B.E. CSE (AI&DS) IV – Semester (AICTE) (Main & Backlog) (New), Aug / Sep ٢٠٢٤

Subject: Foundation of Data Science

Code No: F-١٣٧١٣/N/BL/AICTE – KEY

١. a) Define Data Science.

Data science is the study of data. It involves developing methods of recording, storing, and analyzing data to effectively extract useful information. The goal of data science is to gain insights and knowledge from any type of data — both structured and unstructured.

b) What is a random variable in statistics?

A real-valued function, defined over the sample space of a random experiment, is called a random variable. That is, the values of the random variable correspond to the outcomes of the random experiment. Random variables could be either discrete or continuous.

c)Steps for Simple Linear Regression Model Building

١. Importing libraries
٢. Loading dataset
٣. Split to independent and dependent variables
٤. Splitting data into training and testing data
٥. Choosing the Model
٦. Fit Our model
٧. Predict the output
٨. Plot the graph

d) Give an Example of Vectors and Lists.

A vector in R:

A vector is simply a list of items that are of the same type.

fruits <- c("banana", "apple", "orange")

A list in R:

A list in R can contain many different data types inside it. List١ <-list(١, "a", TRUE, ١+٢i)

e) How to Identify the Independent Variables in a Matrix?

In a matrix, independent variables are the columns that represent predictors or features that influence the dependent variable. Correlation or covariance analysis can help check if a variable is independent.

f) Define Clustering.

Clustering is a technique in machine learning that groups similar data points together based certain characteristics، with the aim of finding structures in data.

g) Write the Steps of Data Science Process.

- Problem understanding
- Data collection
- Data cleaning and preprocessing
- Data exploration and analysis
- Model building and evaluation
- Deployment and monitoring

٢. a) Illustrate the Tools for Data Science Model Building:
Data science model building relies on various tools and platforms. Some of the common used tools include:

- Python: Python is one of the most popular programming languages for data science. It ha a rich set of libraries like
  - o NumPy (for numerical computation)
  - o pandas (for data manipulation and analysis)
  - o matplotlib and seaborn (for data visualization)
  - o scikit-learn (for machine learning
  - o TensorFlow (for deep learning).
- R: R is another powerful language for data science، known for its excellent statistical analysis capabilities. R provides tools like
  - o ggplot٢ for data visualization
  - o packages like dplyr and tidyr for data manipulation.
- SQL: SQL (Structured Query Language) is essential for querying databases. Tools like
  - o MySQL
  - o PostgreSQL
  - o Oracle

  Database allow data scientists to retrieve and manipulate large datasets efficiently.
- Big Data Tools: For handling massive datasets، tools like
  - o Apache Hadoop
  - o Apache Spark

  They allow distributed data storage and processing across multiple nodes.
- Jupyter Notebooks: Jupyter notebooks offer an interactive coding environment where data scientists can write، test، and execute code in a sequential and modular fashion، making them ideal for data exploration. Tableau and Power BI:
- These are business intelligence tools for data visualization. They provide drag-and-drop functionality for creating dashboards and visualizing large datasets without needing to write code.

Each of these tools is chosen depending on the specific requirements of the project، the size of t data، and the type of analysis required.

b) Examine the Different Facets of Data with the Challenges in Their Processing:
Data has several facets that can pose significant challenges:

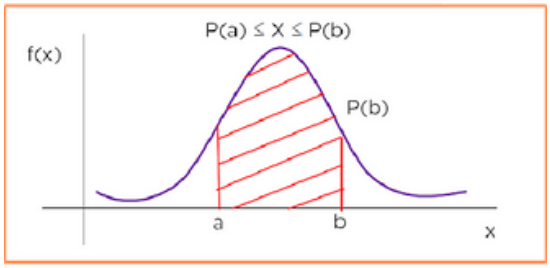
- Volume: Data is being generated at an unprecedented rate. Processing and storing such large volumes of data requires sophisticated tools and technologies like distributed storage (Hadoop HDFS، Google BigQuery) and parallel processing.

