LEVEL 5

Introduction to Data Science and Big Data

Student Guide

Modification History

Version	Date	Revision Description
V1.0	March 2024	For release

Published by: NCC Education Limited, Adamson House, Towers Business Park, Wilmslow Road, Didsbury, Manchester M20 2EZ, UK.

Tel: +44 (0) 161 438 6200 Fax: +44 (0) 161 438 6240 Email: info@nccedu.com
http://www.nccedu.com

CONTENTS

1. Unit Overview and Objectives

This unit will provide students with the foundation of data science and big data, such as data types, data exploration and processing techniques. This unit will also introduce a range of well-known techniques and applications in big data processing, such as dimensionality reduction, feature engineering, machine learning, natural language processing, and visualization.

At the end of the unit, students should be able to:

1. Demonstrate a systematic and thorough understanding of the concept of Data Science and Big Data.
2. Demonstrate proficiency in data collection, design, and modelling for data processing.
3. Select appropriate techniques and tools for data pre-processing.
4. Understand the analytical techniques and software tools to effectively generate insights from data.
5. Understand the suitability of different visualisation methods for different types of Data.
6. Demonstrate a critical understanding of Data Science Ethics

2. Learning Outcomes and Assessment Criteria

Learning Outcomes The Learner will:	**Assessment Criteria** The Learner can:
1. Demonstrate a systematic and thorough understanding of the concept of Data Science and Big Data	1.1 Explain the concept of Data Science and Big Data along with the relationship among two. 1.2 Explain the components of Data science. 1.3 Understand the need and applications of Data Science and Big Data. 1.4Understand the role and responsibility of data scientist, big data expert and data analyst. 1.5 Understand and classify different types of data. 1.6 Understand the data quality and issues.
2. Demonstrate proficiency in data collection, design, and modelling for data processing.	2.1Demonstrate a systematic awareness of the theoretical foundations of data processing. 2.2 Understand the Data Science Process Life Cycle. 2.3 Identify problems and tasks involved in the life cycle of a Data Science Project 2.4 Differentiate between different techniques and tools.
3. Select appropriate techniques and tools for data pre-processing.	3.1 Clean and prepare data for analysis. 3.2 Identify abnormalities such as missing values, outliers, redundant features, etc. in data. 3.3 Model data for a purpose. 3.4 Summarise, and visualise data using suitable tools

4. Understand the analytical techniques and software tools to effectively generate insights from data.	4.1Turn raw data into insightful information. 4.2 Understand a range of data analysis techniques and models. 4.3 Identify the suitability of different data analysis techniques for different types of Data. 4.4 Make use of cutting-edge tools and technologies to analyse data. 4.5 Understand the need of model selection and evaluation. 4.6 Understand the metrics and scoring for the evaluation of selected model.
5. Understand the suitability of different visualisation methods for different types of Data	5.1 Build big data analysis solutions with analysis methods and visualisation tools. 5.2 Understand the need of businesses and present data. 5.3 Select appropriate visualisation tools and visualise aspects of the data in each dataset. 5.4 Systematically interpret and evaluate the results of data analysis solution to inform the decision-making process.
6. Demonstrate a critical understanding of Data Science Ethics	6.1 Understand the ethical challenges and concerns in data science. 6.2 Understand the ethical challenges and concerns in data science. 6.3 Explore practical solutions for implementing responsible data practices. 6.4 Recognize the importance of fostering a culture of data ethics within an organization.

3. Syllabus

Syllabus			
Topic No	Title	Proportion	Content
1	Data Science and Big Data Fundamentals	1/12 2 hours of lectures 3 hours of tutorials	• What is Data Science? • What is Big Data? • What are Data Analytics? • Data Science vs Big Data vs Data Analytics • Components of Data Science • Data Science Process/lifecycle ***Learning Outcome: 1,2***

