Tribhuvan University
Institute of Sciences and Technology
SCHOOL OF MATHEMATICAL SCIENCES
First Assessment 2078

Subject: **Fundamentals of Data Science** *Full Marks:* **45**

Course No: **MDS 501** ***Pass Marks:*** **18**

***Level*: MDS First Year First 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 × 3 = 15]

1. Describe the applications and limitations of Data Science.
2. What is data science lifecycle? Briefly explain two major data science life cycles used by industries.
3. List and highlight the differences between structured, unstructured, and semi-structured data with examples of each.
4. Briefly explain the various methods used to handle missing values during data cleanup.
5. You want to identify global weather patterns that may have been affected by climate change. To do so, you want to use machine learning algorithms to find patterns that would otherwise be imperceptible to a human meteorologist. Discuss what machine learning method (supervised, unsupervised, reinforcement) would you use and why.

Group B [5 × 6 = 30]

6. Explain CRISP-DM, its advantages, and the steps involved in it.

OR

Explain TDSP lifecycle and the steps involved. Mention at least two main advantages of using TDSP lifecycle.

7. With an example, explain how you would determine the True Negative and False Negative data from researchdataset.
8. Explain with examples observation bias and funding bias in a research survey.
9. Describe, at a high-level, the major steps that need to be taken for data cleanup/munging.
10. Describe and explain with examples the various types of machine learning methods (Supervised, Unsupervised, and Reinforcement).

OR

Explain with examples and highlight the relationship between Artificial Intelligence and Machine Learning

***

Scanned with CamScanner

Tribhuvan University
Institute of Science and Technology
SCHOOL OF MATHEMATICAL SCIENCE
First ReAssessment 2078

Subject: Fundamentals of Data Science
Course No: MDS501
Level: MDS /1 year/1 Semester

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

Group A [5 × 3 = 15]

1. List three major limitations of Data Science.
2. Discuss in brief three ethical issues in Data Science.
3. List and highlight the differences between structured and unstructured data with examples.
4. Briefly explain the various methods used to handle missing values during data cleanup.
5. You want to identify global weather patterns that may have been affected by climate change. To do so, you want to use machine learning algorithms to find patterns that would otherwise be impereceptible to a human meteorologist.
   Discuss what machine learning method (supervised, unsupervised, reinforcement) would you use and why.

Group B [5 × 6 = 30]

6. Explain OSEMN framework, its advantages, and the steps involved in it.

   OR

   Explain TDSP lifecycle and the steps involved. Mention at least two main advantages of using TDSP lifecycle.
7. With an example, explain how you would determine the True Negative and False Negative data from researchdataset.
8. Explain with examples negativity bias and bias blind spot in a research survey.
9. Describe, at a high-level, the major steps that need to be taken for data cleanup/munging.
10. Describe and explain with examples the various types of machine learning methods (Supervised, Unsupervised and Reinforcement).

    OR

    How is Artificial Intelligence and Machine Learning related? Explain with examples their inter-connectivity.

Scanned with CamScanner

Second Assessment

Fundamentals to Data Science

MDS 501

Group A (5 questions x 3 marks = 15 marks)

1. What are the differences between linear regression and logistic regression? Explain with an example.
2. What kind of problem can a Decision Tree solve? Explain with an example.
3. Define Support Vector Machines. Describe briefly how SVMs are used for classification.
4. Define and list out major differences between Data Warehouse and Data Lake.
5. What is Hadoop? List and briefly discuss the major parts of a Hadoop system.

Group B (5 questions x 6 marks = 30 marks)

1. Please answer any <u>ONE</u> of the following
   a. What are the advantages of using a random forest algorithm?
   OR
   b. Explain with a real life example where k-NN algorithm can be used to solve problem.
2. List two main issues with privacy and data ethics with examples.
3. Explain how Demographics Parity and Equal Opportunity can help address biases.
4. What is big data? Explain the five Vs of big data.
5. Please answer any <u>ONE</u> of the following
   a. Explain what is Naïve Bayes. Describe its common use cases with examples.
   OR
   b. Explain Decision Trees. Describe its common use cases with examples.

Scanned with CamScanner

TRIBHUVAN UNIVERSITY

INSTITUTE OF SCIENCE AND TECHNOLOGY

SCHOOL OF MATHEMATICAL SCIENCES

First Assessment 2078

Subject: **Mathematics for Data Science** *Full Marks*: 45
Course No.: MDS 504 *Pass Marks*: 22.5
Level: **Master in Data Science/I** Semester *Time*: 2:00 hr

Attempt ***ALL questions****. Write your answer in detail as far as possible.*

Group A [$3 \times 5 = 15$]

1. Show that

   (a) The line $x_2 = ax_1$ (in usual notations, $y = ax$) is a subspace $\mathbb{R}^2$.