- Variety: Data comes in various forms, such as structured, semi-structured, and unstructured data. Structured data, such as that found in relational databases, is easy to store and analyze. However, semi-structured (like JSON, XML) and unstructured data (such as text, images, and videos) are much more difficult to manage. Velocity: The speed at which data is generated and needs to be processed is referred to as velocity. With real-time data from social media, IoT devices, and financial transactions, data scientists need to handle high-velocity data efficiently, often using streaming technologies like Apache Kafka. Veracity: The accuracy and quality of data are crucial. Often, datasets contain noisy, incomplete, or biased data, which can affect the results of a model. Cleaning and preprocessing data to ensure its quality is a major challenge in data science. Value: Not all data is useful. Identifying which data can provide valuable insights is another challenge. It involves separating meaningful signals from noise and ensuring that the models focus on relevant data.

٣. a)Explain Probability Density Function in Detail.

A function that defines the relationship between a random variable and its probability, such that you can find the probability of the variable using the function, is called a Probability Density Function (PDF) in statistics. In the case of a continuous randomvariable, the probability taken by X on some given value x is always ٠. calculate the probability of X lyingin an interval (a, b).

$$\mathbf{P(a \leq X \leq b) = \int_a^b f(x)\,dx}$$



Probability Density Function

b) Define Null Hypothesis and Alternate Hypothesis: Null Hypothesis (H٠) – This can be thought of as the implied hypothesis. "Null" meaning
"nothing." This hypothesis states that there is no difference between groups or no relationship
between variables. The null hypothesis is a presumption of status quo or no change.
Alternative Hypothesis (H١) – This is also known as the claim. This hypothesis should state what you expect the data to show, based on your research on the topic. This is your answer to your research question. Examples:

i. Null Hypothesis: H٠: There is no difference in the salary of factory workers based on gender.
Alternative Hypothesis: H١: Male factory workers have a higher salary than female factory workers.

ii. Null Hypothesis: H٠: There is no relationship between height and shoe size.
Alternative Hypothesis: H١: There is a positive relationship between height and shoe size

٤. a) Define Predictive Modeling:

Predictive modelling is a process used in data science to create a mathematical model that predicts an outcome based on input data. It involves using statistical algorithms and machine learning techniques to analyze historical data and make predictions about future or unknown events.

Types of Predictive Models There are severaltypes of predictive models, each suitable for different types of data and problems. Here are some common types of predictive models: Linear Regression: Linear regression is used when the relationship between the dependent variable and the independent variables is linear. It is often used for predicting continuous outcomes. Logistic Regression: Logistic regression is used when the dependent variable is binary (i.e., has two possible outcomes). It is commonly used for classification problems. Decision Trees: Decision trees are used to create a model that predicts the value of a target variable based on several input variables. They are easy to interpret and can handle both numerical and categorical data. Random Forests: Random forests are an ensemble learning method that uses multiple decision trees to improve the accuracy of the predictions. They are robust against overfitting and can handle large datasets with high dimensionality.

b) Fit a Multiple Linear Regression Model:

y	140	155	159	179	192	200	212	215
x_1	60	62	67	70	71	72	75	78
x_2	22	25	24	20	15	14	14	11

Steps to fit a multiple linear regression model to this dataset

1. calculate $x_1^2, x_2^2, x_1y, x_2y, x_1x_2$

y	x_1	x_2
100	60	22
155	62	25
159	67	24
179	70	20
192	71	15
200	72	14
212	75	14
215	78	11
Sum	1452	555	145
Mean	181.5	69.375	18.125

x_1^2	x_2^2	x_1y	x_2y	x_1x_2
3600	484	8400	3080	1320
3844	625	9610	3875	1550
4489	576	10653	3816	1608
4900	400	12530	3580	1400
5041	225	13632	2880	1065
5184	196	14400	2800	1008
5625	196	15900	2968	1050
6084	121	16770	2365	858
Sum	38767	2823	101895	25364	9859

2) Calculate Regression sums.

