Ex/IT/PC/H/T/421/2022

B.E. INFORMATION TECHNOLOGY FOURTH YEAR SECOND SEMESTER - 2022

DATA SCIENCE (HONS.)

Time: 4 Hours

Full Marks: 70

[Answer all questions]

CO1:

1. Explain about various stages in data science project? [4]
2. Explain in brief four major applications of Data science. [2]
3. Explain the method to collect and analyze data to use social media to predict the weather condition. [2]
4. What is a bias-variance trade-off? [2]

CO2:-

5. What is data Analysis? Why python is used for data analysis? [3]
6. What is dimensionality reduction? Explain the curse of dimensionality. [3]
7. What libraries do data scientists use to plot data in Python? [1]
8. Consider the confusion matrix showing the result on the performance of a classifier. [3]

	Class A	Class B
Class A	80	25
Class B	15	70

Calculate the following (clearly mention the formula for each metric): (i) Precision (ii) Recall (iii) Sensitivity.

CO3:

9. What is stopword removal and stemming? Why are these processes necessary for better information retrieval? [2+2]
10. Define the TF-IDF scheme of determining the weight of a keyword in a document. Why is it necessary to include IDF in the weight of a term? [2+2]
11. What are vocabularies in IR systems? What role do they play in the indexing of documents? [2]

CO4:

12. State different components of a time series. [4]
13. Students believe that the salary they can expect during a placement process is related to their academic performance. The CGPA (indicator of performance) and the salary obtained by six students are (7, 6), (6.8, 5.8), (7.5, 6.5), (8, 7), (8.2, 7.5) and (8.6, 8). Find the salary that a student with CGPA 8.7 can expect? [4]

14. The following table relates to the tourist arrivals during 1990 to 1996 in India: [4]

Years:	1990	1991	1992	1993	1994	1995	1996
Tourists arrivals: (in millions)	18	20	23	25	24	28	30

Fit a straight line trend by the method of least squares and estimate the number of tourists that would arrive in the year 2000.

15. Given the data 92, 93, 92, 91, 93, 94, 92 find the forecast for the eighth period using simple exponential smoothing? Use α = 0.3 and initial forecast using simple average? [4]

16. Calculate the seasonal indices from the following data using the average from the following data using the average method: [4]

	I Quarterly	II Quarterly	III Quarterly	IV Quarterly
2008	72	68	62	76
2009	78	74	78	72
2010	74	70	72	76
2011	76	74	74	72
2012	72	72	76	68

CO5:

17. What are some important features of a good data visualization? [3]
18. List the conventional Data visualization tools. Explain any two. [4]
19. Explain with diagram, the difference between count histogram, relative frequency histogram, cumulative frequency histogram and density histogram? [4]
20. What is the need of data visualization? Also explain the advantages of using data visualization. [2+2]
21. Summarize the importance of visualization in different types of data in exploration in data analysis. [3]
22. When do you use a boxplot and in what situation would you choose boxplot over histograms? [2]

-------X-------

CO1: Describe and discuss the role of Data Science.
CO2: Apply the knowledge of different Data Science tools and techniques for analyzing variety of data.
CO3: Find and interpret various kinds of information retrieval process from different type of domain.
CO4: Choose and examine different intelligent data analytic techniques.
CO5: Assess and operate the different data visualization outcomes