2	Introduction to Data	1/12 2 hours of lectures 3 hours of tutorials	• What is Data? • How big is Big Data? Some examples • Sources of Data • Big Data Challenges • Data quality and issues • 5V's of Big Data • Types of Data: Structured, Unstructured, Semi Structured • Categories of Data Types: Continuous, Categorical, Text Data, Time Series, Binary. ***Learning Outcome: 1,2***
3	Understanding Data & Exploration	1/12 2 hours of lectures 3 hours of lab session	• What is Descriptive Statistic? • What is Inferential Statistic? • What is EDA? Numerical or Graphical methods • Aims of EDA? • Exploratory vs Confirmatory Data Analysis? • Numerical Methods of EDA: Central Tendency, Measurement of Variability • Graphical Methods of EDA: histogram, box plot, scatter plot, stem, and leaf plot. ***Learning Outcome: 2,3,5***
4	Data Pre-Processing I	1/12 2 hours of lectures 3 hours of lab session	• What is Data Pre-processing? • Data Pre-processing Importance? • Data Pre-processing Steps • Data Pre-processing examples? • Data Pre-processing Methods • Missing Values • Categorical Encoding • Feature engineering ***Learning Outcome: 2, 3***
5	Data Pre-Processing II	1/12 2 hours of lectures 3 hours of lab session	• Dimensionality Reduction • Outlier analysis • Class imbalance ***Learning Outcome: 2,3***
6	Data Processing I	1/12 2 hours of lectures 3 hours of lab session	• What is Data Processing? • What is AI and its History? • Difference Between AI, ML and Deep learning • Introduction to Machine Learning • Types of Machine Learning: supervised learning, unsupervised learning. ***Learning Outcome: 1,4***
7	Data Processing II	1/12 2 hours of lectures 3 hours of lab session	• Reinforcement learning • NLP • Deep learning ***Learning Outcome: 4***

8	Model Selection and Evaluation	1/12 2 hours of lectures 3 hours of lab session	• What are model selection and model evaluation? • Types of model selection • Metrics and Scoring ***Learning Outcome: 4***
9	Data Visualisation	1/12 2 hours of lectures 3 hours of lab session	• Data Visualisation • Data Visualisation Process • Advanced data visualisation and processing tools ***Learning Outcome: 4,5***
10	Business Intelligence and Tools	1/12 2 hours of lectures 3 hours of tutorials	• What is Business Intelligence? • Benefits of Business Intelligence • Types of BI tools and software • Example of BI case studies ***Learning Outcome: 1,3,4***
11	Data Science Ethical and Privacy Issues	1/12 2 hours of lectures 3 hours of tutorials	• What is Data Science Ethics? • Accountability and Governance • Data Provenance and Aggregation ***Learning Outcome: 6***
12	Unit Summary	1/12 3 hours of lectures 2 hours of tutorials	***Learning Outcome: ALL***

4. Related National Occupational Standards

The UK National Occupational Standards describe the skills that professionals are expected to demonstrate in their jobs in order to carry them out effectively. They are developed by employers and this information can be helpful in explaining the practical skills that students have covered in this unit.

Related National Occupational Standards (NOS)
Sector Subject Area: ICT Practitioners **Related NOS:** TECDT80841, TECDT80842, TECDT80851, TECIS806401, TECIS805301

5. Resources

Lecturer Guide: This guide contains notes for lecturers on the organisation of each topic, and suggested use of the resources. It also contains all of the suggested exercises and model answers.

PowerPoint Slides: These are presented for each topic for use in the lectures. They contain many examples which can be used to explain the key concepts. Handout versions of the slides are also available; it is recommended that these are distributed to students for revision purposes as it is important that students learn to take their own notes during lectures.

Student Guide: This contains the topic overviews and all of the suggested exercises. Each student will need access to this and should bring it to all of the taught hours for the unit.

5.1 Additional Hardware and Software Requirements

Hardware: Computer with internet access

Software: Python

6. Pedagogic Approach

Suggested Learning Hours						
Guided Learning Hours				Assessment	Private Study	Total
Lecture	Tutorial	Seminar	Laboratory			
25	14	-	21	43	97	200

The teacher-led time for this unit is comprised of lectures, laboratory sessions and tutorials. The breakdown of the hours is also given at the start of each topic, with 5 hours of contact time per topic.

6.1 Lectures

Lectures are designed to introduce students to each topic; PowerPoint slides are presented for use during these sessions. Students should also be encouraged to be active during this time and to discuss and/or practice the concepts covered. Lecturers should encourage active participation and field questions wherever possible.

6.2 Tutorials

Tutorials provide tasks to involve group work, investigation and independent learning for certain topics. The details of these tasks are provided in this guide and also in the Student Guide. They are also designed to deal with the questions arising from the lectures, laboratory sessions and private study sessions.

6.3 Laboratory Sessions

During these sessions, students are required to work through practical tutorials and various exercises. The details of these are provided in this guide and also in the Student Guide. Some sessions will require more support than others as well as IT resources. More detail is given in this guide.

6.4 Private Study

In addition to the taught portion of the unit, students will also be expected to undertake private study. Exercises are provided in the Student Guide for students to complete during this time. Teachers will need to set deadlines for the completion of this work. These should ideally be before the tutorial session for each topic, when Private Study Exercises are usually reviewed.

