The Open University of Sri Lanka
Faculty of Engineering Technology
Department of Electrical and Computer Engineering

Study Programmes	: Bachelor of Software Engineering Honours
Name of the Examination	: Final Examination
Course Code and Title	: **EEX4373- Data Science**
Academic Year	: 2023/24
Date	: 23rd August 2024
Time	: 9:30 – 12:30 hrs

General Instructions

1. Answer all questions.
2. This is a Close Book Test (CBT).
3. Do not use a red pen.
4. The maximum mark allocations are given in square brackets.
5. Read all instructions carefully before answering the questions.
6. Clearly state your assumptions, if any.
7. When writing code snippets, you can decide the Programming language. E.g. – Python/R.

Question 1 [20 Marks]

Write answers with suitable code examples

A. What is meant by data imbalance/class imbalance? [5 Marks]
   a. Explain with a real-world example where data imbalance can occur. Illustrate your explanation with visual aids such as graphs or charts.
   b. Write appropriate code snippets to explain your answer.

B. What do you mean by outlier detection/anomaly detection. [5 Marks]
   a. Explain with a proper real-world example. Hint – Draw a suitable diagram and explain the theory. Write appropriate code snippets to explain your answer.

C. Define "**classification algorithms**" in your own words. [5 Marks]
   a. Provide an example and explain it.

D. Show a basic calculation of how k-mean algorithm performs the clustering. [5 Marks]

Question 2 [20 marks]

Classification algorithms and clustering algorithms are used to solve different machine learning problems. A company wants to understand its customer base better. They use one approach to categorize customers into different risk levels based on their past behaviors.

A. Identify and explain three critical factors to consider when implementing **classification algorithms**. *Hint - For each factor, Provide an example of how it affects the algorithm's performance in the above applications. For each factor, Write sample code snippets for selected factors.* [6 Marks]

B. Compare and contrast two **clustering algorithms**. *Hint - Write appropriate code snippets to explain your answer.* [7 Marks]

C. Differentiate between **classification algorithms** and **clustering algorithms**. *Hint - Use examples to evaluate the effectiveness of each type in specific real-world applications, explaining how the choice of algorithm impacts the results. Provide suitable codes (R or Python) for further explanations.* [7 Marks]

Question 3 [20 marks]

Random forest is a popular ensemble learning method used for both classification and regression tasks. Assume you are analyzing a dataset to optimize a marketing campaign using Decision Trees and Random Forest algorithms. Below is a sample dataset that you received.

Customer Age	Campaign Engagement Score	Purchase History	Gender	Region
25	80	High	Male	Urban
30	65	Medium	Female	Suburban
22	90	High	Female	Urban
35	55	Low	Male	Rural

The meaning of column names is as below.

Customer Age – The age of the customer.

Campaign Engagement Score - A numeric score representing how engaged a customer is with a marketing campaign. Higher scores generally indicate higher engagement levels.

Purchase History - A categorical variable indicating the level of a customer's purchase activity. i.e. usually a high/medium or low number of purchases per given time frame.

Gender - The gender of the customer.

Region - The geographic region where the customer resides.

P.T.O

By considering above information answer the following questions

A. Using the provided dataset, outline the steps you would take to build a Decision Tree model to predict, the 'Campaign Engagement Score' based on Customer Age and other features. [5 marks]

B. How would a Random Forest model handle a given data set differently compared to a single Decision Tree? [5 marks]

C. Evaluate the effectiveness of a Random Forest model in predicting Campaign Engagement Score compared to a Decision Tree model. [5 marks]

D. Propose a strategy to combine Decision Trees and Random Forest models to optimize the marketing campaign? Describe how you would use each model's output to make a final decision on campaign strategies and improvements. [5 marks]

Question 4 [20 marks]

Data imbalance is a fundamental aspect of data science. It affects the accuracy and efficiency of machine learning algorithms. Data imbalance must be properly managed to ensure that models are fair and successful.

