

MASTER OF ENGINEERING IN INFORMATION TECHNOLOGY
FIRST YEAR SECOND SEMESTER EXAMINATION, 2025

Data Science

Time: 3 Hrs.

Full Marks: 100

CO1 - Define and illustrate the data science process (ANY ONE) (20)

1.
 - a) Discuss in detail the entire Data Science process, starting from defining research goals to building applications with a suitable diagram.
 - b) Illustrate each phase with suitable examples.(10+10)
2.
 - a) Explain the benefits and various uses of Data Science in modern applications.
 - b) Explain how Data Mining and Data Warehousing support the data science lifecycle.
 - c) Define the key forms or facets of big data and how they influence decision-making.

CO2 - Illustrate and analyze different types of data description techniques (ANY ONE) (20)

3.
 - a) Explain the different types of data and variables used in data analysis.
 - b) Discuss how data can be described using Summary Statistics.
 - c) Use examples to illustrate how techniques used in summary statistics provide meaningful summaries of datasets.
4.
 - a) Discuss the concept of correlation in quantitative data.
 - b) Explain how scatter plots and correlation coefficients are used to describe relationships between variables.
 - c) Derive and explain the computational formula for the correlation coefficient.

CO3 - Use the Python libraries for data wrangling (ANY ONE) (24)

5.
 - a) Explain the simple linear regression and multiple linear regression with a suitable example and diagram.
 - b) What is a regression line, Provide examples to support your explanation.

- c) Differentiate between loess plot and line plot with proper diagram.

(12+6+6)

6.

- Describe the features of normal, skewed (positive & negative), and bimodal distributions. Provide graphical illustrations for each.
- Explain how central tendency is useful in summarizing a dataset. Differentiate between interpolation vs extrapolation.
- Differentiate between histogram and bar plot with suitable diagram and example.

(12+6+6)

CO4 - Apply visualization libraries in Python to interpret and explore data (AS MENTIONED BELOW)

(13 + 13 = 26)

- What is ROC-AUC? Why do we need it, and when we use it? Explain with suitable example and diagram. Describe the difference between positive, negative, and no correlation with examples.
- Construct box plots from the following data and identify the outliers. Explain each step including calculation of Q1, Q3, IQR, and the fences. Discuss any limitations of box plots in data analysis.
 B1={1.26, 0.34, 0.70, 1.75, 50.57, 1.55, 0.08, 0.42, 0.50, 3.20, 0.15, 0.49, 0.95, 0.24, 1.37, 0.17, 6.98, 0.10, 0.94, 0.38}
 B2= {2.37, 2.16, 14.82, 1.73, 41.04, 0.23, 1.32, 2.91, 39.41, 0.11, 27.44, 4.51, 0.51, 4.50, 0.18, 14.68, 4.66, 1.30, 2.06, 1.19}

OR

- Describe the steps involved in creating a correlation matrix and a heat map from a given multivariate dataset. Explain how a heat map can be used to identify relationships among multiple variables in a dataset.
- Given a dataset of student scores from multiple departments, explain how side-by-side box plots can be used to compare performance across departments.

Student ID	Department	Score
S01	CS	78
S02	CS	85
S03	CS	92
S04	CS	69
S05	CS	88
S06	IT	72
S07	IT	81
S08	IT	76
S09	IT	85
S10	IT	89
S11	ECE	64
S12	ECE	70

[Turn over

S13	ECE	68
S14	ECE	74
S15	ECE	80
S16	ME	60
S17	ME	62
S18	ME	66
S19	ME	71
S20	ME	69

CO5 - Understand data privacy, ethics, importance of data ethics, Data security (ANY THREE)

(5 x 3 = 15)

11. Discuss the risks associated with handling sensitive data and suggest ways to mitigate them.
12. Describe the relationship between data privacy, data anonymization, and modern privacy risks.
13. What is the importance of maintaining data quality and integrity throughout the data lifecycle?
14. Describe some common errors that may occur in data (pre)processing or model design and their ethical implications.