$\Sigma x_1^2 = \Sigma x_1^2 - (\Sigma x_1)^2/n$ } variance of first variable x_1

$= 38767 - (555)^2/8$

$= 263.875$

$\Sigma x_2^2 = \Sigma x_2^2 - (\Sigma x_2)^2/n$ } variance of second variable x_2

$= 2823 - (145)^2/8$

$= 194.875$

$\Sigma x_1 y = \Sigma x_1 y - (\Sigma x_1 * \Sigma y)/n$ } covariance between x_1 and y

$= 101895 - (555 * 1452)/8$

$= 1162.5$

$\Sigma x_2 y = \Sigma x_2 y - (\Sigma x_2 * \Sigma y)/n$ } covariance between x_2 and y

$= 25364 - (145 * 1452)/8$

$= -953.5$

$\Sigma x_1 x_2 = \Sigma x_1 x_2 - (\Sigma x_1 * \Sigma x_2)/n$ } covariance between x_1 and x_2

$= 9859 - (555 * 145)/8$

$= -200.375$

$\Sigma x_1^2 = 263.875$

$\Sigma x_2^2 = 194.875$

$\Sigma x_1 y = 1162.5$

$\Sigma x_2 y = -953.5$

$\Sigma x_1 x_2 = -200.375$

3) calculate b_0, b_1 and b_2

$$b_1 = \frac{(\Sigma x_2^2)(\Sigma x_1 y) - (\Sigma x_1 x_2)(\Sigma x_2 y)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

$$= \frac{(194.875)(1162.5) - (-200.375)(-953.5)}{(263.875)(194.875) - (-200.375)^2}$$

$= 3.148$

$$b_2 = \frac{(\Sigma x_1^2)(\Sigma x_2 y) - (\Sigma x_1 x_2)(\Sigma x_1 y)}{(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2}$$

$$= \frac{(263.875)(-953) - (-200.375)(1152.5)}{(263.875)(194.875) - (-200.375)^2}$$

$= -1.656.$

$b_0 = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2$

$b_0 = 181.5 - 3.148(69.375) - (-1.656)(18.125)$

$= -6.867$

Place b_0, b_1 and b_2 in the estimated linear regression eqn

$\hat{y} = b_0 + b_1 * x_1 + b_2 * x_2$

$\hat{y} = -6.867 + 3.148 x_1 - 1.656 x_2$

- ٥. a) Explain Control Structures Using R Code.

Control structures are essential in programming for directing the flow of execution of the code. They allow the programmer to execute certain blocks of code conditionally، repeatedly، or in loops.

In R، control structures include conditional statements، loops، and branching statement

➢ Conditional Statement:

- if Statement The if statement executes a block of code if a specified condition is true.

```
x <- ١٠
if (x < ٥) {
  print("x is greater than ٥")
}
```

- if-else Statement The if-else statement provides an alternative path if the condition is false.

```
x <- ٣
if (x < ٥) {
  print("x is greater than ٥")
} else {
```

```
    print("x is less than or equal to ٥")
}
```

- ifelse Function ifelse is a vectorized form of the if-else statement.

```
x <- c(٢، ٧، ٥)
result <- ifelse(x < ٥، "Greater"، "Lesser or Equal")
print(result)
```

➢ **Looping Structures**

- Loops allow repetitive execution of a block of code.
- for Loop A for loop is used to iterate over a sequence or vector.

```
for (i in ١:٥) {
  print(i)
}
```

- while Loop A while loop continues to execute as long as a specified condition is true.

```
i <- ١
while (i <= ٥) {
  print(i)
  i <- i + ١
}
```

- repeat Loop The repeat loop is used for infinite loops unless a condition is met to stop it.

```
i <- ١
repeat {
  print(i)
  i <- i + ١
  if (i > ٥) {
    break
  }
}
```

b)How toCombineTwoDataFrames، withExample.