7. Assessment

This unit will be assessed by means of an examination worth 50% of total mark and an assignment worth 50% of the total mark. These assessments will cover the learning outcomes and assessment criteria given above. Sample assessments are available through the NCC Education Virtual Learning Environment (http://vle.nccedu.com/login/index.php) for your reference.

8. Further Reading List

A selection of sources of further reading around the content of this unit must be available in your Accredited Partner Centre's library. The following list provides suggestions of some suitable sources:

Primary Reference:

- Mueller, A. C. & Guido, S. (2016). Introduction to Machine Learning with Python. O′Reilly
- Grus, J. (2019). Data Science from Scratch. O'Reilly Media.
- Holmes, D. E. (2013). Big Data: A Very Short Introduction. Oxford University Press.
- Kirk, A. (2019) Data Visualization: A Handbook for Data Driven Design. United Kingdom, SAGE Publications.
- Russell, S. (2010) Artificial Intelligence: A Modern Approach, 3rd Edition. University of California at Berkeley. Pearson

Additional References

- Patterson, J. & Gibson, A. (2017). Deep Learning: A Practitioners Approach. O'Reilly Media, Inc.
- David, M. (2022). Data Science Ethics: Concepts, Techniques, and Cautionary Tales. Oxford University Press.
- Loukides, M., Mason, H., & Patil, D. (2018). Ethics and Data Science. O'Reilly Media, Inc.
- Sabherwal, R. I. Becerra-Fernandez, I. (2010). Business Intelligence: Practices, Technologies, and Management. John Wiley & Sons Inc. NJ.
- Tufte, E.R. (2001). The Visual Display of Quantitative Information. 2nd Edition. Cheshire, CT: Graphics Press.
- Turban, E., Sharda, R., Delen, E. & Dursun. A., (2011). “Decision Support and Business Intelligence Systems”. Pearson.

Topic 1: Data Science and Big Data Fundamentals

1.1 Learning Objectives

This topic provides an overview of Data Science and Big Data. The detailed introduction, need and applications of both are explained along with the role and responsibility of data scientist and big data professionals. The topic covers the different components of data science along with the classification of data and highlight the issues in data quality.

On completion of the topic, students will be able to:

- Understand the concept of Data Science and Big Data along with the relationship among two.
- Understand the need and applications of Data Science and Big Data.
- Understand the role and responsibility of data scientist, big data expert and data analyst.
- Identify the components and life cycle of Data Science.

1.2 Pedagogic Approach

Information and theory of the topic will be presented to the students during lectures. They will then practise the skills during the tutorial sessions. Students are expected to undertake their own private study to understand the theory fully and put the lectures in context.

1.3 Timings

Lectures:	2 hours
Private Study:	8 hours
Tutorials:	3 hours

1.4 Lecture Notes

The following is an outline of the material to be covered during the lecture time and should be read in conjunction with the slides provided.

The structure of this topic is as follows:

- What is Data Science?
- What is Big Data?
- What are Data Analytics?
- Data Science vs Big Data vs Data Analytics
- Components of Data Science
- Data Science Process/lifecycle

1.5 Tutorial Sessions

The time allocation for this topic is 3 hours.

I. Introduction to Data Science

- What is Data Science, and why is it important in various domains?

- What are the key components of Data Science, and how do they contribute to its success?

- Can you provide examples of real-world applications where Data Science is used?

II. Understanding Big Data

- How would you define Big Data, and what are its defining characteristics?

- What are the challenges and opportunities associated with working with Big Data?

- Can you name some popular tools and technologies used for handling Big Data?

III. Exploring Data Analytics

- What is Data Analytic, and how does it help in extracting insights from data?

- What are the different types of Data Analytics, and how do they differ from one another?

- What are some common techniques used in Data Analytics?

IV. Data Science vs. Big Data vs. Data Analytics

- How would you differentiate Data Science, Big Data, and Data Analytics in terms of their objectives and applications?

- What skills and tools are required in each of these fields?

- Can you provide examples or case studies to illustrate the use of Data Science, Big Data, and Data Analytics in different industries?

V. Data Science Process/Lifecycle

- What are the different stages involved in the Data Science process/lifecycle?

- Can you explain the tasks and challenges associated with each stage?

- How does the Data Science process lead to knowledge and actionable insights?

VI. Essential Skills and Tools in Data Science

- Why are statistics and probability important in Data Science?

- What are some commonly used programming languages in Data Science, and what makes them suitable?

- What are some machine learning techniques and algorithms used in Data Science?

- How can effective data visualization contribute to the communication of insights?

