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

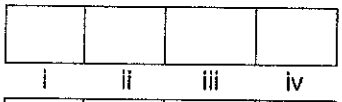
Study Programme	: Bachelor of Software Engineering Honours
Name of the Examination	: Final Examination
Course Code and Title	**: EEX4373 Data Science**
Academic Year	: 2020/21
Date	: 08th January 2022
Time	: 0930-1230hrs
Duration	: **3 hours**

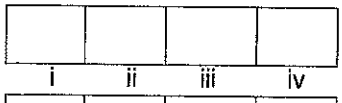
General Instructions

1. Read all instructions carefully before answering the questions.
2. Answer **all** questions in PART A.
3. Questions in PART A carry equal marks.
4. Answer for PART A should mark in this sheet.
5. Attached this answer sheet with PART B answer book

ANSWER SHEET FOR SECTION A

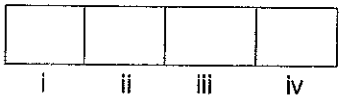
INDEX NO: | | | | | | |

1) 

2) 

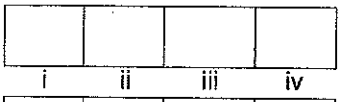
3) 

4) 

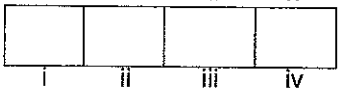
5) 

6) 

7) 

8) 

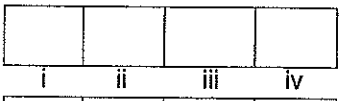
9) 

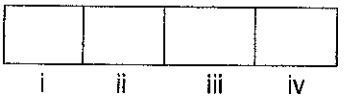
10) 

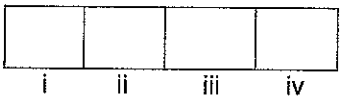
11) 

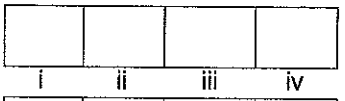
12) 

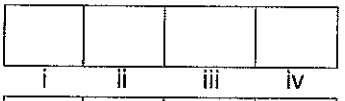
13) 

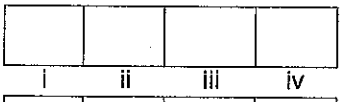
14) 

15) 

16) 

17) 

18) 

19) 

20) 

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

Study Programme	: Bachelor of Software Engineering Honours
Name of the Examination	: Final Examination
Course Code and Title	**: EEX4373 Data Science**
Academic Year	: 2020/21
Date	: 08th January 2022
Time	: 0930-1230hrs
Duration	: **3 hours**

General Instructions

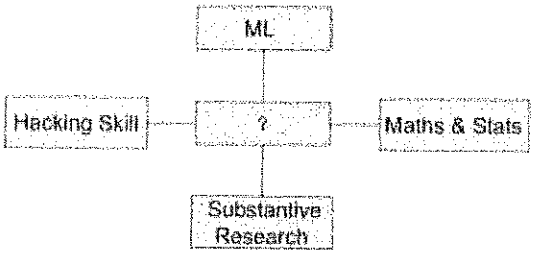
1. Read all instructions carefully before answering the questions.
2. This question paper contains two parts, **Part A** and **PartB. Part A** answers should be marked in the answer sheet provided. **Part B** answers should be written in the answer book provided. The answer for each question should commence from a new page.
3. This question paper consists of **Twenty (20)** MCQ questions in Part A and **Five (5)** essay questions in Part B in **Seven (7)** pages.
4. Answer all questions in part A and any **Four(4)** questions from **Part B**. All questions in part B carry equal marks.
5. This is a Closed Book Test**(CBT).**
6. Answers should be in clear handwriting.
7. Do not use red colour pen.

PART – A

Multiple Choice Questions [40 marks]

1. Which of the following would be more appropriate to be replaced with the question mark in the following figure?

(2 marks)



a. Data Analysis
b. Data Science
c. Descriptive Analysis
d. Artificial Intelligence

2. The most suitable term to describe the data that is huge and complex to store and process is:

(2 marks)

a. Data Science
b. Data Mining
c. Big Data
d. Data Warehouse

3. Which of the following step is performed by a data scientist after acquiring the data?

(2 marks)

a. Data Cleaning
b. Data Integration
c. Data Replication
d. All of the above

4. Which of the following statement is False in the case of the KNN Algorithm?

(2 marks)

a. For a very small value of K, the algorithm is very sensitive to noise.
b. For a very large value of K, points from other classes may be included in the neighborhood.
c. KNN is a lazy learner.
d. KNN is used only for classification problem statements.

5. The robotic arm will be able to paint every corner in the automotive parts while minimizing the quantity of paint wasted in the process. Which learning technique is used in this problem?

(2 marks)

a. Supervised Learning.
b. Reinforcement Learning.
c. Unsupervised Learning.
d. Both (A) and (B).

6. Which one of the following statements is TRUE for a Decision Tree?

(2 marks)

a. The decision tree is only suitable for the classification problem statement.
b. In a decision tree, the entropy of a node decreases as we go down a decision tree.
c. In a decision tree, entropy determines purity.
d. The decision tree can only be used for only numeric valued and continuous attributes.

7. How do you choose the right node while constructing a decision tree?

(2 marks)

a. An attribute having high entropy
b. An attribute having high entropy and information gain
c. An attribute having the lowest information gain.
d. An attribute having the highest information gain.

