Tribhuvan University
Institute of Sciences and Technology

SCHOOL OF MATHEMATICAL SCIENCES

First Assessment 2080

Subject: **Fundamental of Data Science**
Course No: **MDS 501**
Level: **MDS /I Year /I Semester**

Full Marks: **45**
Pass Marks: **22.5**
Time: **2hrs**

Candidates are required to give their answer in their own words as far as practicable.

Attempt All questions.

Group A [5 × 3 = 15]

1. Explain your understanding of data science as multidisciplinary field.
2. What do you mean by data standardization? List out the techniques used for data standardization.
3. How machine learning is different from traditional programming? What actually machine learns through machine learning? Justify your answer.
4. What do you mean by biasness and fairness in data science? Why are they necessary to address in data science?
5. What do you mean by deep learning? How is it similar/different to Neural Network?

Group A [5 × 6 = 30]

6. Explain the CRISP-DM lifecycle for data mining process.

OR

Explain the TDSP lifecycle for data science.

7. What do you mean by missing data? How are they handled in data science process?
8. Fit the linear regression in the given data

X	Y
10	24.06
20	43.91
30	64.1

And predict the output if X = 40.

9. Explain neural network along with forward propagation and back propagation.

OR

Differentiate between Regression and Classification in supervised learning.

10. Explain the different biases that are likely to occur during data science lifecycle.

***

CamScanner

Tribhuvan University
Institute of Sciences and Technology

SCHOOL OF MATHEMATICAL SCIENCES

First Assessment 2080

Subject: **Data Structure and Algorithms** *Full Marks*: **45**
Course: **MDS 202** *PassMarks*:**22.5**
Level: **MDS/I Year/I Semester** *Time*: **2 hrs**

Candidates are required to give their answer in their own words as far as practicable.

Attempt ALL Questions.

Group A [5 x 3 = 15]

1. What is data structure? Explain dynamic memory allocation in brief. (1 + 2)
2. Convert ((A-B)$C - (D - E))$(F+G) to prefix and postfix. (1.5 + 1.5)
3. What is priority queue? Explain.
4. Explain recursive algorithm with example. What is iteration? (2 + 1)
5. Compare linked list with array. What is circular linked list? (2 + 1)

Group A [5 x 6 = 30]

6. Define stack. How do you implement push and pop operations in Stack? Explain. (1 + 5)

OR

How do you implement stack using linked list? Explain. (6)

7. Explain algorithm to convert an infix expression to postfix. Use this algorithm to convert the infix expression (A + B) * C – D to postfix.(3 + 3)

OR

Define queue. How do you implement queue operations in array data structure? Explain. (1 + 5)

8. Define time complexity. What is asymptotic notation? Explain big-oh, omega, and theta notations. (1 + 2 + 3)
9. How do you insert and remove nodes in singly linked list? Explain. (6)
10. Define tail recursion. Explain tail recursion using suitable program. (1.5 + 4.5)

***

CamScanner

Tribhuvan University
Institute of Sciences and Technology

SCHOOL OF MATHEMATICAL SCIENCES

First Assessment 2080

Subject: **Statistical Computing with R**
Course No: **MDS 503**
Level: **MDS /I Year /I Semester**

Full Marks: **45**
Pass Marks: **22.5**
Time: **2hrs**

*Candidates are required to write answers with examples for answering question numbers **1-5**in answer sheets and use laptop for answering question numbers **6-10**. R scripts and outputs/interpretation of question number 6-10 must kintted as HTML file by saving it in a folder with name/exam roll number and submitted for grading.*

Attempt ALL questions.