   (b) The line $x_2 = ax_1 + b$ (perhaps more familiar as $y = ax + b$) is not a subspace $\mathbb{R}^2$ for $b \neq 0$.

2. Show that any vector in $\mathbb{R}^3$ can be expressed as a linear combination of the three unit basis vectors in $\mathbb{R}^3$. Also, show that a linear combination of the three unit basis vectors in $\mathbb{R}^3$ equals to 0 if and only if all coefficients in the linear combination are zeros.

3. What is the parallel coordinates method? Explain with explain with an example. What is the use of this method in data science?

4. Find a basis for the solution space of the equation $x + y - z = 0$.

5. Let $u_1 = (1, 2, 2, -1), u_2 = (1, 1, -1, 1), u_3 = (-1, 1, -1, -1)$ and $B = \{u_1, u_2, u_3\}$ an orthogonal basis for $V = \text{span}(u_1, u_2, u_3)$. Find the projection of $w = (0, 1, 2, 3)$ onto V.

Group B [$6 \times 5 = 30$]

6. (a) By showing that the L_∞-norm satisfies each of the conditions in the definition of a norm prove this is a vector norm for $\mathbb{R}^n$.

   (b) Let $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ be a vector with $x_i = i^{-1}$. Compute the 1-norm, the 2-norm, and the ∞-norm of x.

OR

Prove that if x and y are vectors in $\mathbb{R}^n$, then

(a) $|x \cdot y| \leq \|x\|_2 \|y\|_2$.

Scanned with CamScanner

TRIBHUVAN UNIVERSITY

INSTITUTE OF SCIENCE AND TECHNOLOGY

SCHOOL OF MATHEMATICAL SCIENCES

ॐ नमः शिवाय

Reassessment I 2078

Subject: **Mathematics for Data Science**
Course No.: MDS 504
Level: **Master in Data Science/**I Semester

Full Marks: 45
Pass Marks: 22.5
Time: 2:00 hr

*Attempt **ALL questions**. Write your answer in detail as far as possible.*

Group A [3 × 5 = 15]

1. Let $u = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ and $v = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$.

   (a) Write the vector $\begin{pmatrix} 3 \\ 2 \end{pmatrix}$ in terms of the vectors u and v.

   (b) Show that the vectors u and v span $\mathbb{R}^2$.

2. How many kinds of subspaces of $\mathbb{R}^3$ are there? Mention them. Show that a plane through the origin is a two-dimensional subspace of $\mathbb{R}^3$.

3. What is the angle between the diagonal of the unit cube in the positive orthant and the vector e_1?

4. Define linearly independent vectors. Prove that an orthogonal set of nonzero vectors in a vector space is linearly independent.

5. Let $v_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $v_2 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix}$. Show that $B = \{v_1, v_2\}$ is an orthonormal basis for $\mathbb{R}^2$. Find a vector $x \in \mathbb{R}^2$ with respect to the basis B.

Group B [6 × 5 = 30]

6. Define the span of a set. Prove that $\text{span}(\{v_1, v_2, \ldots v_k\}) \subseteq V$ is a subspace of a vector space V. Also, show that if $x \in \mathbb{R}^2$, such that $x \neq 0$, then $x^\perp$ is a subspace of $\mathbb{R}^2$.

7. Explain four major ways to view a matrix. Prove that if $S : \mathbb{R}^l \to \mathbb{R}^m$ and $T : \mathbb{R}^n \to \mathbb{R}^l$ are linear transformations, given by matrices A and B, respectively, then, the composition $S \circ T : \mathbb{R}^n \to \mathbb{R}^m$ is a linear transformation and is given by AB.

8. By showing that the L_1-norm satisfies each of the conditions in the definition of a norm prove this is a vector norm. First do this for $\mathbb{R}^2$, and then do this for $\mathbb{R}^n$.

Scanned with CamScanner

OR

Prove that if $x \in \mathbb{R}^n$, then $\|x\|_\infty \leq \|x\|_1 \leq n\|x\|_\infty$.

9. Let

$$V = \left\{ \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \in \mathbb{R}^3 : \; x_1 + x_2 + x_3 = 0 \right\}, \qquad B = \left\{ \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} \right\}$$

Show that B is a basis for V.

OR

Let V be a subspace of $\mathbb{R}^n$ and w a vector in $\mathbb{R}^n$.

