Reg. No ..... Name **22U306**

END SEMESTER EXAMINATION : OCTOBER 2022

SEMESTER 3 : INTEGRATED M.Sc. PROGRAMME COMPUTER SCIENCE - DATA SCIENCE

COURSE : **21UP3CRMCP07 : INTRODUCTION TO DATA SCIENCE**

(For Regular - 2021 Admission)

Time : Three Hours Max. Weightage: 30

PART A

Answer any 8 questions

1. List the various types of data that are collected for building recommender systems.
2. Define data munging.
3. State any two constituent technologies of data science.
4. Define the term 'metadata' in the context of a dataset.
5. Mention a drawback of SVD.
6. State the difference between relational and non - relational databases.
7. List any two factors that affects the quality of features.
8. Define the term 'population' with respect to bigdata.
9. List any two commonly used metrics that are used to test the accuracy of a machine learning algorithm.
10. State how a NULL hypotheses is defined in a hypotheses testing.

(1 x 8 = 8 Weight)

PART B

Answer any 6 questions

11. Form a decision tree based on a real-life example.
12. Given the following data that relates the number of users and total revenue of an online news site, predict the number of users if the total revenue for a particular month is Rs. 12500:
    (X, Y) = {(1, 25), (10, 250), (100, 2500), (200, 5000)}
13. Discuss how contextuality and temporality affects the peformance of a recommender system.
14. Discuss the importance of dimensionality reduction.
15. List the steps involved in implementing a linear regression algorithm.
16. Differentiate between descriptive statistics and inferential statistics in data analysis.
17. Briefly explain the format of an ANOVA table for a one-way classification.
18. Define data science technology.

(2 x 6 = 12 Weight)

PART C

Answer any 2 questions

19. Given below is the relationship between age and glucose level of certain people:

Age	43	21	25	42	57	59
Glucose Level	99	65	79	75	87	81

    Calculate the Pearson's correlation coefficient for the above data, perform hypotheses test and test its significance.

20. The following data shows the sales and advertising expenses of an Italian clothing company. If the amount spent on advertising in year 2010 is 62 million Euro, predict the Sales expenses in the year.

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009
Sales (Million Euro)	651	762	856	1063	1190	1298	1421	1440	1518
Advertising (Million Euro	23	26	30	34	43	48	52	57	58

21. The table given below displays data about Age and Glucose Level. Apply a suitable machine learning algorithm to predict the Glucose Level of a man aged 32. Also, calculate the error rate in the prediction.

Age	43	21	25	42	57	59
Glucose Level	99	65	79	75	87	81

22. Decompose the matrix $\begin{bmatrix} -3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix}$ using any decomposition technique.

(5 x 2 = 10 Weight)