

Dr. Babasaheb Ambedkar Open University
Term End Examination January – 2026

Course	: MSCDS	Date	: 18/02/2026
Subject Code	: MSCDS-103	Time	: 11:30am to 01:45pm
Subject Name	: Principles of Data Science	Duration	: 02.15 Hours
		Max. Marks	: 70

Section A

Answer the following (Attempt any three)

(30)

1. Define Data Science. Explain the Data Science Life Cycle in detail, covering all stages from business understanding to deployment.
2. What is Exploratory Data Analysis (EDA)? Explain the various techniques used for data visualization and summary statistics.
3. Explain the concept of Data Wrangling. Discuss various methods for handling missing values and removing duplicates in a dataset.
4. Describe the different types of Data Sources (Structured, Unstructured, and Semi-structured) and explain the ETL process.
5. What is Big Data? Explain the 5 V's of Big Data and discuss the role of Hadoop and Spark in the data science ecosystem.

Section B

Answer the following (Attempt any four)

(20)

1. Explain the difference between Data Science, Data Analytics, and Machine Learning.
2. Describe the use of Jupyter Notebooks in Data Science and explain its key features.
3. What is Data Normalization and Standardization? Why are they important for machine learning models?
4. Explain the role of Python as a primary language for Data Science and list its key libraries (NumPy, Pandas, Matplotlib).
5. Discuss the importance of Data Ethics and Privacy in the field of Data Science.
6. What is "Chunking" in data processing? How does it help in managing large datasets?

Section C

(10)

1. Which stage of the Data Science life cycle involves cleaning and transforming raw data?
A. Modeling B. Data Preparation C. Deployment D. Business Understanding
2. Which Python library is primarily used for data manipulation and analysis?
A. Matplotlib B. Scikit-learn C. Pandas D. TensorFlow
3. A dataset that has no fixed schema and includes text, images, and videos is called:
A. Structured B. Semi-structured C. Unstructured D. Relational
4. Scaling data to a [0, 1] range is known as:
A. Standardization B. Normalization C. Transformation D. Encoding
5. Which of the following is an example of an open-source IDE for Data Science?
A. MS Word B. Jupyter Notebook C. Adobe Reader D. Oracle
6. The "V" of Big Data that refers to the speed at which data is generated is:
A. Volume B. Variety C. Velocity D. Veracity

- 7 Which technique is used to find the middle value of a dataset?
A. Mean B. Mode C. Median D. Variance
- 8 What does ETL stand for?
A. Extract, Transform, Load B. Execute, Transfer, List
C. Edit, Trace, Locate D. Entry, Transfer, Log
- 9 Which library is used for creating static, animated, and interactive visualizations in Python?
A. NumPy B. Matplotlib C. Requests D. OS
- 10 Identifying and removing identical rows in a dataset is called:
A. Imputation B. Deduplication C. Binning D. Outlier Detection

Part – B (Do as Directed)

(10)

State whether the following statements are true or false

- 1 Data Science is only concerned with building predictive models.
- 2 Standardization adjusts data to have a mean of 0 and a standard deviation of 1.
- 3 CSV stands for Comma Separated Values.
- 4 Qualitative data is numerical and can be measured.
- 5 Exploratory Data Analysis is usually performed after the modeling stage.
- 6 A histogram is used to visualize the distribution of a single numerical variable.
- 7 Missing values should always be deleted from a dataset.
- 8 SQL is used to communicate with relational databases.
- 9 Data wrangling consumes the majority of a data scientist's time.
- 10 Machine learning is a subset of Artificial Intelligence.