1.6 Private Study

The time allocation for private study in this topic is expected to be 8 hours.

The private study is divided into sections that cover various aspects of the data science process and related concepts. Each section focuses on a specific topic or stage, providing explanations, examples, and tasks to complete.

Stage 1: Problem Definition
- Why is problem definition important in data science?
- Task: Select a real-world problem and clearly define the problem statement and objectives.

Stage 2: Data Acquisition
- What is data acquisition, and why is it necessary for data science projects?
- Task: Identify potential data sources for the selected problem in Stage 1 and list the types of data that could be collected.

Suggested Answer:

Example 1: Customer Churn Prediction for a Telecommunications Company

Stage 3: Data Preparation
- Why is data preparation crucial in the data science process?
- Task: Given a small dataset below, clean the data by handling missing values and removing duplicates.

Dataset: Student Grades

Student ID	Name	Age	Gender	Grade
1	Alice	19	F	90
2	Bob	20	M	85
3	Charlie		M	75
4	Diana	21	F	
5	Eric	18	M	88
6	Fiona	19	F	92
2		20	M	85
8	Hannah	22	F	78
9	Isaac	23		80
10	Jack	20	M	85

Stage 4: Exploratory Data Analysis (EDA)
- What is the purpose of exploratory data analysis (EDA) in data science?
- Task: Perform basic EDA on the cleaned dataset by visualising key features (i.e., Histogram or Bar chart) and summarising statistical measures.

Stage 5: Modelling
- What is modelling in data science, and how does it help solve problems?

Stage 6: Model Evaluation
- Why is model evaluation important in the data science process?

Stage 7: Model Deployment and Maintenance
- What is model deployment, and what considerations are involved?

Topic 2: Introduction to Data

2.1 Learning Objectives

This topic provides an overview of Data. The detailed introduction on Big Data along with examples and how the data sources can be explored and what are the associated challenges. The topic covers the data quality challenges and explain the 5V's of Big Data in detail. The topic covers the different types of data along with the classification of data in different categories.

On completion of the topic, students will be able to:

- Understand the need and applications of Data Science and Big Data.
- Understand and classify different types of data.
- Understand the data quality and issues.
- Demonstrate a systematic awareness of the theoretical foundations of data processing.

2.2 Pedagogic Approach

Information and theory of the topic will be presented to the students during lectures. They will then practise the skills during the tutorial sessions. Students are expected to undertake their own private study to understand the theory fully and put the lectures in context.

2.3 Timings

Lectures:	2 hours
Private Study:	8 hours
Tutorials:	3 hours

2.4 Lecture Notes

The following is an outline of the material to be covered during the lecture time and should be read in conjunction with the slides provided.

The structure of this topic is as follows:

- What is Data?
- How big is Big Data? Some examples
- Sources of Data
- Big Data Challenges
- Data quality and issues
- 5V's of Big Data
- Types of Data: Structured, Unstructured, Semi Structured
- Categories of Data Types: Continuous, Categorical, Text Data, Time Series, Binary.

2.5 Tutorial Sessions

The time allocation for this topic is 3 hours.

1. What is data, and why is it important in today's digital world?
2. What are the different types of data? Provide examples of each type.
3. What are the characteristics of Big Data, and what sets it apart from traditional data?
4. Can you give examples of sources that contribute to Big Data? How does Big Data impact various industries?
5. What are the challenges associated with Big Data? How does volume, velocity, variety, veracity, and value contribute to these challenges?
6. Why is data quality crucial, and what are some common data quality issues?
7. How can we address data quality issues and improve the overall quality of data?
8. Explain the concept of the 5 V's of Big Data (Volume, Velocity, Variety, Veracity, Value) and their significance. How do the 5 V's apply to real-world scenarios and different industries?

2.6 Private Study

The time allocation for private study in this topic is expected to be 8 hours.

Questions:

1. What are the three main types of data? Describe the characteristics of each type (structured, unstructured, semi-structured) and provide examples.

2. What is continuous data? Give examples of continuous data and explain its significance in statistical analysis.

3. Define categorical data and provide examples of categorical variables. Discuss the importance of categorical data in classification and grouping.

4. How is text data defined, and what are its applications? Provide examples of text data and explain how it is used in natural language processing and sentiment analysis.

5. What is time series data? Give examples of time series data and discuss its use in forecasting and trend analysis.

6. Define binary data and provide examples. Explain how binary data is used to represent yes/no or true/false information.
7. Why is it important to understand different types and categories of data for effective data analysis?

