Tribhuvan University
Institute of Sciences and Technology

SCHOOL OF MATHEMATICAL SCIENCES

Second Assessment 2080

Subject: **Fundamentals of Data Science** | *Full Marks*: 45
Course No: **MDS 501** | *Pass Marks*: 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 × 3 = 15]

1. Elaborate on agility (agile implementation) of CRISP-DM.
2. Explain why data preprocessing is a major job of a data scientist.
3. Why is logistic regression a regression despite its use for classification tasks? Illustrate the sigmoid curve along with its mathematical function.
4. Define entropy. How is it used with decision tree?
5. Explain why data ethics is crucial in data science.

Group B [5 × 6 = 30]

6. Explain the OSEMN framework for data science project implementation with any suitable example of your own.

OR

Discuss the scope and limitations for data analysis of election data. What responsibilities should be fulfilled by a data scientist for such projects. Discuss with one example of your own.

7. Discuss the problems caused by data quality during the training of machine learning algorithms. Do you recommend data first approach or model first approach for any of the data science problems? Explain with reason.
8. Apply Naïve bayes algorithm to decide if Married Female with salary of 42000 is probable to have illness or not based on data given below:

Marital Status	Gender	Income	Illness
Married	Male	40000	Yes
Unmarried	Male	35000	No
Married	Male	60000	Yes
Married	Female	61000	Yes
Unmarried	Female	36000	Yes
Married	Female	47500	No
Unmarried	Female	32000	No

OR

Explain the forward propagation and backward propagation of neural networks.

9. What do you mean by Random Forests? Why is random forest commonly used for feature selection despite being a machine learning model? Is random forest prone to over fitting ? Why or why not?

10. A healthcare organization developed a machine learning model to predict patients' risk of developing certain medical conditions based on their electronic health records (EHR). The model was trained using historical patient data, including diagnoses, treatments, and outcomes. However, it was later discovered that the model exhibited bias against patients from lower socioeconomic backgrounds. The training data disproportionately represented atients from wealthier neighborhoods who had better access to healthcare services and resources. As a result, the model erroneously associated higher socioeconomic status with lower health risks, leading to underestimating the risk of certain conditions for patients from disadvantaged backgrounds.

Explain how you can address the biases in such a situation. Be specific and suggest remedies to avoid biases.

***

Tribhuvan University
Institute of Sciences and Technology

SCHOOL OF MATHEMATICAL SCIENCES

Second Assessment 2080

Subject: **Data Structures and Algorithms**
Course No: **MDS 502**
Level: **MDS /I Year/I Semester**

Full Marks: 45
Pass Marks: 22.5
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. Explain importance of sorting. Explain insertion sort. (1.5 + 1.5)
2. Compare linear search with binary search? What are their time complexities? (2 + 1)
3. What is hashing? Explain open hashing. (1 + 2)
4. Explain almost complete binary tree with example. (3)
5. What is adjacency matrix representation of the graph? (3)

Group B [5 × 6 = 30]

6. Explain merge sort. Trace the execution of merge sort algorithm with the array of numbers34, 23, 17, 31, 45, 7, 21, 15, 8, and 1. (2 + 4)

OR

Explain shell sort. Trace the execution of shell sort algorithm with the array of numbers 34, 23, 17, 31, 45, 7, 21, 15, 8, and 1. (2 + 4)

7. Define searching. Explain binary search algorithm with suitable example:(1 + 5)

OR

How do you implement both sequential search and binary search algorithms? (6)

8. 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. (2 + 4)
9. What is AVL tree. Construct AVL tree for the sequence 21, 26, 30, 9, 4, 14, 28, and 18. (1 + 5)
10. What is graph traversal? Explain both depth first search (DFS) algorithm for traversing graphs with example. (1 + 5)

***

Tribhuvan University
Institute of Sciences and Technology

SCHOOL OF MATHEMATICAL SCIENCES

Second 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 the answer sheet and use laptop for answering question numbers* ***6-10 with R scripts****. R scripts must be knitted as HTML with the outputs/interpretation of question number 6-10 and it must be saved in a folder with the HTML file and the name/exam roll number for grading.*

Attempt ALL Questions.