(a) If $\{v_1, v_2, \ldots, v_k\}$ is an orthogonal basis for V, derive the expression for the projection of w onto V.

(b) If $\{v_1, v_2, \ldots, v_k\}$ is an orthonormal basis for V, derive the expression for the projection of w onto V.

10. Describe the Gram-Schmidt Process to transform a basis first for $\mathbb{R}^2$ and $\mathbb{R}^3$ with illustrations and then for $\mathbb{R}^n$ to an orthonormal basis.

***

Scanned with CamScanner

Tribhuvan University
Institute of Science and Technology
SCHOOL OF MATHEMATICAL SCIENCES
First Re Assessment 2078

Subject: Data Structures and Algorithms (MDS502) — Full Marks: 45
Course No: MDS 502 — Time: 2 Hrs.
Level: MDS 1st year 1st Semester — Pass Marks: 22.5

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

Group A [5 × 3 = 15]

1. Define ADT. What are the benefits of using ADT? (1 + 2)
2. ConvertA$B*C–D+E/F/(G+H)to prefix and postfix. (1.5 + 1.5)
3. What is priority queue? How can we make priority queue? (1 + 2)
4. What is recursion? Compare it with iteration. (1 + 2)
5. Define recursion. What are the benefits of using recursive algorithms? (1 + 2)

Group B [5 × 6 = 30]

6. What is complexity of algorithms? Explain Big-oh, Theta, and Omega notation in detail. (1.5 + 4.5)

OR

What is data structure? Why do we need it? Explain dynamic data structure and static datastructure with example. (1 + 2 + 3)

7. What are different applications of stack? How do you implement push and pop operations in Stack? Explain. (2 + 4)

OR

Explain algorithm to convert an infix expression to postfix with suitable example. (6)

8. Explain algorithm for evaluating postfix expression using suitable example (6)
9. What are different applications of queue. How do you implement enqueue and dequeue operations? Explain. (1 + 5)
10. Explain tail recursion with suitable program. (6)

***

Scanned with CamScanner

Master in Data Science
Pre-board Examination

Course Title: Data Structures and Algorithms (MDS502) Time: 2 Hrs.
Full Marks: 45 Pass Marks: 22.5

Group A
Attempt all questions. (5 × 3 = 15)

1. What is linked list? Compare singly linked list with double linked list. (1 + 2)
2. Why do we need header node in linked list? Explain circular linked list. (1.5 + 1.5)
3. Explain importance of sorting. Explain insertion sort. (1.5 + 1.5)
4. Compare linear search with binary search? What are their time complexities? (2 + 1)
5. What is hashing? Explain open hashing. (1 + 2)

Group B
Attempt all questions. (5 × 6 = 30)

6. How can you implement stack using linked list? Explain using suitable program. (6)

   OR

   How can you implement queue using linked list? Explain using suitable program. (6)

7. What is singly linked list? How can you insert and remove nodes in a singly linked list? Explain. (2 + 4)

   OR

   Com Com What is doubly linked list? How can you insert and remove nodes in a doubly singly linked list? Explain. (2 + 4)

8. Explain bubble sort. Hand test bubble sort algorithm with the array of numbers 34, 23, 17, 31, 45, 7, 21, 15, 8, and 1. (2 + 4)
9. Define searching. Explain binary search algorithm with suitable example. (1 + 5)
10. Explain linear probing. Suppose, the set of keys is {7, 12, 14, 10, 49, 58, 9, 50}, m = 10, and ***h(x) = x mod 10***. Show the effect of successively inserting these keys using linear probing.

Scanned with CamScanner

Master in Data Science
Mid-term Examination

Course Title: Data Structures and Algorithms (MDS502) Time: 2 Hrs.
Full Marks: 45 Pass Marks: 22.5

Group A
Attempt all questions. (5 × 3 = 15)

1. What is data type. How is it different from ADT? (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. What is recursion? Compare it with iteration. (1 + 2)
5. Explain recursive algorithm using suitable example. (3)

Group B
Attempt all questions. (5 × 6 = 30)

6. What is asymptotic notation? Explain Big-oh, Theta, and Omega notation in detail. (1.5 + 4.5)

   OR

   What is data structure? Why do we need it? Explain dynamic data structure and static datastructure with example. (1 + 2 + 3)
7. Define stack. How do you implement push and pop operations in Stack? Explain. (1 + 5)

   OR

   Explain algorithm to convert an infix expression to postfix with suitable example. (6)
8. Explain algorithm for evaluating postfix expression using suitable example (6)
9. Define queue. How do you implement queue operations using array data structure? Explain. (1 + 5)
10. Explain tail recursion with suitable program. (6)

