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

186

Study Programme	: Bachelor of Software Engineering Honours
Name of the Examination	: Final Examination
Course Code and Title	**: EEX4373 Data Science**
Academic Year	: 2022/23
Date	:04th November 2023
Time	: 0930-1230hrs
Duration	: **3 hours**

General Instructions

1. Read all instructions carefully before answering the questions.
2. This question paper consists of **Five (5) questions** in **five (5)** pages.
3. Answer **ALL** questions. All questions carry equal marks.
4. This is a Closed Book Test **(CBT).**
5. Answers should be in clear hand writing.
6. Do not use Red colour pen.

Question 1

Machine learning automates the process of data analysis. It helps predictions based on data in real time without human intervention. There are three (03) basic types of Machine learning methods namely, supervised, unsupervised and reinforcement learning methods.

(a) Describe the key features in reinforcement learning with an example. [04 marks]

(b) Describe the clustering and association learning methods in unsupervised machine learning methods with an example for each. [06 marks]

(c) Figure 1 contains the process from preparing the data to implementation of a new model. Copy the labels into the answer book and name them and briefly explain the task carried out in each step. [07 marks]

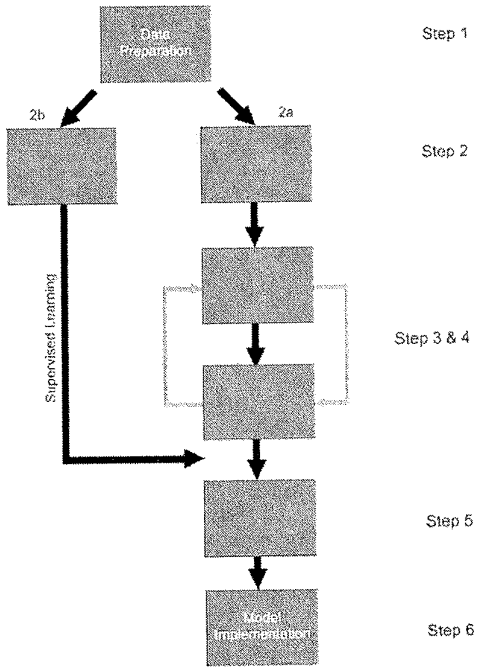

Figure 1- Machine Learning process in Data Science

(d) When developing a machine learning model, the dataset is split into training and testing sets. Explain why this step is important in machine learning. [03 marks]

Question 2

Assume that you are a data scientist working for a retail company, and you've been provided with a dataset containing information about the company's sales and customer behavior. The dataset includes various attributes such as customer demographics, purchase history, and product details etc. Your goal is to perform exploratory data analysis to gain insights from the data.

(a) Describe the process of Exploratory Data Analysis (EDA) that you can carry out for the above-described company data set. [04 marks]

(b) Describe four (04) key statistics, including their key characteristics, that can be used in exploratory data analysis. [04 marks]

(c) Clearly describe three (03) techniques that can be used to handle missing data with advantages and disadvantages of each of the methods. [06 marks]

(b) Assume that you need to create a few charts for each attribute for better understanding. Assuming that you have identified histograms, scatter plots and bar charts as the most suitable types of charts. Compare and contrast each of the selected charts taking ino consideration the attributes from the given scenario. [06 marks]

Question 3

Imagine you are working as a data scientist for a social media company, and your task is to build a machine learning model to classify user posts into two categories: "Spam" and "Not Spam." The company is facing a growing issue of getting more and more spam posts, which are being manually identified at the present. The company wants an automated solution to filter out spam content. You decide to use the k-Nearest Neighbors (KNN) algorithm for this classification task.

(a) Explain the steps of the KNN algorithm in the context of this classification problem. [06 marks]

(b) Describe how you would choose the optimal value of 'k' for the KNN model. [04 marks]

(c) What are the advantages and disadvantages of using KNN for this classification task? [04 marks]

(d) During model development, it is noted that the model is overfitted.

(i) Explain the concept of regularization in the context of model fitting. [02 marks]

(ii) Describe two specific techniques that can be used to mitigate overfitting problem in this scenario. [04 marks]

Question 4

Assume that you are a data scientist working for a large super market chain, and you are tasked with analyzing customer purchase behavior. The marketing team is interested in segmenting customers based on their purchasing habits to create targeted marketing campaigns. You decide to use the k-Nearest Neighbors (KNN) clustering algorithm for this purpose.

(a) Explain the steps on how the KNN clustering algorithm works to segment the customers in the above given context. [04 marks]

(b) Describe the key factors to consider when choosing the optimal value of 'k' for KNN clustering. [02 marks]

(c) Write two (02) advantages and two (02) disadvantages of using KNN clustering for customer segmentation? [04 marks]

(d) Assume that you are given a set of data points to form two clusters using K-means clustering method. The data points are,

(0,0), (0,1), (1,1), (1,0), (o.5,0.5), (5,5), (5,6), (6,6), (6,5), (5.5,5.5)

Choose the first two data points as seed points and initialize the clusters as $C_1=\{(0,0)\}$ and $C_2=\{(0,1)\}$.

(i) Write the data points in each of the clusters C1 and C2 considering the Euclidean distance. [04 marks]

(ii) Write any assumptions you made while assigning data points to clusters. [02 marks]

(iii) Calculate the 2nd stage cluster centers. [02 marks]

(iv) When do you stop clustering? [02 marks]

Question 5

Table 1 shows the details of six students. Write R codes for each of the following.

Table 1: Students' details

ID	Name	Gender	Degree	Mathematics	Statistics	English
A1	Kasun	F	CS	86	78	80
A2	Asanka	M	CS	89	78	76
A3	Nayomi	M	SE	70	80	90
A4	Amali	F	CS	84	92	86
A5	Chandi	M	SE	86	78	80
A6	Bhagya	F	SE	50	68	76

(a) Represent the above data in a data frame. [03 marks]

(b) Name the data frame as ***Marks***. [02 marks]

(c) Add the details of the following student to the above ***Marks*** data frame. [02 marks]

ID: A7 Name: Pubudu

Gender: M Degree: CS

Mathematics: 80 Statistics: 90

English: 85

(d) Add 5 marks for ***English*** for all the students. [02 marks]

(e) Obtain mean, range and summary statistics for the module ***Mathematics***. [03 marks]

(f) Obtain the details of the students who are following CS degree. [02 marks]

(g) Obtain the average mark of each student for the three subjects and add to the data frame as a new variable with the name ***AverageMark***. [03 marks]

(h) Obtain the names of the female students whose average mark is above 70. [03 marks]

End of the Paper