Group A **[5 × 3 = 15]**

1. Describe supervised learning with focus on:
   a) Grammar of graphics
   b) Layers in grammar of graphics
   c) Statistical transformations in grammar of graphics
2. Describe supervised learning linear regression model with focus on:
   a) Pre-requisites before fitting this type of model
   b) Multicollinearity and its importance in the multiple linear regression model
   c) Best regularized regression model to control the multicollinearity in multiple linear regression model
3. Describe supervised learning classification regression model with focus on:
   a) Model fit indices b) Confusion matrix c) Prediction accuracy indices
4. Describe supervised learning method with focus on:
   a) Single layer, feed-forward neural network
   b) Activation functions used in the neural network models
   c) Network model with input, hidden and output layers
5. Describe decision tree classification model with focus on:
   a) Bagging b) Improved bagging c) Boosting

Group B **[5 × 6 = 30]**

6. Do the following in R Studio using ggplot2 package with R script:
   a) Create a dataset with following variables: age (18-99 years), sex (male/female), educational levels (No education/Primary/Secondary/Beyond secondary), socio-economic status (Low, Middle, High) and body mass index (14 – 38)with random 100 cases of each variable. Your roll number must be used to set the random seed.
   b) Create a line chart of age variable using ggplot2 package and interpret the result carefully
   c) Create scatter plot of age and body mass index variables using ggplot2 package and interpret the result carefully.
   d) Create classes of body mass index variable as: <18, 18-24, 25-30, 30+ and show it as pie chart using ggplot2 package and interpret it carefully
   e) Create classes of age variable as <15, 15-59 and 60+ and show it as bar diagram using the ggplot2 package and interpret it carefully
7. Do the followings in R Studio using "Bfox" dataset with R script:
   a) Divide the Bfox data into train and test datasets with 70:30 random splits
   b) Fit a supervised linear regression model and KNN regression model on train data with "debt" as dependent variable and all other variables as independent variable
   c) Get the summary of the model, fit indices and interpret them carefully
   d) Predict the debt variable in the test data and interpret the result carefully
   e) Get the fit indices (R-square, MSE, RMSE)of the predicted model and interpret them carefully

8. Do the following in R Studio using "Arrests" dataset with R script:
   a) Divide the mtcars data into train and test datasets with 80:20 random splits
   b) Fit a supervised logistic regression model and naïve bayes classification models on train data with "released" as dependent variable and colour, age, sex, employed and citizen as independent variable
   c) Get the confusion matrix, sensitivity, specificity of the fitted model and interpret them carefully
   d) Predict the transmission variable in the test data 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 "mtcars" dataset with R script:
   a) Create a "car scale" withall the variables variables using the Principal Component Analysis (PCA) model
   b) Interpret the results of the PCA carefully in terms of Kaiser's criteria
   c) Get scree plot and decide the number of components to be retained
   d) Get the bi-plot of the fitted model and interpret it carefully
   e) 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:
   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. Use the "mtcars" data and do as follows with R Script:
   a) Fit a hierarchical clustering model using single linkage and get the dendogram for this model
   b) Fit a hierarchical clustering model using complete linkage and get the dendogram for this model
   c) Fit a hierarchical clustering model using average linkage and get the dendogram for this model
   d) Show the number of clusters (k) to retain for the data using ablines in the dendogram of the best model
   e) Get the best value of number of clusters to form (k) using the fitted model above

OR

Load the "USArrests" data file into R Studio and do as follows with R script:
   a) Fit a k-means clustering model in the data with k=2 and 3
   b) Get the cluster means and within sum of square value each model and interpret them carefully
   c) Plot the clusters formed by k=2 and 3 and interpret them carefully
   d) Add cluster centers for plot with k=3 and interpret it carefully
   e) Visualize the clusters for k=3 and interpret it carefully

***

Tribhuvan University
Institute of Sciences and Technology

SCHOOL OF MATHEMATICAL SCIENCES

Second Assessment 2080

Subject: **Mathematics for Data Science** — *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. Prove that any two eigenvectors of a symmetric matrix corresponding to different eigenvalues are orthogonal.

