The Open University of Sri Lanka
Faculty of Engineering Technology
Bachelor of Software Engineering Honours

Department	: Electrical and Computer Engineering
Level	: 4
Name of the Examination	: Final Examination
Course Title and - Code	: EEX4373 – Data Science
Academic Year	: 2024/25
Date	: 17/08/2025
Time	: 13.30 hrs – 16.30 hrs
Duration	: 3 hrs

General Instructions

1. Read all the instructions carefully before answering the questions.
2. This question paper consists of **five (5)** questions in **four (4)** pages.
3. Answer all **five (5)** questions. Each question carries equal marks.
4. Answer for each question should commence on a new page.
5. Draw fully labelled diagrams where necessary.
6. Write answers in clear and concise format appropriately in points/bullet form.
7. This is a Closed Book Test (CBT).
8. Involvement in any activity that is considered as an exam offense will lead to punishment.
9. Use blue or black pen to answer the questions.
10. Clearly state your index number in your answer script on every page..
11. When writing code snippets, you can decide the language. E.g. – Python/R

QUESTION 01 [20 Marks]

a) Discuss two methods used for data collection in data science. [2 Marks]

b) Name four commonly used machine learning algorithms. [4 Marks]

c) What is outlier detection? Explain with a suitable example. [4 Marks]

d) Discuss the importance of data visualization in data science. [4 Marks]

e) What are the key steps involved in a data science project? [6 Marks]

QUESTION 02 [20 Marks]

a) Briefly explain what is meant by 'Feature Engineering'. [4 Marks]

b) Briefly define 'feature dimensionality reduction'. [4 Marks]

c) Briefly describe two data quality issues you might encounter in a data science project. [4 Marks]

d) What is the relationship between data science and artificial intelligence? [4 Marks]

e) Discuss the relationship between statistics and data science. [4 Marks]

QUESTION 03 [20 Marks]

You are working with a dataset that contains anonymous employee records. Your task is to perform Exploratory Data Analysis (EDA) to gain a better understanding of the dataset and identify basic patterns.. This exercise does not require any prior domain-specific knowledge and focuses purely on data analysis techniques.

Dataset Columns

- ID: Unique identifier
- Age: Age in years
- Salary: Monthly salary in USD
- Experience: Years of professional experience
- EducationLevel: Highest education level (e.g., Bachelor's, Master's, PhD)
- RemoteWork: Whether the person works remotely (Yes/No)

- Department: Department they belong to (e.g., HR, IT, Sales)

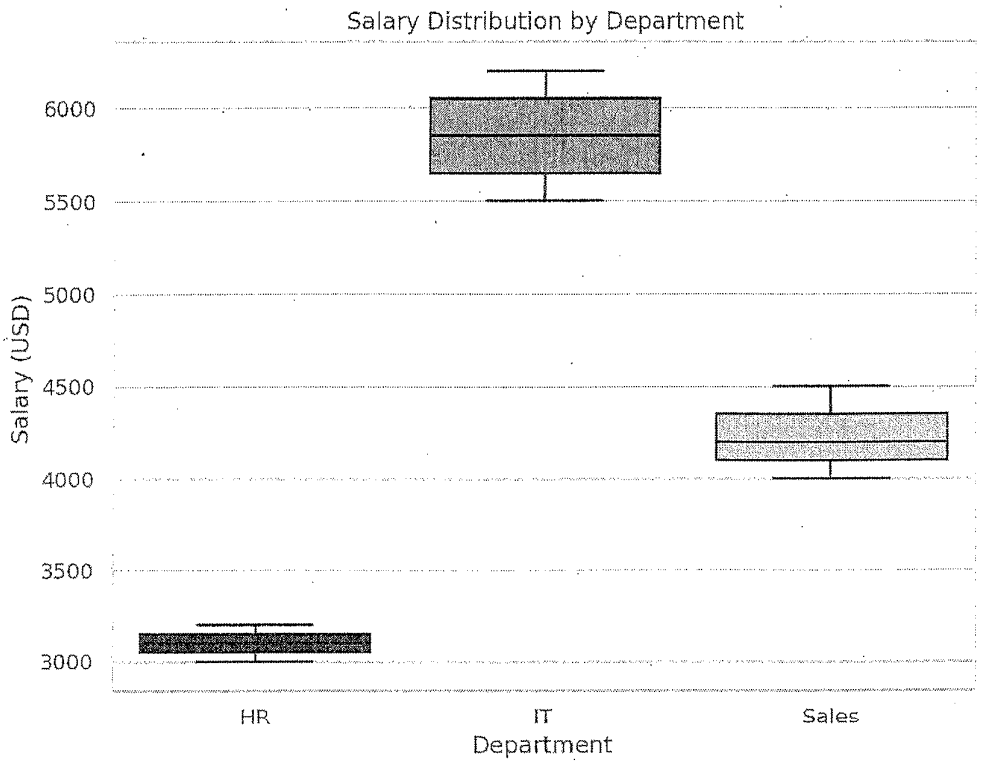

Figure 01 – Salary Distribution by Department

a) Based on Figure 01, which department has the highest median salary? What does this suggest? [4 Marks]

b) What steps would you follow to check for missing values in the dataset? How would you handle missing entries in the 'Salary' column? [4 Marks]

c) What summary statistical features would help you to understand a given column? Name two of the statistical values you may use here. [4 Marks]

d) You need to explore the relationship between 'Experience' and 'Salary.' Explain what kind of techniques you can use. [4 Marks]

e) What type of plot would help you understand how many people work remotely in each department? [4 Marks]

QUESTION 04 [20 Marks]

a) List and explain three unsupervised learning algorithms. [3 Marks]

b) How do unsupervised learning algorithms work? Explain the difference between clustering and dimensionality reduction. [5 Marks]

c) How do supervised learning algorithms handle imbalanced datasets, and what techniques can be used to improve model performance in such cases? [5 Marks]

d) You have been hired as a data scientist by a professional sports club tasked with enhancing decision-making through data-driven insights. One of your core responsibilities is to evaluate and improve the performance of teams, individual players, and athletes over time.

   i. What data science methodologies would you consider most suitable for assessing and measuring performance at each of these levels? [3 Marks]

   ii. How would you tailor these methods to account for sport-specific variables, dynamic in-game conditions, and contextual performance factors? [4 Marks]

QUESTION 05 [20 Marks]

a) Why is Mean Absolute Error (MAE) commonly used for evaluating regression models, and how is it calculated? [4 Marks]

b) What is the purpose of using a validation set when training machine learning models, and how does it help prevent overfitting? [4 Marks]

c) How does cross-validation improve model evaluation compared to a simple train-test split? [4 Marks]

d) When would you choose accuracy as an evaluation metric, and why might it be misleading for imbalanced classification problems? [4 Marks]

e) How would you decide which model performs better: one with lower MAE on training data or one with lower MAE on validation data? Explain your reasoning. [4 Marks]

~~ End of the Paper ~~