We use the rbind() and the cbind() functions to combine two data frames together in R.
rbind() – combines two data frames vertically
cbind() – combines two data frames horizontally
Combine Dataframe Vertically Using rbind() in R
If we want to combine two data frames vertically، the column name of the two data frames

must be the same.
For example،

```
create a data frame
dataframe١ <- data.frame(

  Name = c("Juan"، "Alcaraz")،
  Age = c(٢٢، ١٥)
)
create another data frame
dataframe٢ <- data.frame(
```

```
  Name = c("Yiruma", "Bach", "Ludovico"),
  Age = c(٤٦, ٨٩, ٧٢)
)

combine two data frames vertically
updated <- rbind(dataframe١, dataframe٢)
print(updated)
```

output:

```
     Name Age
١    Juan  ٢٢
٢ Alcaraz  ١٥
٣  Yiruma ٤٦
٤    Bach  ٨٩
٥ Ludovico ٧٢
```

Combine Dataframe Horizontally Using cbind() in R
The cbind() function combines two or more data frames horizontally.
For example,

```
create a data frame
dataframe١ <- data.frame(
  Name = c("Juan", "Alcaraz"),
  Age = c(٢٢, ١٥)
)

create another data frame
dataframe٢ <- data.frame(
  Hobby = c("Tennis", "Piano")
)

combine two data frames horizontally
updated <- cbind(dataframe١, dataframe٢)
print(updated)
```

output:

```
     Name Age Hobby
١    Juan  ٢٢ Tennis
٢ Alcaraz  ١٥ Piano
```

٦. a) Explain K-NN Algorithm with an Example.

K-Nearest Neighbors (K-NN) Algorithm: KNN is a simple, non-parametric, and lazy learning algorithm used for classification and regression. It assumes that similar data points are close to each other in a multi-dimensional space, so the algorithm predicts the class of a new data point based on its K nearest neighbors from the training dataset. Working of the K-NN Algorithm

١. Data Representation:
K-NN assumes that data points can be represented in a multi-dimensional spac
where each point has coordinates corresponding to feature values.

২. Choosing K:
The user specifies the number of nearest neighbors, K, to consider. A smaller value of K may lead to more specific (and potentially noisy) predictions, while a larger K may smooth out predictions but might lose detail.
৩. Distance Metric:
K-NN uses a distance metric (usually Euclidean distance) to determine how close the training points are to the test point. Thedistance between two points

$$d(A, B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

৪. Prediction:
For Classification: The new data point isassigned to the class that is most frequent among its K nearest neighbors.
For Regression: The predicted value isthe average of the K nearest neighbors' values.

b) What is a Confusion Matrix? Write ৪ Performance Measures for Classification.

A confusion matrix shows the performance of a classification algorithm by comparing th predicted classes to the actual classes.

		ACTUAL	
		Negative	Positive
PREDICTION	Negative	TRUE NEGATIVE	FALSE NEGATIVE
	Positive	FALSE POSITIVE	TRUE POSITIVE

Where:

True Positive (TP): The model correctly predicted a positive outcome.
True Negative (TN): The model correctly predicted a negative outcome.
False Positive (FP): The model incorrectly predicted a positive outcome (Type I error).
False Negative (FN): The model incorrectly predicted a negative outcome (Type II error).
Performance measures:

- Accuracy:
  Accuracy measures the proportion of correctly predicted instances (both true positives and true negatives) out of the total instances.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

- Precision:
  Precision (also called positive predictive value) measures the proportion of correctly predicted positive observations out of all predicted positives.

$$Precision = \frac{TP}{TP + FP}$$

- Recall:
  Recall (or sensitivity) measures the proportion of actual positives that were correctly predicted.

$$Recall = \frac{TP}{TP + FN}$$

- F1-Score:
  The F1-score is the harmonicmean of precision and recall, providing a balanced measure that considers bothfalse positives and false negatives.

$$F1 - score = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$$

v. a) Applications of Data Science:

- Healthcare

Data science aids in medical imaging, predictive analytics for disease outcomes, and drug discovery by analyzing patient data and medical records.