Group A [5 × 3 = 15]

1. Explain how to import these types of data in R using base R functions:
   a) Comma separated values text file
   b) Excel data file
   c) SPSS data file

2. Explain following data types in R with examples and R codes:
   a) Numeric and integer
   b) Categorical and factor
   c) Data and Date as well as time

3. Explain these terms with examples for R:
   a) Arrays and matrices
   b) List and unlist
   c) Data frame and data table

4. Explain the followings with examples for R:
   a) Base R code vs pipe operator code
   b) For loop code vs pipe operator code
   c) "tee" pipe operator vs "exposition" pipe operator

5. Explain the following in R with example:
   a) Package
   b) Installation of package
   c) Development of package

Group B [5 × 6 = 30]

6. Open the R studio and do the followings with R script and knit HTML output:
   a) Define integers from 1 to 15 using three different coding approaches as I
   b) Define these five numbers: 1.1, 2.2, 3.3, 4.4 and 5.5 and save it as N
   c) Add, subtract, multiply and divide I from N and interpret the results carefully
   d) Perform matrix multiplication of I and N and interpret the result carefully
   e) Transpose I and N, perform matrix multiplication and interpret the result carefully

CamScanner

7. Do the following in R studio and with R script to knit HTML output:
   a) Load the in-built "air quality" data available in base R as AQ object
   b) Check the structure of AQ and explain class of each variable
   c) Replace missing values of Ozone variable with median of this variable
   d) Replace missing values of Solar. R variable with mean of this variable
   e) Create "Date" variable on AQ object using "Month" and "Day" variables for year 2020
8. Use the cleaned "AQ" object to do following in R Studio with R script to knit HTML output:
   a) Create line plot of "Temp" with "Day" as the row index and interpret it carefully
   b) Create bar plot of "Temp" variable after defining class intervals systematically
   c) Create histogram of "Temp" variable and compare it with the bar plot of "Temp" variable
   d) Plot Normal Q-Q plot of "Temp" variable and interpret it carefully
   e) Create a scatter plot of "Temp" and "Wind" variables and interpret it carefully
9. Do the following in R Studio with tidy verse package using R Script to knit HTML output:
   a) Define a tibblehaving country, year, cases and population variables with 10 random data each
   b) Transform this tibble to long format and interpret it carefully in terms of tidy data format
   c) Transform the cases variable as log of cases (LnCase) and population variable as log of population (LnPop)
   d) Create scatter plots of 1. Cases and population, 2. LnCase and population, 3. Cases and LnPop and 4.LnCase and LnPopin a single graph window and interpret it carefully

OR

Use the cleaned "AQ" file in R studio and do as follows with R Scripts and HTML outputs:
   a) Get reference range of "Temp"variable using mean and standard deviation
   b) Plot histogram of "Temp" variable and show the outliers of "Temp" with reference range limits
   c) Get reference range of "Temp"variable using median and inter-quartile range
   d) Plot box plot of "Temp" variable and show the outliers of "Temp" with reference range limits
   e) Which measure of central tendency and dispersion should be used for this variable? Why?
10. Load the "igraph" package in R studio and do the basic SNA as follows with R script and HTML output:
   a) Define g as graph object with (1,2) as its elements
   b) Plot the g and interpret it carefully
   c) Define g1 as graph object with ("R", "S", "S", "T", "T", "R", "R", "T", "U", "S") as its elements
   d) Plot g1 with node color as green, node size as 30, link color as red and link size as 5 and interpret it
   e) Get degree, closeness and betweenness of g1 and interpret them carefully.

OR

Do as follows in R Studio and do as follows with R script and HTML outputs:
   a) Open R and then go to Help and Manuals if PDF and open "An Introduction to R" file
   b) Import this pdf file in R using "pdftools" package
   c) Perform pre-processing and create 'corpus' afterwards
   d) Find the most frequent terms and create histogram of the most frequent
   e) Create word cloud of the corpus, color it using rainbow or RColorBrewer package
   f) Perform topic modelling and interpret the result carefully

Note: Save the R scripts and knitted HTML files of Group B questions with your name/roll number!

***

CamScanner

Tribhuvan University
Institute of Sciences and Technology

SCHOOL OF MATHEMATICAL SCIENCES

First Re-assessment 2080

Subject: **Statistical Computing with R**
Course No: **MDS 503**
Level: **Masters in Data Science /I Year /I Semester**

Full Marks: 45
Pass Marks: 22.5
Time: 2hrs

Candidates are required to write answers with examples for answering question numbers ***1-5****in answer sheets and use laptop for answering question numbers* ***6-10****. R scripts and outputs interpretation of question number 6-10 must be knitted as HTML file. Save the HTML file in a folder with name exam roll number and submit them for grading.*

Attempt ALL questions.