Scanned with CamScanner

Master in Data Science
Pre-board Examination

Course Title: Data Structures and Algorithms (MDS502) Time: 2 Hrs.
Full Marks: 45 Pass Marks: 22.5

Group A
Attempt all questions. (5 × 3 = 15)

1. What is linked list? Compare singly linked list with double linked list. (1 + 2)
2. Why do we need header node in linked list? Explain circular linked list. (1.5 + 1.5)
3. Explain importance of sorting. Explain insertion sort. (1.5 + 1.5)
4. Compare linear search with binary search? What are their time complexities? (2 + 1)
5. What is hashing? Explain open hashing. (1 + 2)

Group B
Attempt all questions. (5 × 6 = 30)

6. How can you implement stack using linked list? Explain using suitable program. (6)

   OR

   How can you implement queue using linked list? Explain using suitable program. (6)

7. What is singly linked list? How can you insert and remove nodes in a singly linked list?

   Explain. (2 + 4)

   OR

   What is doubly linked list? How can you insert and remove nodes in a doubly singly linked list? Explain. (2 + 4)

8. Explain bubble sort. Hand test bubble sort algorithm with the array of numbers 34, 23, 17, 31, 45, 7, 21, 15, 8, and 1. (2 + 4)
9. Define searching. Explain binary search algorithm with suitable example. (1 + 5)
10. Explain linear probing. Suppose, the set of keys is {7, 12, 14, 10, 49, 58, 9, 50}, m = 10, and $h(x) = x \bmod 10$. Show the effect of successively inserting these keys using linear probing.

Scanned with CamScanner

Tribhuvan University
Institute of Sciences and Technology

SCHOOL OF MATHEMATICAL SCIENCES

First Re Assessment 2078

Subject: **Statistical Computing with R**

Full Marks: **45**

Course No: **MDS 503**

Pass Marks: **22.5**

Level: **MDS /I Year /I Semester**

Time: **2hrs**

Candidates are required to write answers with examples for answering question numbers 1-5in answer sheets and use laptop foranswering question numbers 6-10. R scripts and outputs/interpretation of question number 6-10 must be saved as R markdown filein a folder with name/exam roll number and submitted for grading.

Attempt ALL Questions.

Group A **[5 × 3 = 15]**

1. Explain how can you import following types of text files into the R software with codes:
   a) Tab separated text file
   b) Comma separated value text file
   c) Semi colon separated text file
2. Explain how youcan dosub-settingin R software with codes:
   a) Define the 5x5 matrix and select last two rows
   b) Select second and fourth row with third and fifth column
   c) Add 3 new rows in this matrix
3. Explain how to do these with codes in R:
   a) Define "gender" variable with male and female attributes as factor
   b) Check the attributes of the gender variable
   c) Check how the male and female values are stored in R
4. An object called "best_practice" is stored in R. now do as follows with codes:
   a) Define "Let", "the", "computer", "do", "the", "work" as elements of the "best_practice" object
   b) Write a function to print words (elements) of this object
   c) Write an improved functionwith loop to print the words (elements) of this object
5. Explain different types of pipe operators with codes and examples:
   a) Compound assignment operator
   b) Tee operator
   c) Exposition operator

Group B **[5 × 6 = 30]**

6. Do the followings with R script in R Studio:
   a) Define a column vector X with numbers between 1 and 30
   b) Define another column vector Y with cubes of X
   c) Combine the two column vectors in a new data frame called DF
   d) Get plot X and Y variables and decide which type of relationship is seen
   e) Get the appropriate correlation coefficient for this plot and interpret it carefully
7. Create a function and do as follows:
   a) Define a function: "roll"of a fair "die" twice with random sampling with replacement as true
   b) Get the first roll and interpret the result
   c) Get the second roll and interpret the result
   d) Get the third roll and interpret the result
   e) Write a summary of the results obtained in the earlier steps with conclusion

Scanned with CamScanner

8. Import the "covid_tbl.csv" data file in R studio as data frame and do as follows with R script:
   a) Check the structure of the data frame
   b) View the data frame: remove the first row and last column
   c) Change column names by adding underscore for the spaces
   d) Remove "+" and "%" from the columns where they appear
   e) Change attributes of the number variables from characters to numbers

