U.S.N.										

B.M.S. College of Engineering, Bengaluru-560019

Autonomous Institute Affiliated to VTU

February / March 2023 Semester End Main Examinations

Programme: B.E. **Semester: VII**
Branch: Institutional Elective **Duration: 3 hrs.**
Course Code: 21CS7OEDAS **Max Marks: 100**
Course: Data Science **Date: 22.02.2023**

Instructions: 1. Answer any FIVE full questions, choosing one full question from each unit.
2. Missing data, if any, may be suitably assumed.

UNIT - I

1 a) EMC's Global Innovation Network and Analytics (GINA) team is a group of senior technologists located in Centers of Excellence (COEs) around the world. This team's charter is to engage employees across global COEs to drive innovation, research, and university partnerships. In 2012, a newly hired director wanted to improve these activities and provide a mechanism to track and analyze the related information. In addition, this team wanted to create more robust mechanisms for capturing the results of its informal conversations with other thought leaders within EMC, in academia, or in other organizations, which could later be mined for insights. **14**

The GINA team thought its approach would provide a means to share ideas globally and increase knowledge sharing among GINA members who may be separated geographically. It planned to create a data repository containing both structured and unstructured data to accomplish three main goals.

- Store formal and informal data.
- Track research from global technologists.
- Mine the data for patterns and insights to improve the team's operations and strategy.

For the above case study brief out the design of data analytics lifecycle to analyze innovation data at EMC

b) Give a detailed note on the common tools used for the Model Building Phase **06**

UNIT - II

2 a) Calculate the following statistics for the variable Age (as shown in the Table) : (a)Mode (b) Median (c) Mean (d) Range (e) Variance (f) Standard deviation **12**

Variables *Name* and *Age*

Name	Age
P. Lee	35
R. Jones	52
J. Smith	45
A. Patel	70
M. Owen	24
S. Green	43
N. Cook	68
W. Hands	77
P. Rice	45
F. Marsh	28

b) Explain HYPOTHESIS TESTS with suitable example **08**

UNIT - III

3 a) A set of 10 hypothetical patient records from a large database is presented in Table below. Patients with a diabetes value of 1 have type-II diabetes and patients with a diabetes value of 0 do not have type-II diabetes. **10**

1. Create a new column by normalizing the Weight (kg) variable into the range 0–1 using the min–max normalization
2. Create a new column by binning the Weight (kg) variable into three categories: low (less than 60 kg), medium (60–100 kg), and high (greater than 100 kg).

3. Create an aggregated column, body mass index (BMI), which is defined by the formula: BMI = Weight(kg) / (Height(m))2

Table of Patient Records

Name	Age	Gender	Blood Group	Weight (kg)	Height (m)	Systolic Blood Pressure (mm Hg)	Diastolic Blood Pressure (mm Hg)	Temperature (°F)	Diabetes
P. Lee	35	Female	A Rh+	50	1.52	68	112	98.7	0
R. Jones	52	Male	O Rh−	115	1.77	110	154	98.5	1
J. Smith	45	Male	O Rh+	96	1.83	88	136	98.8	0
A. Patel	70	Female	O Rh−	41	1.55	76	125	98.6	0
M. Owen	24	Male	A Rh−	79	1.82	65	105	98.7	0
S. Green	43	Male	O Rh−	109	1.89	114	159	98.9	1
N. Cook	68	Male	A Rh+	73	1.76	108	136	99.0	0
W. Hands	77	Female	O Rh−	104	1.71	107	145	98.3	1
P. Rice	45	Female	O Rh+	64	1.74	101	132	98.6	0
F. Marsh	28	Male	O Rh+	136	1.78	121	165	98.7	1

b) Explain the following: **10**

a. Converting text to numbers
b. Converting continuous data to categories

UNIT - IV

4 a) Explain agglomerative hierarchical clustering and differentiate between single linkage, average linkage and complete linkage. **10**

b) Explain Chi-Square test with suitable example. **10**

OR

5 a) Generate frequent 1 itemset, frequent 2 itemset of the following transactions (minimum support=2) **10**

tid	Items brought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk

Calculate the support and confidence for the association rule Diaper ->Beer and Beer->Diaper take minimum support=50% of the above data set.

b) Explain K-Means clustering with suitable example **10**

UNIT - V

6 a) With suitable example Explain Linear and Logistic Regression **10**

b) Explain Bayes classifier with suitable example **10**

OR

7 a) Explain K-Nearest Neighbours method with example. **10**

b) What is the need for Decision Tree? Explain the concept of learning Decision Trees from data **10**

******