• Finance:
Used for fraud detection, algorithmic trading, and credit scoring, leveraging patterns in transaction data to predict risks and improve decision-making.

• Retail & E-commerce:
Data science powers recommendation systems, inventory management, and customer sen analysis, enhancing customer experience and operational efficiency.

• Transportation & Logistics:
Enables route optimization, predictive maintenance, and autonomous vehicles, improving logistics, safety, and fuel efficiency.

• Marketing & Advertising:
Helps in customer segmentation, ad targeting, and campaign optimization, personalizing marketing efforts to improve engagement and ROI.

• Social Media:
Used for content recommendations, sentiment analysis, and influencer analytics, enhancin engagement and brand strategies.

• Manufacturing:
Powers predictive maintenance, quality control, and supply chain optimization, increasing productivity and reducing downtime in production processes.

b) Debugging and Simulation in R Programming.

Debugging is a process of cleaning a program code from bugs to run it successfully. While writing codes, some mistakes or problems automatically appears after the compilation of code and are harder to diagnose. So, fixing it takes a lot of time and after multiple levels of calls.

Debugging in R is through warnings, messages, and errors. Debugging in R means debugg functions. Various debugging functions are:

Editor breakpoint

traceback()
browser()
recover()

- Editor Breakpoints

Editor Breakpoints can be added in RStudio by clicking to the left of the line in RStudio or pressing Shift+F9 with the cursor on your line. A breakpoint is same as browser() but it doesn't involve changing codes. Breakpoints are denoted by a red circle on the left side, indicating that debug mode will be entered at this line after the source is run traceback() Function

- The traceback() function is used to give all the information on how your function arrived at an error. It will display all the functions called before the error arrived called the "call stack"
- in many languages, R favors calling traceback.

browser() Function

browser() function is inserted into functions to open R interactive debugger. It will stop the execution of function() and you can examine the function with the environment of itself. In debug mode, we can modify objects, look at the objects in the current environment, and also continue executing.

Simulation Generating random data or repeating experiments using functions like sample() morm() runif() in R. Simulations can be used to model complex systems.

Code No: F-13713/N/BL/AICTE

FACULTY OF ENGINEERING

B.E. CSE (AI&DS) IV - Semester (AICTE) (Main & Backlog) (New) Examination, August / September 2024

Subject: Foundation of Data Science

Time: 3 Hours **Max. Marks: 70**

Note: (i) First question is compulsory and answer any four questions from the remaining six questions. Each question carries 14 Marks.
(ii) Answer to each question must be written at one place only and in the same order as they occur in the question paper.
(iii) Missing data, if any, may be suitably assumed.

1. a) Define Data Science.
   b) What is random variable in statistics?
   c) Write the steps for simple linear regression model building.
   d) Write an example of Vectors and lists.
   e) How to identify the independent variables in a matrix?
   f) Define Clustering.
   g) Write the steps of data science process.

2. a) Illustrate the tools for data science model building.
   b) Examine the different facets of data with the challenges in their processing.

3. a) Explain Probability density function in detail.
   b) Define null hypothesis and Alternate hypothesis.

4. a) Define Predictive Modeling.
   b) Consider the following dataset with one response variable y and two predictor variable x1 and x2.

Y	140	155	159	179	192	200	212	215
x1	60	62	67	70	71	72	75	78
x2	22	25	24	20	15	14	14	11

   Fit a multiple linear regression model to this dataset.

5.a)Explain control structures using R code.
   b)How to combine two data frames, explain with an example.

6.a)Explain K-NN algorithm with an example.
   b)What is confusion matrix. Write any 4 performance measures for classification.

7.a)Discuss about the applications of Data Science.
   b)Write about debugging and simulation in R programming.

**

14 (a) Discuss basic Decision Tree Learning Algorithm with an example.
(b) Explain with example KNN Algorithm.