Group A [5 × 3 = 15]

1. Explain how to import these types of data in R using "**dplyr**" package:
   a) Tab separated values text file
   b) Comma separated values text file
   c) SPSS data file

2. Explain following data types in R with examples:
   a) Integer variable is different than number variable
   b) Categorical variable is different than factor variable
   c) Date variable is different than Date as well as time variable

3. Explain these terms with examples for R:
   a) Getting multi-way table with array
   b) Creating class intervals of continuous variable
   c) Missingness vs nothingness

4. Explain the followings with examples for R:
   a) Reference range based on mean
   b) Reference range based on median
   c) Outliers and extreme values

5. Explain the following in R with example:
   a) Nodes and edges
   b) Diameter
   c) Edge density

Group B [5 × 6 = 30]

6. Open the R studio and do the followings with R script and knit HTML output:
   a) What happens when 4L is multiplied by 3.2?
   b) What happens when 4L is multiplied by 2L?
   c) Define blood with O, O, A, A, B, B and check its type and attributes with your comments
   d) Define x with 1,2,NA,8,3,NA,3 and get its mean with or without pipes.
   e) Get the first and sixth elements of x using sub-setting codes and its explanation.

7. Do the following in R studio and with R script to knit HTML output:
   a) Define an object "rating" with 9, 2, 5, 8, 6, 1, 3, 2, 8, 4, 6, 8, 7, 1, 2, 6, 10, 5, 6, 9, 6, 2, 4, 7
   b) Replicate the given table obtained from SPSS software for the rating object in R

rating

CamScanner

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	2	8.3	8.3	8.3
	2	4	16.7	16.7	25.0
	3	1	4.2	4.2	29.2
	4	2	8.3	8.3	37.5
	5	2	8.3	8.3	45.8
	6	5	20.8	20.8	66.7
	7	2	8.3	8.3	75.0
	8	3	12.5	12.5	87.5
	9	2	8.3	8.3	95.8
	10	1	4.2	4.2	100.0
	Total	24	100.0	100.0	

8. Use the "air quality" data as AQ to do following in R Studio with R script to knit HTML output:
   a) Replace missing values of Ozone variable with the best measure of central tendency
   b) Create a Date variable in AQ using Month and Day variable for year 2022.
   c) Create line plot of "Ozone" variable with "Date" as the row index and interpret it carefully
   d) Get class intervals of the cleaned Ozone variable using range, its square root and zero rounding
   e) Get frequency distribution (n and %) of Ozone variable class intervals and interpret it carefully
9. Do the following in R Studio with **tidyverse** package using R Script to knit HTML output:
   a) Define a tibblehaving country, year, cases and population variables with 100 random data each
   b) Transform the cases variable as log of cases (LnCase) and population variable as log of population (LnPop)
   c) Create scatterplots of 1. Cases and population, 2. LnCase and population, 3. Cases and LnPop and 4.LnCase and LnPop in a single graph window **with base R plot code** and interpret it carefully.

OR

Load the "igraph" pac
kage in R studio and do the basic SNA as follows with R script and HTML output:

   a) Define g as graph object with (1,2,2,3,3,4,4,1) as its elements
   b) Plot g with node color as green, node size as 30, link color as red and link size as 5 and interpret it
   c) Plot the g as undirected arguments and interpret it carefully
   d) Plot g with seven nodes and interpret it carefully
   e) Get degree, closeness and betweenness of g and interpret them carefully.