9. Use the "mtcars" dataset of R and do as follows:
   a) Plot histogram of mpg variable and interpret it carefully
   b) Refine the histogram by filling the bars with "blue" color and changing number of bins to 10
   c) Add a vertical abline at mean of the mpg variable
   d) Plot Q-Q plot of mpg variable, add normal Q-Q line of red color on it and interpret it carefully
   e) Plot density plot of mpg variable without the border, fill it with yellow color and interpret it

OR

Use the ggplot2" package and do as follow in R studio:
   a) Define first layer with diamond data,carat as x-axis and price as y-axis
   b) Add layer with geometric aesthetic as "point", statistics and position as "identity"
   c) Add layers with scale of y and x variables as continuous
   d) Add layer with coordinate system as Cartesian
   e) Add layer with appropriate title and interpret the resulting graph carefully

10. Load the "igraph" package in R studio and do the basic SNA as follows with R scripts to:
   a) Define g as graph object with (1,2,3,4) as its elements
   b) Plot the g and interpret it carefully
   c) Define g1 as graph object with ("Sita", "Ram", "Rita", "Gita", "Gita", "Sita", "Sita", "Gita", "Anita", "Rita", "Ram", "Sita") as its elements
   d) Plot g1 with node color as green, node size as 20, link color as red and link size as 10 and interpret it
   e) Get degree, closeness and betweenness of g1 and interpret them carefully.

OR

Load the "rdm Tweets. rdata" file in R studio and do as follows with "tm" and "tweetR" packages:
   a) Convert twitter list as data frame and assign it as "df" object
   b) Create corpus using the"text" column of the data frame
   c) Perform pre-processing to clean the corpus for text mining
   d) Create term document matrix using the cleaned corpus
   e) Find the most frequent terms using the term document matrix
   f) Find the co-occurrence of the term "r" with filter of 0.1 and above.

***

Scanned with CamScanner

Tribhuvan University
Institute of Sciences and Technology
SCHOOL OF MATHEMATICAL SCIENCES
First Assessment 2078

Statistical Computing with R

Full Marks: **45**

Pass Marks: **22.5**

No: MDS 503

Time: **2hrs**

MDS /I Year /I Semester

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

Attempt ALL Questions.

Group A **[5 × 3 = 15]**

1. Explain how can you import following types of data into the R software with simple examples/codes:
   a) a text file saved in the local computer
   b) a table embedded in any webpage
   c) json file with web API
2. Explain the logic behind extraction of the following subsets from a 5x5 data frame in R software:
   a) First two rows
   b) Third and fifth row with second and fourth column
   c) Add 5 new rows in this data frame
3. Explain data mining in data science with focus and examples on:
   a) Tasks
   b) Analytics
   c) Learning's
4. Explain how to work efficiently with "big data" in R software in relation to the:
   a) Subsetting with base R and dplyr packages
   b) ff, ffbase and ffbase2 packages
   c) data. table package
5. Explain social network analysis and describe its use in a real-life situation with:
   a) Nodes
   b) Links
   c) Attributes

Group B **[5 × 6 = 30]**

6. Open the Ror R studio software and do the followings with R script:
   a) Define integers from 1 to 15 using three different coding approaches in R
   b) Define these five numbers: 1.1, 2.2, 3.3, 4.4 and 5.5 and save it as column vector N
   c) Add, subtract, multiply and divide vector R from vector N and interpret the results carefully
   d) Define a listusing "This" "is" "my" "first" "programming" "in" "R"and save it as L
   e) Transform these list elements as characters of UL object.
7. Import the "pollution.csv" file into R studio and do as follows with R script:
   a) Check the structure of the data and explain class of each variable
   b) Change the attributes of "particulate matter", "date time" and "value" variables
   c) Get the summary of all the variables and replace the outliers as missing value
   d) Get summary statistics of "value" variables by "particulate matter" variable categories
   e) Write a summary of the results obtained in the earlier steps with interpretation and conclusion

Scanned with CamScanner

8. Use the "pollution.csv" file imported and cleaned in R studio and do as follows with R script:
   a) Create bar plot of "particulate matter" variable
   b) Create histogram of "value" variable
   c) Create line plot of "date time" and "value" variables
   d) Create histogram of "value" variable by particulate matter categories
   e) Write a summary of the results obtained in the earlier steps with interpretation and conclusion

9. Load the "term Doc Matrix. R data" file into R studio and do as follows with R script:
   a) Define the term document matrix data object as matrix and store itas "m" object
   b) Define the frequencies of the terms using "row Sums" function and get the term frequencies
   c) Create a histogram of the term frequencies using ggplot2 package
   d) Create a histogram of the terms with 10 or more frequencies using ggplot2 package
   e) Create word cloud of term frequencies using word cloud package and interpret it carefully

OR

Load the "rdm Tweets. rdata" file in R studio and do as follows with "tm" and "tweetR" packages:
   a) Convert twitter list as data frame and assign it as "df" object
   b) Create corpus using the "text" column of the data frame
   c) Perform pre-processing to clean the corpus for text mining
   d) Create term document matrix using the cleaned corpus
   e) Find the most frequent terms using the term document matrix
   f) Find the co-occurrence of the term "r" with filter of 0.1 and above.

