support the statement using **TWO (2)** examples.

[8 marks]

b. Choose and justify **ONE (1)** learning approach suitable for sentiment analysis.

[4 marks]

c. Determine and explain **TWO (2)** machine learning techniques that are appropriate for sentiment analysis.

[8 marks]

-END OF PAPER-