10. Use the cleaned "AQ" file in R studio and do as follows with R Scripts and HTML outputs:
   a) Get reference range of "Ozone" variable using mean and standard deviation
   b) Plot histogram of "Ozone" variable and show the outliers of "Ozone" with reference range limits
   c) Get reference range of "Ozone" variable using median and inter-quartile range
   d) Plot boxplot of "Ozone" variable and show the outliers of "Ozone" with reference range limits
   e) Write a summary of the results obtained from the histogram and boxplot

OR

Do as follows in R Studio and do as follows with R script and HTML outputs:

   a) Open R and then go to Help and Manuals if PDF and open "An Introduction to R" file
   b) Import this pdf file in R using "pdftools" package
   c) Perform pre-processing and create 'corpus' afterwards
   d) Find the most frequent terms and create histogram of the most frequent
   e) Create word cloud of the corpus, color it using rainbow or R Color Brewer package
   f) Perform topic modelling and interpret the result carefully

Note: Save the knitted HTML files of Group B questions with your name/roll number!

***

CamScanner

Tribhuvan University
Institute of Sciences and Technology

SCHOOL OF MATHEMATICAL SCIENCES

First Assessment 2080

Subject: **Mathematics for Data Sciences** *Full Marks*: 45
Course No: **MDS 504** *Pass Marks*: 22.5
Level: **MDS /I Year /I Semester** *Time*: 2hrs

Candidates are required to give their answer in their own words as far as practicable.

Attempt All questions.

Group A [5 × 3 = 15]

1. Are the following sets form the subspace of R^2? Justify.

a) The set S of all solutions of homogeneous equation $AX = 0$ of any 2×2 matrix A and $X=\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in R^2$.

b) The closed L_2 ball $= B(0,1)=\{X=\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in R^2 : \|X\| \leq 1\}$. [1.5+1.5]

2. Define an involutory matrix. Is the matrix $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ involutory? Justify. Prove that the inverse of the transpose of a non–singular matrix A is the transpose of its inverse. [0.5+1+1.5]

3. Define linear transformation. Let $T: R^2 \to R^2$ be defined by $T(0, 1) = (2, 1)$, $T(1, 4) = (0, -2)$. If T is linear, find the formula for $T(x, y)$. Also, find the matrix represented by T relative to the standard basis. [0.5+1.5+1]

4. What do you mean by a symmetric bilinear form on a vector space V over the field F? Let A be a symmetric matrix. Prove or disprove that the mapping $B_A(u,v) = u^T Av$, $\forall u,v \in V$ is a symmetric bilinear form on V. [1+2]

5. Let $v_1, v_2, \ldots, v_n$ be the eigenvectors associated with the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of a $n \times n$ symmetric matrix A respectively, then prove that

$$A = \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T + \ldots + \lambda_n v_n v_n^T$$

Group B [5 ×6 = 30]

6. Define L_1, L_2 and L_∞ norms on a vector n-space R^n. Let $\|.\|$ be the Euclidean norm ,and X & Y be two vectors in R^n. Prove the Cauchy -Schwarz inequality $|X.Y| \leq \|X\| \|Y\|$. Verify this property for $X = (1, 3)$ and $Y = (2, 1)$. [1+4+1]

7. Distinguish between linear dependent and independent vectors. Prove that the linear hull of a given set of vectors $v_1, v_2, \ldots, v_n$ in a vector space V is a subspace of V. Also, the representation of any vector in a vector space in terms of its basis vectors is unique. [1+2.5+2.5]

OR

CamScanner

Define a basis and dimension of a vector space. Show that the set B= {(1, 1, 1), (1, −1, 1), (2, 0, 3)} forms a basis for R^3. Also, find the co–ordinates of $v = (1, 3, 2) \in R^3$ with respect to the basis B. [1+3+2]

8. Define fourier coefficient of a vector u on a vector v. Prove that a set of non–zero orthogonal vectors is linearly independent. Also, find an orthonormal basis from the basis {(1, 0, 1),(1, 1, 0), (1, 1, 1)} of R^3 using Gram Schimdt Orthogonalization Process. [1+1.5+3.5]

9. What are the conditions necessary for a matrix to possess an inverse? Prove that the inverse of a square matrix if it exists, is unique. If A, B ,and C are matrices of order $m \times n$, $n \times p$, $p \times q$ respectively, then prove that $(AB)C = A(BC)$.. [1+1+4]