2. Define the rank of a matrix. Reduce the matrix $A = \begin{pmatrix} 6 & 1 & 3 & 8 \\ 4 & 2 & 6 & -1 \\ 10 & 3 & 9 & 7 \\ 16 & 4 & 12 & 15 \end{pmatrix}$ to Echelon form and hence find the rank. [1+2]

3. Under what circumstance a system of homogeneous linear equations has a non-zero solutions? Find the values of k such that the system of homogeneous equations
$x + ky + 3z = 0,\ 4x + 3y + kz = 0,\ 2x + y + 2z = 0$
has non–zero solutions. [1+2]

4. State the conditions for the various types of definiteness of a quadratic form? For what value of α a quadratic form $Q(x) = \alpha x_1^2 - 6x_1x_2 + \alpha x_2^2$ is positive semi definite? [1+2]

5. Find the inverse of the matrix $A = \begin{bmatrix} 2 & 4 & 3 \\ 0 & 1 & 1 \\ 2 & 2 & -1 \end{bmatrix}$ by elementary row transformation.

Group B [5 ×6 = 30]

6. Define a bilinear form. Give an example. Given a symmetric bilinear map B: $\boldsymbol{R}^n \times \boldsymbol{R}^n \to \boldsymbol{R}$ defined by $B(u, v) = B_A(u, v) = u^T Av$ for all $u, v \in \boldsymbol{R}^n$ for some symmetric matrix A, then A is unique. [1+1+4]

7. Under what circumstance a system of homogeneous linear equations has a non-zero solutions? Determine the values of α and β for which the system
$3x - 2y + z = \beta,\quad 5x - 8y + 9z = 3,\quad 2x + y + \alpha z = -1$
has a unique solution, no solution and infinitely many solutions. [1+5]

OR

Under what circumstance a system of linear equations has a unique solution, no solution and infinitely many solutions? Find the values of k such that the system of homogeneous equations $x + ky + 3z = 0,\ 4x + 3y + kz = 0,\ 2x + y + 2z = 0$
has non–trivial solutions. [2+4]

8. Explain in brief the purpose of the inclusion of Linear Algebra in the Data Science curriculum? Describe various techniques Spectral theories are applied in Data Science and their works. [2+4]

OR

Describe the role of basic linear algebra techniques which are useful in the study of data science. Describe how the various concepts of System of Linear Equations are applied in Machine Learning. [2+4]

9. What is singular value decomposition? Find the singular value decomposition of the matrix

$$\begin{pmatrix} 0 & 2 \\ 1 & -2 \\ 1 & 1 \end{pmatrix}$$

[1+5]

10. Define four fundamental subspaces of a matrix A. Find the null space of the matrix $A = \begin{pmatrix} 2 & 3 \\ 8 & 12 \end{pmatrix}$. Find the quadratic form determined by A and remove the cross term of the Quadratic form. [1.5+1.5+1.5+1.5]

***

Tribhuvan University

SCHOOL OF MATHEMATICAL SCIENCES

Second Assessment 2080

Subject: **Database Management Systems** | *Full Marks*: 45
Course No: **MDS 505** | *Pass Marks*: 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 × 3 = 15]

1. How wait for graph is used to determine deadlock?
2. Why group by and having clauses are used in SQL? Support your answer with suitable examples.
3. How division operation is done in relational algebra? Illustrate with example.
4. When a relation is said to be in BCNF?
5. Define database schema and instance with example.

Group B [5 × 6 = 30]

6. How security in database can be ensured using discretionary and mandatory access control mechanisms. [6]

OR

How database is created using XML? Create a XML file to store information about student. Use your own assumptions if required. [2+4]

7. What is distributed database? Differentiate vertical fragmentation from horizontal fragmentation using suitable example. [2+4]