8. How can knowledge of data types and categories enhance decision-making and insights in various fields?

9. **Task: Analysing Data Types**

a) Given a dataset below, identify its data type (structured, unstructured, semi-structured).

b) Categorise the data types within the dataset (continuous, categorical, text data, time series, binary).

c) Justify your classifications based on the characteristics of each data type and category.

d) Discuss potential challenges or considerations when analysing the dataset based on its data types.

Dataset: Customer Feedback Survey

Customer ID	Feedback	Age	Satisfaction Rating	Purchase Amount	Timestamp
1	Great service!	35	4	150	2023-07-01 10:15:00
2	Needs improvement	42	2	80	2023-07-02 14:30:00
3	Excellent experience	28	5	231	2023-07-03 09:45:00
4	Highly satisfied	51	5	120	2023-07-03 18:20:00
5	Slow delivery	37	3	367	2023-07-04 12:00:00

Topic 3: Understanding Data & Exploration

3.1 Learning Objectives

This topic presents Exploratory Data Analysis (EDA), utilising numerical and graphical methods to gain insights, understand data distributions, and recognize outliers. This lecture aims to clarify the aims of EDA and the distinction between exploratory and confirmatory data analysis. Students will also grasp fundamental EDA techniques, both numerical (central tendency, variability) and graphical (histograms, box plots), essential for preliminary data understanding and visualization.

On completion of the topic, students will be able to:

- Demonstrate a systematic awareness of the theoretical foundations of data processing.
- Understand and apply descriptive statistics to summarize and describe data.
- Utilise numerical and graphical methods in Exploratory Data Analysis (EDA) to gain insights from data.
- Differentiate between exploratory and confirmatory data analysis approaches and recognize their objectives.

3.2 Pedagogic Approach

Information and theory of the topic will be presented to the students during lectures. They will then practise the skills during the lab sessions. Students are expected to undertake their own private study to understand the theory fully and put the lectures in context.

3.3 Timings

Lectures:	2 hours
Private Study:	8 hours
Laboratory:	3 hours

3.4 Lecture Notes

The following is an outline of the material to be covered during the lecture time and should be read in conjunction with the slides provided.

The structure of this topic is as follows:

- What is Descriptive Statistic?
- What is Inferential Statistic?
- What is EDA? Numerical or Graphical methods
- Aims of EDA?
- Exploratory vs Confirmatory Data Analysis?
- Numerical Methods of EDA: Central Tendency, Measurement of Variability
- Graphical Methods of EDA: histogram, box plot etc.

3.5 Laboratory Sessions

The time allocation for this topic is 3 hours.

The lab for this session is divided into two parts. This is the first time; the practical sessions are going to start. Therefore, the lab 1 is just to provide them basic understand about the Google Collaboratory and how to program in python.

Lab 1: Introduction to Google Collaboratory
There is no need for answers to this lab. The step-by-step guide in included for the students. The lecturer is expected to help students if they get stuck at some point.

Lab 2: Basic coding in Python.
There is no need for answers to this lab. The step-by-step guide in included for the students to understand how to write code and comments in Python. The lecturer is expected to help students if they get stuck at some point.

The Python tutorial starts from scratch. This will be trivial for some students who have already studied programming, but the unit can be taken by other students who may not have selected the programming unit before.

Refer to the respective lab exercises and private study resources provided.

3.6 Private Study

The time allocation for private study in this topic is expected to be 8 hours.

The private study exercise is based on further enhancing students' skills in Python. All the steps required to perform the exercise are included in the study manual along with dataset. However, there are some extra tasks that student is expected to do and explore Python at their own time. The solution and corresponding Python code is already added in the folder for the lecturer to show to students in the next session.

Refer to the respective lab exercises and private study resources provided.

Topic 4: Data Pre-processing I

4.1 Learning Objectives

This topic provides an overview of data pre-processing. Students will cover the essentials of cleaning and preparing data for analysis. They will learn the significance of this process in ensuring data quality and reliable results. The lecture will detail key steps in data pre-processing, including handling missing values, encoding categorical data, and outlier detection.

On completion of the topic, students will be able to:

- Demonstrate a systematic awareness of the theoretical foundations of data processing.
- Understand and apply data pre-processing methods to clean and prepare data for analysis.
- Identify abnormalities such as missing values, encoding, feature engineering, etc. in data.
- Differentiate between different techniques and tools for data pre-processing.

4.2 Pedagogic Approach

Information and theory of the topic will be presented to the students during lectures. They will then practise the skills during the tutorial sessions. Students are expected to undertake their own private study to understand the theory fully and put the lectures in context.

