18U281　　　　　　　(Pages: 2)　　　　　Name : …………….……….

Reg. No. :…………………..

Time: 3 Hours　　　　　　　　　　　　　　　　　　　Maximum: 80 Marks

PART A
Answer *all* questions. Each question carries 1 mark.

1. Matrix that has nonzero entries only in the diagonal is called _____

2. The smallest eigenvalue for every Laplacian matrix is _____

3. The weight of a newly born baby is an example _____ of type of random variable.

4. A variable representing the number of accidents occurring in a district on a day follows _____ distribution.
   a) Binomial　　　　b) Poisson　　　　c) Normal　　　　d) None of these

5. Which of the following is most important language for Data Science?
   a) Java　　　　b) Ruby　　　　c) R　　　　d) None of the above

6. Which of the following is used to identify the relationship between two numerical variables graphically?
   a) A scatter diagram　　　　　　b) Frequency Curve
   c) Ogive　　　　　　　　　　　d) Determining the Y intercept

7. Expansion of KDD is :
   a) Knowledge Database　　　　　b) Knowledge Discovery Database
   c) Knowledge Data House　　　　d) Knowledge Data Definition

8. _____ technique is also used in product advertisement.
   a) Phishing　　　b) Cookies　　　c) e-Banners　　　d) Spamming

9. Which of the following adds marginal sums to an existing table?
   a) par()　　　　　　　　　　　b) prop.table()
   c) addmargins()　　　　　　　　d) all of the mentioned

10. What would be the relation between the time taken by 1-NN,2-NN,3-NN.
    a) 1-NN >2-NN >3-NN　　　　　b) 1-NN < 2-NN < 3-NN
    c) 1-NN ~ 2-NN ~ 3-NN　　　　d) None of these

(10 x 1 = 10 Marks)

PART B

Answer any *eight* questions. Each question carries 2 marks.

11. What is classification?
12. Define statistical inference.
13. Define feature selection.
14. Define Poisson distribution.
15. Define discrete and continuous random variable.
16. Differentiate Population and sample.
17. What is systematic sampling?
18. What is datafication?
19. Define wrappers.
20. Define Eigen value and Eigenvector.
21. What is normalized cut?
22. What is data visualization?

(8 x 2 = 16 Marks)

PART C

Answer any *six* questions. Each question carries 4 marks.

23. Explain types of operators in R.
24. Explain how APIs and different tools used for web scraping.
25. Explain linear regression.
26. Explain naïve bayes classification.
27. What are the different sampling techniques?
28. Explain Girvan-Newman Algorithm.
29. Explain singular value decomposition.
30. What is spam? How do you identify spam emails?
31. Explain Exploratory Data Analysis.

(6 x 4 = 24 Marks)

PART D

Answer any *two* questions. Each question carries 15 marks.

32. a) Define social network. What are its characteristics?

    b) Briefly explain the different types of social networks.

33. a) Explain Dimensionality Reduction.

    b) Explain principal component analysis with example.

34. Define probability. Explain various probability distributions.

35. Explain decision tree algorithm. How to calculate entropy and information gain for attributes (features) of a dataset?

(2 x 15 = 30 Marks)

****