OR

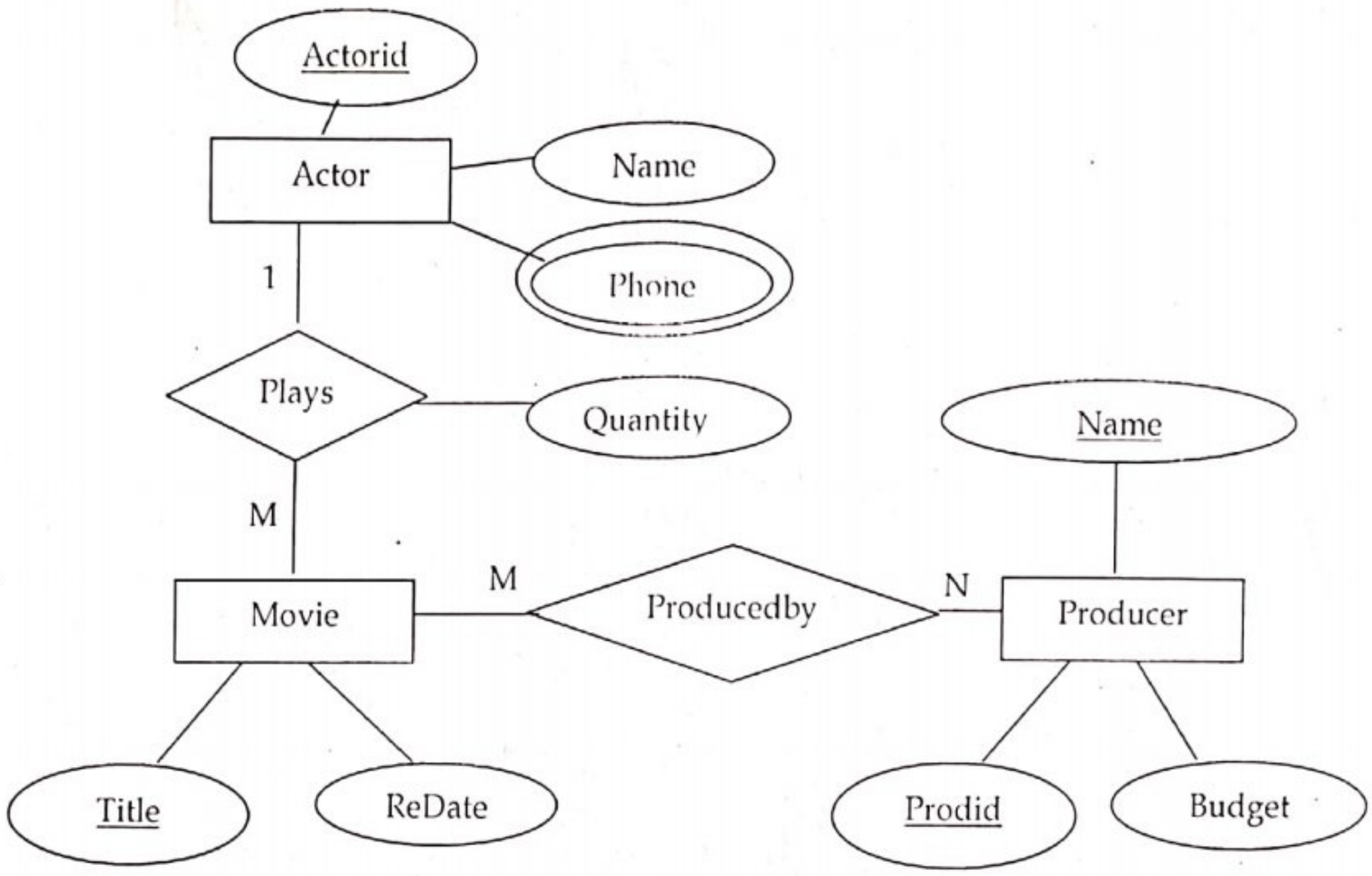
What is indexing? How sparse indexing is different from dense indexing? How B+ tree is used in indexing? [1+2+3]

8. What is query optimization? How materlization is different from pipelining? Prepare the query tree and optimized query tree for following SQL statement;

SELECT FNAME, ADDRESS FROM EMPLOYEE INNER JOIN DEPARTMENT ON DNO = DNUMBER WHERE DEPARTMENT.DNAME='Research';

[1+2+3]

9. Describe weak entity with an example. Convert following ER diagram into the equivalent relational schema. [2+4]



10. What is serializability? How can you determine whether a schedule is conflict serializable? Given following schedule, determine whether it is conflict serializable or not and justify your answer. [1+2+3]

T_1	T_2
read(A)	
$A := A - 50$	
write(A)	
	read(A)
	$temp := A * 0.1$
	$A := A - temp$
	write(A)
read(B)	
$B := B + 50$	
write(B)	
commit	
	read(B)
	$B := B + temp$
	write(B)
	commit

***