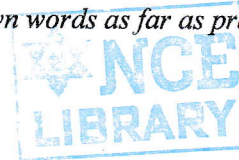


Exam.	Regular (New Course)		
Level	BE	Full Marks	60
Programme	BCT	Pass Marks	24
Year / Part	II / I	Time	3 hrs.

Subject: - Foundation of Data Science (ENCT 202)

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ Attempt All questions.
- ✓ The figures in the margin indicate Full Marks.
- ✓ Assume suitable data if necessary.



1. How is a programmer different from a data scientist? Explain the differences in terms of skill requirements and the various roles available for data scientists in the job market. [5]
2. a) Why is K-fold cross validation considered as the standard method for validating model? Explain all the terms; training data, testing data, validating data and fold and overall steps. [3]
- b) Considering the test data size as 50,000, a typical classification model experimentation results are as shown on the confusion matrix.
 - i) Calculate Accuracy, Precision and F_1 -score measure data as per table below. [3]
 - ii) Explain the difference on resulted accuracy and F_1 -score value that you have calculated. [2]

	PREDICTED CLASS	
	Yes	No
ACTUAL CLASS	Yes 1000	No 1000
	No 1000	47000

3. a) Prepare a note on Central Limit Theorem. Why this is considered as important in data analysis and other tasks? Briefly explain. [4]
- b) An e-mail filter is planned to separate valid e-mails from spam. The word free occurs in 70% of the spam messages and only 4% of the valid messages. Also, 25% of the messages are spam. Determine the following probabilities: [6]
 - (i) The message contains free. (ii) The message is spam given that it contains free.
 - (iii) The message is valid given that it does not contain free.
4. a) What is semi-structured data? How it is different than structured and unstructured data? Explain briefly. [3]
- b) How do you handle missing data during data preprocessing? Briefly explain two methods illustrating with a small sample data example. [4]
5. Here is the 10 random sample data of graduate student age from Sociology and Political Science program class.
 - a) Calculate normalized ages of 2-students (Data #1, both Women and Men) using z-score and min-max of Zero-to-one. [4]

Data#	1	2	3	4	5
Age (Women)	35	29	32	23	36
Age (Men)	28	25	29	24	23

(Show your all calculations and consider all 10 data as a single sample.)

- b) Using the same data above, draw a box-plot for depiction of all 10 student age distribution. You need to show all calculations and label the plot properly. [4]

6. a) What are the purposes of data visualization? Explain with an example of heat-map based visualization. [5]
b) What is regression, and how does it support predictive analytics? Explain with a small sample business application. [5]
7. Write brief notes on: (Any Three) [3×4]
- | | |
|------------------------|------------------------------------|
| (i) GDPR | (iii) Generative AI |
| (ii) Feature selection | (iv) Performance measure using MAE |