10. Load the "igraph" package in R studio and do the basic SNA as follows with R scripts to:
   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 ("S", "R", "R", "G", "G", "S", "S", "G", "A", "R") 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

Load the "term Doc Matrix. R data" file into R Studio and do as follows with R script:
   a) Define term Doc Matrix as matrix m
   b) Transform it into adjacency matrix
   c) Build an undirected SNA graph with the adjacency matrix data
   d) Remove loops and plot the SNA graph again
   e) Interpret all the results carefully

***

Scanned with CamScanner

Tribhuvan University
Institute of Sciences and Technology
SCHOOL OF MATHEMATICAL SCIENCES
Second Assessment 2078 (Open Book)

Subject: **Statistical Computing with R** *Full Marks*: **45**

Course No: **MDS 503** *Pass Marks*: **22.5**

Level: **MDS /I Year /I Semester** *Time*: **2hrs**

*Candidates are required to write answers on **their own language** for answering question numbers **1-5** in answer sheet and covert them as PDF file and use **R script** in R studio of their laptop for answering question numbers **6-10**. The **R scripts** of question 6-10 **must be knitted as PDF** file showing all the outputs/interpretations. **These PDF files must be uploaded/attached in the Second Assessment 2078 assignment of the MS Teams software**.*

Attempt ALL Questions.

Group A **[5 × 3 = 15]**

1. Compare following methods used in supervised learning models:
   a) Ordinary least square
   b) Gradient descent
   c) Maximum likelihood
2. Explain advantage and limitations of these concepts used in supervised learning:
   a) Validation
   b) Cross-validation
   c) Cross-validation with repetitions/replications
3. Explain when these models are used in supervised learning:
   a) Log-transformed models
   b) Polynomial models
   c) Neural network models
4. Differentiate supervised classification (decision tree) model using:
   a) Bagging
   b) Improved bagging
   c) Boosting
5. Explain unsupervised association rules mining with focus on:
   a) Method
   b) Use/example
   c) Limitations

Group B **[5 × 6 = 30]**

6. Do the following in R Studio with R script so that it can be knitted as PDF for review/grading:
   a) Prepare a data with 50 random observations and two variables: miles per gallon (mpg) with random range between 10 to 50 and transmission (am) as random binary variable (0=automatic, 1=Manual), **do not forget to use your class roll number as random seed to replicate the result**
   b) Perform goodness-of-fit test on miles per gallon (mpg) variable to check if it follows normal distribution or not
   c) Perform goodness-of-fit test on miles per gallon (mpg) variable to check if the variances of mpg are equal or not on am variable categories
   d) Perform the best independent sample t-test based on goodness-of-fit results and interpret it carefully
   e) Can you use the independent sample t-test for this data? Why?

Scanned with CamScanner

7. Do the followings in R Studio using R script so that it can be knitted as PDF for review/grading:
   a) Prepare a data with 100 random observations and four variables: miles per gallon (mpg) with random range between 10 to 50; transmission (am) as random binary variable (0=automatic, 1=Manual), weight (wt) with random range of 1 to 10 and horse power (hp) with random range of 125 and 400, **do not forget to use your exam roll number as random seed to replicate the result**
   b) Divide this data into train and test datasets with 70:30 random splits with your roll number as random seed
   c) Fit supervised linear regression model and KNN regression model on train data with miles per gallon (mpg) as dependent variable and all other variables as independent variable
   d) Get the summary of the model, fit indices and select the best model
   e) Predict the mpg variable in the test data with best model, get fit indices and interpret the results carefully

8. Do the following in R Studio with R script so that it can be knitted as PDF for review/grading:
   a) Prepare a data with 150 observations and four variables: transmission (am) as random binary variable (0=automatic, 1=Manual), miles per gallon (mpg) with random range between 10 to 50; weight (wt) with random range of 1 to 10 and horse power (hp) with random range of 125 and 400, **do not forget to use your exam roll number as random seed to replicate the result**
   b) Divide this data into train and test datasets with 80:20 random splits with your roll number as random seed
   c) Fit supervised binary logistic regression model and naïve Bayes classification model on train data with transmission (am) as dependent variable and miles per gallon (mpg), weight (wt) and horse power (hp) as independent variable and select the best model
   d) Predict the transmission variable in the test data using best model and interpret the result carefully
   e) Get the confusion matrix, sensitivity, specificity of the predicted model and interpret them carefully