4.3 Timings

Lectures:	2 hours
Private Study:	8 hours
Laboratory:	3 hours

4.4 Lecture Notes

The following is an outline of the material to be covered during the lecture time and should be read in conjunction with the slides provided.

The structure of this topic is as follows:

- What is Data Pre-processing?
- Data Pre-processing Importance?
- Data Pre-processing Steps
- Data Pre-processing examples?
- Data Pre-processing Methods
- Missing Values
- Categorical Encoding
- Feature engineering

4.5 Laboratory Sessions

The time allocation for this topic is 3 hours.

The lab for this session is consist of Outlier detection and handling missing values. The step-by-step guide is provided to students along with the code and description so that they will be able to understand in depth what is taught in the lecture. The corresponding datasets and python code is also provided for the lecturer so that if some student experience any issue lecturer will be able to see the running code and help. Lecturer can run the Python code is already added in the folder at the end of session to show students and explain if needed.

Refer to the respective lab exercises and private study resources provided.

4.6 Private Study

The time allocation for private study in this topic is expected to be 8 hours.

The private study exercise is based on further enhancing students' skills in outlier detection. All the steps required to perform the exercise are included in the study manual along with dataset (if required). However, there are some extra tasks that student is expected to do and explore detection at their own time. The solution and corresponding Python code is already added in the folder for the lecturer to show to students in the next session.

Refer to the respective lab exercises and private study resources provided.

Topic 5: Data Pre-processing II

5.1 Learning Objectives

This topic provides an overview of data pre-processing. Students will cover the essentials of cleaning and preparing data for analysis. They will learn the significance of this process in ensuring data quality and reliable results. The lecture will detail key steps in data pre-processing, including handling missing values, encoding categorical data, and outlier detection.

On completion of the topic, students will be able to:

- Demonstrate a systematic awareness of the theoretical foundations of data processing.
- Understand and apply data pre-processing methods to clean and prepare data for analysis.
- Learn techniques to reduce data, select features and handle class imbalance.
- Differentiate between techniques and tools for data pre-processing.

5.2 Pedagogic Approach

Information and theory of the topic will be presented to the students during lectures. They will then practise the skills during the tutorial sessions. Students are expected to undertake their own private study to understand the theory fully and put the lectures in context.

5.3 Timings

Lectures:	2 hours
Private Study:	8 hours
Laboratory:	3 hours

5.4 Lecture Notes

The following is an outline of the material to be covered during the lecture time and should be read in conjunction with the slides provided.

The structure of this topic is as follows:

- Dimensionality Reduction
- Feature Engineering
- Class imbalance

5.5 Laboratory Sessions

The time allocation for this topic is 3 hours.

The lab for this session is consist of Dimensionality Reduction in Python using PCA. The step-by-step guide is provided to students along with the code and description so that they will be able to understand in depth what is taught in the lecture. The corresponding datasets (link is available within the guide) and python code is also provided for the lecturer so that if some student experience any issue lecturer will be able to see the running code and help. The lecturer can run the Python code that is already added in the folder at the end of session to show students and explain if needed.

Refer to the respective lab exercises and private study resources provided.

5.6 Private Study

The time allocation for private study in this topic is expected to be 8 hours.

The private study exercise is based on further enhancing students' skills in dimensionality reduction and based on feature engineering. All the steps required to perform the exercise are included in the study manual along with dataset (if required). Moreover, an overview of different dimensionality reduction techniques is provided so that students can understand when and where to use the required techniques. Students can try different techniques to the dataset provided in their own time and explore them further.

However, there are some extra tasks that student is expected to do and explore feature engineering at their own time. The task includes implementing linear Support Vector Machines (SVM) instead of logistic regression. The solution and corresponding Python code is already added in the folder for the lecturer to show to students in the next session.

Refer to the respective lab exercises and private study resources provided.

Topic 6: Data Processing I

6.1 Learning Objectives

This topic introduces the concept of AI, tracing its historical development and its pivotal role in modern technology. Students will distinguish between AI, Machine Learning (ML), and Deep Learning, gaining insights into the varying levels of complexity in AI systems. The lecture will further delve into Machine Learning, exploring its basic principles and the distinction between supervised and unsupervised learning, providing a foundational understanding of how AI and ML are reshaping various industries through data-driven decision-making and predictive analytics.

On completion of the topic, students will be able to:

- Demonstrate a systematic awareness of the theoretical foundations of data processing.
- Understand the need of data processing and machine learning.
- Differentiate between different terms such as machine learning and Artificial Intelligence.
- Understand the basic categories of machine learning techniques.