OR

Define quadratic form. Let A be a 3×3 square matrix with a quadratic form in 3 variables. Then there exists a 3×3 symmetric matrix B such that $X^TAX = X^TBX$, $\forall X \in R^3$. Further, express the quadratic form $x_1^2 + x_1x_2 - 4x_3x_1 + 2x_2x_3 - 4x_3^2$ as the difference of squares. [1+2.5+2.5]

10. Find the eigenvalues and the corresponding eigenvectors of the matrix $A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$.

a) Diagonalize the matrix A.

b) Find the quadratic form determined by A and test its definiteness.

c) Remove the cross term of the quadratic form.

d) Examine the maximum and minimum value of quadratic form subject to the constraint $\| X^TX \| = 1$. [2+1.5+1+1.5]

***

CamScanner

Tribhuvan University
Institute of Sciences and Technology

SCHOOL OF MATHEMATICAL SCIENCES

First Assessment 2080

Subject: **Data Base Management Systems** | *Full Marks*: **45**
Course No: **MDS 505** | *Pass Marks*: **22.5**
Level: **MDS /I Year /I Semester** | *Time*: **2hrs**

Candidates are required to give their answer in their own words as far as practicable.

Attempt All questions.

Group A [5 × 3 = 15]

1. What is database? List the characteristics of database system. [2+1]
2. What is data independence? How three schema architecture ensures logical and physical data independence? [1+2]
3. Create a relation of your choice and illustrate the concepts of attribute, tuple, and primary key. [3]
4. Why aliasing in SQL is needed. Show aliasing of relation in SQL query. [1+2]
5. What is trigger? Illustrate with example, how can you define before trigger on insert operation. [1+2]

Group B [5 × 6 = 30]

6. What do you mean by normalization? Discuss 2NF and 3NF with examples. [2+4]

OR

What is functional dependency? Discuss 1 NF and 2NF with examples. [2+4]

7. What is assertion? Mention its use. Given following relations, create an assertion to ensure that there is a person who is student. [1+1+4]

Person(*pid, pname, dob, padd*)
Student(*roll, pid.*)

OR

How assertion differ from trigger? Given following relation, create an assertion to ensure that there is no person with name starting with "An" and dob = 2020. [2+4]

Person(*pid, pname, dob, padd*)

8. Design an ER diagram for following scenario: [6]

In film industry, producers produce movies. Producers have their name, age and budget as attributes. They are uniquely identified by prod_id. All the movies have their title, year, and release date. No movies can have same title. Every movies must be played by actor. An actor can play many movies. Actors have Fname and Lname to uniquely identify them. The actors have charge_rate as well. A single movie can have many producers and a producer can produce zero or many movies. There is an identifying relationship between actor and vanity van. Vanity van has partial attribute van_id.

CamScanner

9. Consider the following relations containing airline flight information. [6]

Flights(*flno*, *from*, *to*, *flight_distance*, *departs_at*, *arrives_at*, *aid*)
Aircraft(*aid*, *aname*, *cruising_range*)
Certified(*pid*, *aid*, *date*, *certified_by*)
Pilot(*pid*, *pname*, *psalary*)

Write the **SQL** statements for following;

a) Find all aircrafts.
b) Find the name of flights that arrive at 1:00 PM.
c) Find the names of certified pilots
d) Find names of aircrafts that can be used on flights from KTM to DHI.
e) Find the average salary of pilots.
f) Use left outer join between Aircraft and Certified.

10. Consider the following relations containing airline flight information. [6]

Flights(*flno*, *from*, *to*, *flight_distance*, *departs_at*, *arrives_at*, *aid*)
Aircraft(*aid*, *aname*, *cruising_range*)
Certified(*pid*, *aid*, *date*, *certified_by*)
Pilot(*pid*, *pname*, *psalary*)

Write the **Relational Algebra** statements for following;

a) Find all aircrafts.
b) Find the name of flights that arrive at 1:00 PM.
c) Find the names of certified pilots
d) Find names of aircrafts that can be used on flights from KTM to DHI.
e) Find total number of pilots.
f) Use natural join between Aircraft and Certified.

***

CamScanner