9. Do as follows using "USArrests" dataset in R studio with R script so that it can be knitted as PDF for review/grading:
   a) Check the head and the structure of the dataset and interpret them carefully
   b) Use all the variables of this dataset in the Principal Component Analysis (PCA) to create "criminality score" for the US states
   c) Interpret the results of the PCA model carefully
   d) Get scree-plot and explain how many components must be retained
   e) Get the bi-plot of the fitted model and interpret it carefully
   f) Improve the fitted model with VARIMAX process and interpret the results carefully

OR

Do as follows using "USArrests" dataset in R studio with R script so that it can be knitted as PDF for review/grading:
   a) Get dissimilarity distance as "state.dissimilarity" object
   b) Fit a classical multidimensional model using the "state.dissimilarity" object
   c) Get the summary of the model and interpret it carefully
   d) Get the plot of the model and interpret it carefully
   e) Compare this model with the first two components from principal component analysis model in this data

10. Do as follows in the R Studio with R Script so that it can be knitted as PDF for review/grading:
   a) Define a 2-column matrix "x" with 50 normally distributed random observations for each column, **do not forget to set random seed as your roll number for replication**
   b) Define "dist" object of x matrix to compute 50x50 inter-observation Euclidean distance matrix
   c) Fit a hierarchical clustering model using "dist" object and single linkage and, get dendogram
   d) Fit a hierarchical clustering model using "dist" object and complete linkage and, get dendogram
   e) Show the number of clusters (k) to retain for the data using ablines in the dendogram of these models
   f) Get the best value of number of clusters to form (k) using the best model fitted above

Scanned with CamScanner

OR

Do as follows in the R Studio with R Script so that it can be knitted as PDF for review/grading:

a) Define a 2-column matrix "x" with 50 normally distributed random observations for each column, **do not forget to set random seed as your roll number for replication**
b) Assign the first 25 observation of the "x" matrix data as "1" and next 25 observation as "2" of a new column of this matrix
c) Fit a k-means clustering model in the "x" matrix data with k=2 and nstart = 20
d) Fit a k-means clustering model in the "x" matrix data with k=3 and nstart = 20
e) Plot the clusters formed by k=2 and k=3, compare the results and interpret them carefully
f) Visualize the clusters for best k value and interpret it carefully

***

Scanned with CamScanner

Tribhuvan University
Institute of Sciences and Technology

SCHOOL OF MATHEMATICAL SCIENCES

First Assessment 2078

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

ject: **Data Base Management System**
urse No: **MDS 503**
ee: **MDS /I Year /I Semester**

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

mpt *ALL* Questions.

Group A [5 × 3 = 15]

1. What is weak entity and weak relation? Explain with an example.
2. Differentiate between total participation and partial participation.
3. Why Normalization is carried out?
4. Differentiate between logical design and physical design.
5. Explain different types of cardinalities.

Group B [5 × 6 = 30]

6. Justify why database management system is efficient than file management system.
7. Find Attribute Closure of the following relation

STU_ID	NAME	Age	FACULTY	ADDRESS
1	Bikash	20	Computer	Kathmandu
2	Anju	21	Electronics	Patan
3	kiran	22	computer	Bhaktapur
4	Anju	23	Computer	Patan

8. Convert the given relation to 3 NF.

ID	ROLL	HOSTEL ROOM	FACULTY	PRICE
1	1	112	Computer	1000
2	1	113	Electronics	2000
3	2	112	Data Science	1000
1	3	114	Data Science	5000