6.2 Pedagogic Approach

Information and theory of the topic will be presented to the students during lectures. They will then practise the skills during the tutorial sessions. Students are expected to undertake their own private study to understand the theory fully and put the lectures in context.

6.3 Timings

Lectures:	2 hours
Private Study:	8 hours
Laboratory:	3 hours

6.4 Lecture Notes

The following is an outline of the material to be covered during the lecture time. Please also refer to the slides.

The structure of this topic is as follows:

- What is Data Processing?
- What is AI and its History?
- Difference Between AI, ML and Deep learning
- Introduction to Machine Learning
    - Types of Machine Learning: supervised learning, unsupervised learning

6.5 Laboratory Sessions

The time allowance for tutorials in this topic is 3 hours.

The lab for this session is consist of ML in Python using K-NN algorithm. The step-by-step guide is provided to students along with the code and description so that they will be able to understand in depth what is taught in the lecture. The corresponding datasets (link is available within the guide if needed) and python code is also provided for the lecturer so that if some student experience any issue lecturer will be able to see the running code and help. Lecturer can run the Python code is already added in the folder at the end of session to show students and explain if needed. Students are advised to run the algorithm with different K values and observe the results. Encourage student to also explore other ML algorithms.

Refer to the respective lab exercises and private study resources provided.

6.6 Private Study

The time allocation for private study in this topic is expected to be 8.5 hours.

The private study exercise is based on further enhancing students' skills in ML and based on Random Forest Classifier algorithm implementation in Python. All the steps required to perform the exercise are included in the study manual along with dataset (if required). The solution and corresponding Python code is already added in the folder for the lecturer to show to students in the next session. There is some portion in the manual on model evaluation, however, you can mention to students that it will be covered in next lecture. It has already mentioned in the guide as well.

Refer to the respective lab exercises and private study resources provided.

Topic 7: Data Processing II

7.1 Learning Objectives

This topic introduces the cutting-edge fields of Reinforcement Learning, Natural Language Processing (NLP), and Deep Learning. They will explore the concept of Reinforcement Learning, understanding how agents learn to make decisions by interacting with their environments. The lecture will also introduce NLP, highlighting its role in enabling computers to understand and generate human language, with applications in chatbots and language translation. Deep Learning, a subset of machine learning, will be discussed in depth, emphasising its use in training artificial neural networks for tasks like image and speech recognition.

On completion of the topic, students will be able to:

- Demonstrate a systematic awareness of the theoretical foundations of data processing.
- Understand the need of data processing and machine learning.
- Understand the basics of learning techniques such as reinforcement learning, natural language processing, deep learning.

7.2 Pedagogic Approach

Information will be transmitted to the students during the lectures. They will then practise the skills during the laboratory sessions and extend their understanding during private study time. The tutorial will then provide an opportunity to review the key ideas and obtain further guidance and support.

7.3 Timings

Lectures:	2 hours
Private Study:	8 hours
Laboratory:	3 hours

7.4 Lecture Notes

The following is an outline of the material to be covered during the lecture time. Please also refer to the slides.

The structure of this topic is as follows:

- Reinforcement learning
- Natural language processing (NLP)
- Deep learning

7.5 Laboratory Sessions

The time allowance for tutorials on this topic is 3 hours.

In this lab session, students will learn the basics of reinforcement learning using Python. The step-by-step guide is provided to students along with the code and description so that they will be able to understand in depth what is taught in the lecture.

The corresponding datasets (link is available within the guide if needed) and python code are also provided for the lecturer, so that if students experience any issues, the lecturer will be able to see the running code and help. The lecturer can run the Python code that is already added in the folder at the end of session to show students and explain if needed. Students are advised to learn more about the Taxi-V3 scenario to fully understand the code.

Refer to the respective lab exercises and private study resources provided.

7.6 Private Study

The time allocation for private study in this topic is expected to be 8 hours.

The private study exercise is based on further enhancing students' skills in ML and students will learn the basics of NLP using Python. They will perform text preprocessing, text classification, and sentiment analysis on a dataset of movie reviews.

All the steps required to perform the exercise are included in the study manual along with dataset (if required). The solution and corresponding Python code is already added in the folder for the lecturer to show to students in the next session.

Refer to the respective lab exercises and private study resources provided.

Topic 8: Model Selection and Evaluation

8.1 Learning Objectives

This topic provides an overview on model selection and evaluation, students will explore the significance of choosing the right machine learning models and assessing their performance. They will learn about various evaluation techniques, including cross-validation along with the use of performance metrics like accuracy, precision, and recall gauging model effectiveness. This lecture equips students with the essential skills for informed model decision-making in machine learning applications.

