26/04/23
A.N

(Pages : 3) **R – 6721**

Reg. No. : .....

Name : .....

Second Semester M.A. Degree Examination, April 2023

Behavioural Economics and Data Science

BEDS 523 – FOUNDATIONS OF DATA SCIENCE

(2020 Admission onwards)

Time : 3 Hours Max. Marks : 75

SECTION – A

Answer **all** questions from this section. **Each** question carries **1** mark.

1. What is Data Science?
2. What is meant by hypothesis testing?
3. _____ Routines attempt to fill in missing values, smooth out noise while identifying outlines, and correct inconsistencies in the data.
4. _____ combines data from multiple sources into a coherent store
5. Using a decision tree, only categorical variables would be modeled. (True/False).
6. _____ Regression involves finding the "best" line to fit two variables so that one variable can be used to predict the other.
7. K-means is not deterministic and it also consists of number of iterations. (True/False).
8. Which clustering algorithm follows a top to bottom approach?

P.T.O.

9. Which visualization plot helps in determining outliers?

10. Which plot is appropriate to graph a single categorical variable?

(10 × 1 = 10 Marks)

SECTION – B

Answer any **seven** questions. **Each** question carries **5** marks.

11. What are the five Vs of big data?

12. How PCA is useful for attribute reduction?

13. How do you fit a data model? Explain.

14. Explain in detail about the process of binning. Smooth the below given data using median binning technique.

    3050,210, 350, 533,490, 533, 574, 625, 666, 697, 779,738, 943, 910, 984, 1025, 1068, 2007

15. What is exploratory data analysis? Explain

16. What is meant by logistic regression? Give suitable example.

17. Explain about Decision trees.

18. Explain how the value of k the number of clusters is determined in K means algorithm.

19. What is the difference between agglomerative clustering and divisive clustering?

20. Explain in detail about different tools used for data visualization.

(7 × 5 = 35 Marks)

SECTION – C

Answer any **three** questions. **Each** question carries **10** marks.

21. Use these methods to normalize the following sample of data: 350,450,550,750,1500

    (a) min-max normalization setting min=0 and max = 1

    (b) z-score normalization

    (c) z-score normalization using the mean absolute deviation instead of standard deviation.

22. What is the need of data reduction process? Explain the different techniques used for data reduction.

23. Draw the optimal decision tree model that can predict whether a student will pass the examination based on the given instance of dataset. The sample dataset contains 3 fields: CGPA (Low, Medium or High), Revised (Yes or No) and Passed (Yes or No). Explain the step-by-step process of optimal decision tree construction with calculations based on the sample dataset.

CGPA	REVISED	PASSED
LOW	NO	NO
LOW	YES	YES
MEDIUM	NO	NO
MEDIUM	YES	YES
HIGH	NO	YES
HIGH	YES	YES

24. Explain in detail DBSCAN algorithm.

25. Explain different methods used for visualization.

(3 × 10 = 30 Marks)