15 (a) Explain K-Means Algorithm.
(b) How association rules are generated using Apriori Algorithm?

16 (a) Explain useful measures of Association Rules
(b) Distinguish Data Mining, Machine Learning, Deep Learning, AI and Data Science.

17 (a) Write short notes on Hyper planes and Half spaces.
(b) Write short notes on Sample Statistic.

****

Code No. D-2407/AICTE

FACULTY OF ENGINEERING

B.E. VII - Semester (AICTE) (Main) Examination, March / April 2022

Subject: Open Elective – II
Data Science and Data Analytics

Time: 3 Hours **Max. Marks: 70**

(Missing data, if any, may be suitably assumed)

PART – A

Note: Answer all questions. **(10 x 2 = 20 Marks)**

1 Why Data Science?
2 What are Orthogonal Vectors? Give Example.
3 How do you characterise a Random phenomenon?
4 What is the motivation for Hypothesis Testing?
5 Explain response and a predictor variable.
6 Which function is used to implement Logistic Regression?
7 What do you mean by attribute-value pair?
8 Give the formula for calculating Information Gain.
9 Calculate the Support for pen, pencil and notebook?

Transactions	Item sets
T1	{pen, pencil, notebook}
T2	{pen, pencil, eraser}
T3	{sharpener, pencil, notebook}

10 What is Synthetic Data?

PART – B

Note: Answer any five questions. **(5 x 10 = 50 Marks)**

11 (a) "Mathematics in particular Linear Algebra is the language of Data Science"- Elucidate.
(b) Discuss how extracting solutions for matrix equations is the most fundamental aspect of Data Science.

12 (a) Elucidate why do we need a notion of a random variable and how it is useful in Data Science.
(b) "Hypothesis Testing is an important activity when you want to make decision from a set of data"-Discuss.

13 (a) Calculate the Kendall Rank Correlation Coefficient for the following data of two experts ranking the food items.

Food Item	Expert1	Expert2
1	1	1
2	2	3
3	3	6
4	4	2
5	5	7
6	6	4
7	7	5

(b) Explain with an example how to build multiple linear regression model.

Code No: 15263/AICTE

FACULTY OF ENGINEERING

BE VI-Semester (AICTE) (Main) Examination, October 2021
Subject: Data Science using R Programming
(Open Elective – II)

Time: 2 Hours **Max. Marks: 70**

(Note: Missing data, if any, may be suitably assumed)

Answer any five questions. **PART - A** **(5 x 2 = 10 Marks)**

1 List the advantages of R programming and over other programming languages?
2 What are the different R packages?
3 How do you create a Data frame in R?
4 Distinguish between invalid values and outliers.
5 What is the binary logistic regression?
6 How do you do a regression analysis in R?
7 What is decision tree method?
8 How do you write time series data in R?
9 Define Clustering.
10 Differentiate supervised learning and unsupervised learning?

Answer any four questions. **PART - B** **(4 x 15 = 60 Marks)**

11 (a) Discuss datatypes in R with examples.
(b) Explain the challenges of Analytical Data Processing.

12 (a) Discuss briefly the R function for understanding Data in data frames?
(b) Explain the concept of finding the missing values in R?

13 (a) Describe briefly about the assumptions of linear regression?
(b) Describe about multinomial logistic repression model?

14 (a) What are the issues in decision tree Learning? Explain them in detail.
(b) Explain the method of decomposing the Time series data?

15 (a) Explain the K-mean algorithm?
(b) Discuss the sequential process of text mining process?

16 (a) Discuss the steps involved in Downloading and Installing R?
(b) Describe the concept of load Data frames?