8. What kind of distance metric(s) are suitable for categorical variables to find the closest neighbors?

(2 marks)

a. Hamming distance.
b. Euclidean distance.
c. Manhattan distance.
d. Minkowski distance.

9. In the Naive Bayes algorithm, suppose that prior for class w1 is greater than class w2, would the decision boundary shift towards the region R1(region for deciding w1) or towards region R2(region for deciding w2)?

(2 marks)

a. Towards region R1.
b. Towards region R2.
c. No shift in decision boundary.
d. It depends on the exact value of priors.

10. Which of the following is FALSE about Correlation and Covariance? (2 marks)

a. A zero correlation does not necessarily imply independence between variables.
b. The covariance and correlation are always the same sign.
c. Correlation and covariance values are the same.
d. Correlation is the standardized version of Covariance.

11. Which of the following is FALSE for neural networks?

(2 marks)

a. Artificial neurons are similar in operation to biological neurons.
b. Training time for a neural network depends on network size.
c. Neural networks can be simulated on conventional computers.
d. The basic unit of neural networks are neurons.

12. Which of the following gave rise to the need for graphs in data analysis?

(2 marks)

a. Data visualization
b. Communicating results
c. Decision making
d. All of the above

13. ———— is/are an example of human-generated unstructured data.

(2 marks)

a. YouTube data
b. Satellite data
c. Sensor data
d. All of the above

14. JSON file data is an example of ————

(2 marks)

a. Structured data
b. Un-Structured data
c. Semi-Structured data
d. Scattered data

15. What function/s do we use to import a CSV file into a R session?

(2 marks)

i. read.table()
ii. read.csv()
iii. scan()

a. Only (i) is true
b. Only (ii) is true
c. Only (i) and (ii) are true
d. All the statements are true

16. What is not an instance of the R recycling rule?

(2 marks)

a. seq(0,6,2)
b. c(1,2,4) + c(6,0,9,20,22)
c. cbind(1:6,1:3)
d. c(11, 2) < c(10, 20, 30)

17. R will represent Dividing by zero by the symbol,
    a. (2 marks)
    b. NA
    c. NaN
    d. Null
    e. 0

18. Which of the following businesses would be least likely to use a Recommendation System?

(2 marks)

a. Facebook
b. WhatsApp
c. Instagram
d. LinkedIn

19. What does the characteristics "Velocity" in Big Data represents?

(2 marks)

a. Speed of input data generation
b. Speed of individual machine processors
c. Speed of only storing data
d. Speed of storing and processing data

20. What statement/s correctly describe R Lists, Matrices, and Data frames

(2 marks)

i. The list is a one-dimensional data structure, where Matrices and Data frames are two-dimensional.
ii. The list can consist of elements of the same data types, where Matrices and Data frames can contain objects of multiple data types.
iii. It is possible to perform indexing and subletting techniques for all the above three data structures.

a. Only (iii) is true
b. Only (iii) and (ii) are true
c. Only (iii) and (i) are true
d. All the statements are true

PART – B

(Essay Questions) Answer ANY 4 Questions (60 Marks)

1. Write short answers for the following questions.
   a. Briefly explain why Data Science is regarded as a Multidisciplinary subject. (2 Marks)
   b. Describe three (03) reasons why data cleansing is important in data analysis. (3 Marks)
   c. Briefly describe the role of Machine Learning in Data Science? (5 Marks)
   d. Describe 03 applications of Machine Learning in Data Science? (5 Marks)

2. Decision Trees are one of the popular algorithms which can be used for Classification. It usually mimics human thinking ability while making a decision.
   a. Briefly explain the procedure of creating a decision tree in 5 steps. (5 Marks)
   b. Draw a simple decision tree for enrolling in a MSc degree programme after your graduation? (3 Marks)
   c. Name 2 other Classification algorithms except Decision Tree algorithm? (2 Marks)
   d. Explain the difference between classification and regression in machine learning? (5 Marks)

3. Structured data is clearly defined and searchable types of data, while unstructured data is usually stored in its native format.
   a. Give 2 examples each for structured data and unstructured data? (2 Marks)
   b. Describe whether it is possible to use R object type Matrix to store unstructured data? Justify your answer with reasons? (3 Marks)
   c. What is the difference between array and matrix types in R. Write a simple code block to explain the same? (5 Marks)
   d. Briefly explain the 5Vs in Big Data? (5 Marks)

4. Supervised Learning is a subcategory of Machine Learning in Artificial Intelligence.
   a. What are the 2 types of Supervised Learning? (2 Marks)
   b. Name 3 unsupervised learning algorithms. (3 Marks)
   c. Summarize the importance of visualization of data in exploration data analysis. (5 Marks)
   d. Describe what is clustering in Machine Learning briefly explaining with a suitable example? (5 Marks)

5. This question is based on your knowledge in R scripting language.
   a. Which data object in R is used to store and process categorical data? Explain why? (2 Marks)
   b. Below table represents the marks of 5 students for 3 subjects. Refer to the table and write R Script to (6 Marks)
      i. Store these data in a Matrix?
      ii. Find the average mark of Java subject?
      iii. Find the maximum mark of C language?

Student Name	Java	Data Science	C language
John	90	67	51
Harry	95	64	59
Niko	92	60	67
Tom	80	52	40
Jennifer	93	70	77

   c. Write a R function to check if the given number is odd or even. Then it should print the following. (7 Marks)

      Eg: 7 is an odd number
      10 is an even number

~End of the paper~