STELLA MARIS COLLEGE (AUTONOMOUS) CHENNAI –600 086
(For candidates admitted from the academic year 2019 – 2020)

SUBJECT CODE: 19BI/PC/BD44

M. Sc. DEGREE EXAMINATION, APRIL 2022
BIOINFORMATICS
FOURTH SEMESTER

COURSE : CORE
PAPER : BIG DATA ANALYSIS
TIME : 3 HOURS **MAX. MARKS: 100**

SECTION – A

ANSWER ALL QUESTIONS **(20 X 1=20)**

Choose the correct answer:

1. Which is the correct order for medical image analysis
   a) Preprocessing, Segmentation, Thresholding, and feature detection
   b) Preprocessing, Thresholding, Segmentation, and feature detection
   c) Preprocessing, Segmentation, feature detection and Thresholding
   d) Preprocessing, feature detection, Segmentation and Thresholding
2. Handling large volumes of data involves
   a) Using correct algorithms b) Choosing the right data structures
   c) Using right tools d) All the above
3. Volume, velocity and variety are ____ to big data,
   a) Intrinsic b) Extrinsic c) Both a and b d) none of the above
4. Which is not an discretization technique?
   a) Normalization b) Histogram c) Binning d) Regression
5. _____ function is responsible for consolidating the results produced by each of the Map() functions/tasks.
   a) Reduce b) Map c) Task d) All of them

Fill in the Blanks:

6. _____ is the process of reducing a group of words into their lemma or dictionary form in NLP.
7. _____ is the python library will take care of the matrix splitting, and linear regression variable weights can be calculated using matrix calculus.
8. Variety describes one of the biggest challenges of ____.
9. PCA stands for ____.
10. ___ and ___ are the two ways to implement DFS.

Define in single line:

11. Define IoT.
12. Give an example of graph database
13. What is the five step process to structure your data analysis?
14. Define Binning.
15. List two features of DFS.

Say True or False:

16. *Dummy variables* can only take two values: true(1) or false(0).
17. Is sparse matrix also known as dense matrix?
18. Veracity makes sure that the data is accurate?
19. Is smoothing technique can be employed to remove the noise in data?
20. Blocks, name node and data node are the concepts of HDFS?

SECTION – B

ANSWER ANY FOUR QUESTIONS. EACH ANSWER SHOULD NOT EXCEED 500 WORDS. ALL QUESTIONS CARRY EQUAL MARKS. (4 x 10 = 40)

21. Comment on implementation of Network modelling in biological concepts.
22. Elaborate the data integration process briefly.
23. Explain the 3V's of big data.
24. Highlight the importance of data reduction process with PCA as example.
25. Brief about the following: a) Data Scaling, b) Data Transformation.
26. Discuss the key points in analyzing the data and communicating the results.
27. Illustrate the importance of MapReduce.

SECTION – C

ANSWER ANY TWO QUESTIONS. EACH ANSWER SHOULD NOT EXCEED 1200 WORDS. ALL QUESTIONS CARRY EQUAL MARKS. (2 x 20 = 40)

28. Illustrate the facets of big data in detail.
29. Discuss the machine generated data and their advantages
30. Enumerate the data per-processing steps in detail.
31. Comment on HDFS and YARN as resource manager.

****