17 Write short notes on:
(a) CURE Algorithm
(b) Pattern evaluation method

* * *

Time: 3 Hours **Max. Marks: 70**

Note: ***Answer all questions from Part-A & answer any five questions from Part-B and each question carries 10 Marks.***

PART – A (20 Marks)

1. What is Data Science.? How it is different from Data Analysis?
2. Why Linear Algebra is significant in Data Science?
3. List out the differences between Probability Mass Functions and Probability Density Functions.
4. What is a hypothesis and how it is tested?
5. Define Predictive modeling. List out the types of Predictive modeling.
6. Write the possible ways of improving the accuracy of a linear regression model.
7. Briefly describe the data structures in R Programming Language.
8. Write a program to find the factorial of a number using R.
9. Define Object. List the methods for measuring Distance between Objects.
10. What is classification? Draw and explain learning and classification process.

PART – B (5 x 10 = 50 Marks)

11. Draw and explain the life cycle of Data Science. 10
12. Explain K-Means Algorithm and its implementation in R Programming Language. 10
13. Why logistic regression is used for classification. Explain model building strategies for logistic regression. 10
14. Define list and data frame in R and explain various operations on lists and data frames with suitable examples. 10
15. Explain K-Nearest Neighbours Algorithm and its implementation in R Programming Language. 10
16. What is the purpose sample statistics? Explain the properties of sample statistics? 10
17. What is statistical hypothesis? Briefly describe the various test statistics? 10

****

Answer any five questions. **(5x2 = 10 Marks)**

1. What is Data Science? How it is different from Business Intelligence?
2. List out the applications of Eigen vectors and Eigen values in Data Science?
3. How random variables are different from traditional variables used in algebra?
4. List out the properties of Probability Mass and Density functions?
5. What is linear regression? List out the critical assumptions of linear regression?
6. What is logistic regression? Explain with an example?
7. List out the various control structures supported by R Programming Language?
8. Write a program to find a given number is prime or not using R Programming Language?
9. List out the various performance metrics for classification?
10. Define clustering. List out the applications of clustering technique?

PART – B

Answer any four questions. **(4x15 = 60 Marks)**

11. Why Linear Algebra is significant in Data Science? How Linear Algebra is applied in Data Science?
12. What is statistical hypothesis? Briefly describe the various test statistics?
13. Explain K-Nearest Neighbours Algorithm and its implementation in R Programming Language?
14. Write the syntax for various data structures supported by R and explain with suitable examples?
15. Explain K-Means Algorithm and its implementation in R Programming Language?
16. What is Predictive modeling? Discuss about evaluation of Predictive models?
17. Why logistic regression is used for classification? Explain model building strategies for logistic regression.

****

(b) Write R-programming script to compute the sum of squares of N numbers. ($S_n=1^2+2^2+3^2+....n^2$)

15. Write the concept of normal distribution and explain an example with R code.

16.(a) Two teams, say the Cavs ant the Warriors, are playing a seven game championship series. The first to win four games, therefore, wins the series. The teams are equally good so that each have a 50-50 chance of winning each game. If the Cavs lose the first game, what is the probability that they win the series? Demonstrate with R..

(b) Describe an experiment of tossing a coin 80 times and prepare its frequency distribution.

17. Explain K-nearest neighbor technique and implementation of KNN in R.

******

PART – A

Note: Answers any five questions. **(5x2=10 Marks)**

1. What do you mean by Data Science?
2. Define Eigen Vector.
3. Write the basic features of R.
4. Draw a box plot of the following observations
   28, 42, 25, 34, 37, 26, 33, 28, 36, 33, 22.
5. Write R code to return a complex object.
6. What is correlation analysis?
7. What is Regression?
8. Define classification.
9. Differentiate table and data frame in R.
10. Write the purpose of clustering.

PART - B

Note: Answers any four questions. **(4x15=60 Marks)**

11. How is linear algebra used in Data Science? Describe the objects that operate on Vectors and Matrices.

12. (a) Define hyperplanes. Demonstrate the usage of hyperplane in data science with an example.

    (b) Write R program to create Pie chart for the following data.

    Houseing-600, Food-300, Clothes-150, Entertainment-100, Others-200.