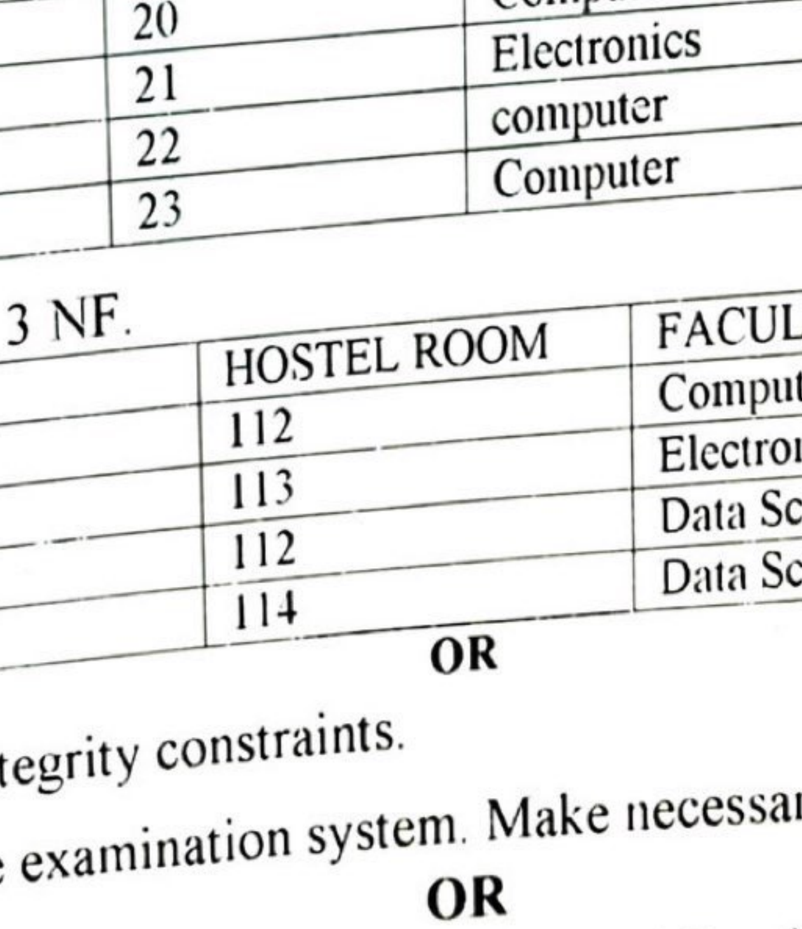
OR

Explain different types of Integrity constraints.

9. Draw ER diagram for online examination system. Make necessary assumptions.

OR

Assume attributes of CUSTOMER, ORDER, PRODUCT and Employee with required primary key.
Convert the ER diagram to relational model.



Scanned with CamScanner

Tribhuvan University
Institute of Science and Technology
SCHOOL OF MATHEMATICAL SCIENCE
First ReAssessment 2078

Course No: MDS 505
Subject: Data Base Management System
Level: MDS 1year/I Semester

Full Mark: 45
Pass Mark: 22.5
Time: 2 Hrs

Candidates are required to give their answer in their own words as far as practicable. The figures in the margin indicate
marks

Group A (5x3=15)

1. What is partial dependency?
2. Illustrate the use of ISA relation.
3. How can we reduce redundancy?
4. Differentiate between one to many and many to many relation with an example.
5. Mention the rules for a table to be in 3NF.

Group B (5x6=30)

5 questions 6 marks, 2 or questions on same chapter

Explain different types of Integrity constraints

1. Draw ER diagram of online movie management system. Make necessary assumptions yourself
2. Convert the ER diagram of question no. 2 to Relational Model

OR

Mention any five rules that are required for conversion of ER diagram to Relational Model

3. Find Attribute Closure of the following relation

ID	P_NAME	LOCATION	COST
1	P1	L1	100
2	P2	L2	200
3	P1	L1	300
4	P3	L3	300

OR

Convert table of question 4 to 3 NF

4. What are the practical implications of Functional Dependencies? Explain its different types.

Scanned with CamScanner

Faculty: Data Science

FM: 45

Program: Masters

PM:

Subject: Database Management Systems

Time: 2 Hrs.

Group A (5*3=15)

1. What are the advantages of B tree index files?
2. Differentiate between database and data warehouse.
3. Explain slotted page structure of variable-length record.
4. Explain desirable properties of transaction.
5. Explain any three database access control mechanism.

Group B (5*6=30)

1. Highlight on different types of RAID.
2. How do we define equivalence of schedules? Provide examples.
3. What is a schedule? Define the concept of recoverable, cascade less, and strict schedule. Also compare them in terms of their recoverability.

OR

Which of the following schedule is (conflict) serializable? For each serializable schedule, determine the equivalent serial schedules.

i) r1(X):r2(x):w1(x):r2(y):w2(x)
ii) r2(x):r1(x):r3(x):w2(x):w1(x):w3(x)

4. Differentiate between supervised and unsupervised algorithm. What are the disadvantages of Apriori algorithm and how it can be solved?

OR

Divide the data points {(1, 1), ((2, 1), (4, 3), (5, 4)} into two clusters.

5. Write short notes on
   a. Information retrieval
   b. Properties of parallel database

Scanned with CamScanner