A. How can you identify a situation where there is a data imbalance issue? Illustrate your explanation with visual aids such as graphs or charts. [5 Marks]

B. What methods/techniques (Minimum 2 techniques) can be used to handle data imbalance? Hint - Provide the proper technical terms and explain each method. [5 Marks]

C. Mr. ABC suggests that using a **log transformation** is the best way to handle data imbalance. Do you agree with Mr. ABC? First, state whether you agree or disagree with him. Then, explain your reasoning. [5 Marks]

D. Mr. XYZ proposes that **data imputation** can solve the data imbalance problem. Do you agree with Mr. XYZ? First, state whether you agree or disagree with his statement. Then, explain your reasoning. [5 Marks]

Question 5 [20 marks]

Assume that you received marks for a continuous assessment conducted two days ago at OUSL. Initially, you created a graph representing the marks distribution as part of your exploratory data analysis. The generated graph can be found in figure 01.

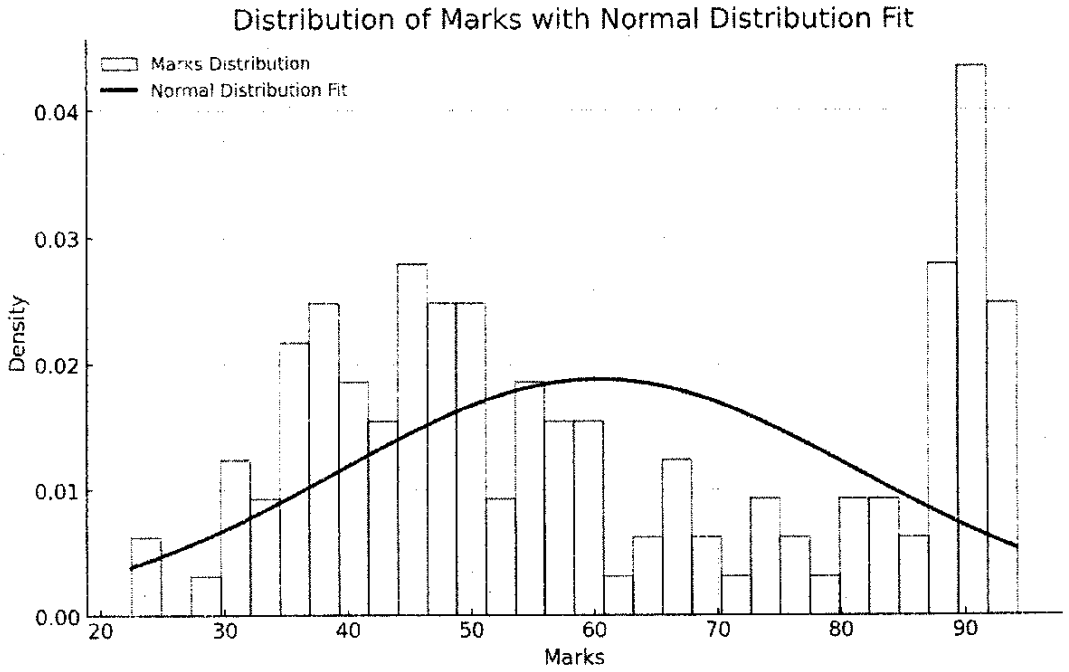

Figure 01 - Histogram with a normal distribution curve

Hint - Mean (μ) is 60.35. Standard Deviation (σ) is 21.24. The lowest mark is 22.5. The highest mark is 94.17.

Based on the above figure and the given statistics, answer the following questions.

A. As a data scientist, If you found outliers in a dataset, what is the approach that you need to follow to clean a data set. Explain clearly with a proper real-world example. Hint – Draw a suitable diagram and explain the theory. Write sample code snippets for a complete answer. [5 Marks]

B. Based on the above graph (histogram with a normal distribution curve), can you find any outliers? Any suspicions in the data set? [5 Marks]

C. Do you think the graph given in figure 1 would be easy to explain the outliers in a data set? If not, what kind of visualization is the best graph to explain outliers. [5 Marks]

D. Provide recommendations for the lecturer based on your analysis to improve the selected subject in future. [5 Marks]