13. Describe the different types of Statistical Testing methods. Demonstrate T-test in R with an example.

14. (a) Write the steps in R to create a data frame containing name and income of father for 5 individuals using edit command.

.....2

Code No.D-2

FACULTY OF ENGINEEERING

B.E. CSE (AI & DS) IV - Semester (AICTE) (Main) (New) Examination, September/ October – 2022

Time: 3 Hours **Subject: Foundation of Data Science** **Max. Marks: 70**

Note: (i) First question is compulsory and answer any four questions from the remaining six questions. Each Question carries 14 Marks.
(ii) Answer to each question must be written at one place only and in the same order as they occur in the question paper.
(iii) Missing data, if any, may be suitably assumed.

1. (a) Define Rank Nullity theorem.
(b) What are orthogonal vectors? Give an example.
(c) Define Probability/mass density function.
(d) How matrices are differ from Data Frames?
(e) Define random variable? Give one example.
(f) Write the R code to read the data from SQL databases.
(g) List out any four performance measures.

2. (a) What are various Data Science Toolkits?
(b) Explain different cases of solving Linear equations.

3. (a) Explain about the Hypothesis Testing and mention its error.
(b) Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.
   i. What is the mean of the data? What is the median?
   ii. What is the mode of the data? Comment on the data's modality
   iii. What is the midrange of the data?
   iv. Can you find (roughly) the first quartile (Q1) and the third
   v. quartile (Q3) of the data?
   vi. Give the five-number summary of the data.
   vii. Show a boxplot of the data

4. (a) Explain Multiple linear regression model with a suitable example.
(b) Differentiate between Logistic regression model and linear regression model.

5. (a) How to declare functions in R Language? Explain with suitable example?
(b) What are the data frames? Write its significance in R-Language?

6. (a) Describe about the k-NN algorithm & its implementation in R.
(b) Describe about the k-mean algorithm & its implementation in R.

7. (a) Discuss about Data Science Process with neat diagram.
(b) Write short notes on Hyper planes & Half spaces?

****

Code No: E-5713/N/BL/AICTE

FACULTY OF ENGINEERING

B.E. CSE(AI&DS) IV - Semester (AICTE) (Main & Backlog) (New) Examination, September /October 2023

Subject: Foundation of Data Science

Time: 3 Hours

Max. Marks: 70

Note: (i) First question is compulsory and answer any four questions from the remaining six questions. Each questions carries 14 Marks.
(ii) Answer to each question must be written at one place only and in the same order as they occur in the question paper.
(iii) Missing data, if any, may be suitably assumed.

1. a) Distinguish between business intelligence and data science.
   b) What are various types of data used in Data Science?
   c) Define Statistical model.
   d) Difference between linear regression and logistic regression.
   e) What are various features of R programming?
   f) Training data and testing data should be in the ratio of 80:20. Justify this statement.
   g) Compare in between clustering & classification.

2. a) Explain in detail about the stages of a data science project.
   b) Consider A with attribute {x1,x2,x3}.

$$A = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 4 & 0 \\ 3 & 6 & 1 \end{bmatrix}$$

such that Identify Rank of A & Nulity. Write R code for it.

3. a) Explain in detail NULL hypothesis and Alternate Hypothesis with suitable example.
   b) Explain Binomial Mass Function, Normal Density Function & Chi square Density function with example.

4. a) What is correlation coefficient? Explain different types of correlation coefficients used in data science.
   b) Write the assumptions of binomial, multinomial logistic regression and multiple linear regression.

5. a) Write a R program reverse of given number 47896 .
   b) What are the various operators support to the R language?

6. a) Explain briefly K-mean algorithm with implementation of R code.
   b) Explain briefly K-NN algorithm with implementation of R code.

7. a) Explain about eigenvalues and eigenvectors. Find eigenvalues & eigenvectors of matrix.

$$A = \begin{bmatrix} 5 & 4 \\ 1 & 2 \end{bmatrix}$$

   b) Explain briefly various performance measures for classification.

**