On completion of the topic, students will be able to:

- Understand the need of model selection and evaluation.
- Understand the metrics and scoring for the evaluation of selected model.

8.2 Pedagogic Approach

Information will be transmitted to the students during the lecture. They will then practise the skills during the laboratory sessions and extend their understanding during private study time. The tutorial will then provide an opportunity to review the key ideas and obtain further guidance and support.

8.3 Timings

Lecture:	2 hours
Private Study:	8 hours
Laboratory:	3 hours

8.4 Lecture Notes

The following is an outline of the material to be covered during the lecture time. Please also refer to the slides.

The structure of this topic is as follows:

- What are model selection and model evaluation?
- Types of model evaluation
- Metrics and Scoring

8.5 Laboratory Sessions

The time allowance for tutorials in this topic is 3 hours.

In this lab session, students will learn how to evaluate the performance of machine learning models for both classification and regression tasks using Python. The step-by-step guide is provided to students along with the code and description so that they will be able to understand in depth what is taught in the lecture. The corresponding datasets (link is available within the guide if needed) and python code is also provided for the lecturer so that if some students experience any issue lecturer will be able to see the running code and help. Lecturer can run the Python code is already added in the folder at the end of session to show students and explain if needed.

Refer to the respective lab exercises and private study resources provided.

8.6 Private Study

The time allocation for private study in this topic is expected to be 8 hours.

The private study exercise is based on Training a machine learning Walkthrough example. The complete step-by-step guide is provided to help students to go through this. There is no code provided as all the code is available within the guide. The lecturer can help student if they get stuck but at this point in this module, encourage students to explore online resources to address the issues.

Refer to the respective lab exercises and private study resources provided.

Topic 9: Data Visualization

9.1 Learning Objectives

This topic provides an overview to students to learn the process of presenting data clearly and effectively. They'll explore advanced data visualization tools like Python libraries, enabling them to create impactful visuals for informed decision-making. This equips students with the skills to transform complex data into compelling, actionable insights.

On completion of the topic, students will be able to:

- Understand the need of data visualization.
- Understand the analytical techniques and software tools to effectively generate insights from data.
- Understand the suitability of different visualization methods for different types of Data.

9.2 Pedagogic Approach

Information will be transmitted to the students during the lectures. The tutorial will then provide an opportunity to review the key ideas and obtain further guidance and support.

9.3 Timings

Lectures:	2 hours
Private Study:	8 hours
Laboratory:	3 hours

9.4 Lecture Notes

The following is an outline of the material to be covered during the lecture time. Please also refer to the slides.

The structure of this topic is as follows:

- Data Visualization
- Data Visualization Process
- Advance data visualization and processing tools

9.5 Laboratory Sessions

The time allocation for this topic is 3 hours.

In this lab session, students will learn Data Visualization with Matplotlib. The step-by-step guide is provided to students along with the code and description so that they will be able to understand in depth what is taught in the lecture.

The corresponding datasets (link is available within the guide if needed) and python code is also provided for the lecturer so that if some students experience any issues the lecturer will be able to see the running code and help. The lecturer can run the Python code is already added in the folder at the end of session to show students and explain if needed.

Refer to the respective lab exercises and private study resources provided.

9.6 Private Study

The time allocation for private study in this topic is expected to be 8 hours.

The private study exercise is based on further enhancing students' skills in visualization and students will learn data visualization on real datasets in Python. All the steps required to perform the exercise are included in the study manual along with dataset (if required). The solution and corresponding Python code is already added in the folder for the lecturer to show to students in the next session.

Refer to the respective lab exercises and private study resources provided.

Topic 10: Business Intelligence and Tools

10.1 Learning Objectives

This topic provides an overview about the concept of Business Intelligence (BI) and its pivotal role in harnessing data for strategic insights. The lecture will highlight the numerous benefits of BI, including improved decision-making, efficiency, and competitiveness. Students will also explore various types of BI tools and software, gaining insight into the diversity of options available. Real-world BI case studies will be presented to demonstrate how organizations have leveraged data to achieve their objectives, providing valuable insights into the practical applications of BI in modern business environments.

On completion of the topic, students will be able to:

- Understand the need for computerised decision making in organisations.
- Understand the concept of business intelligence and its importance.
- Understand the suitability of different methods and tools for Business Intelligence.
- Gain insight into the practical use of BI.

10.2 Pedagogic Approach

Information will be transmitted to the students during the lectures. They will then practise the skills during the laboratory sessions and extend their understanding during private study time. The tutorial will then provide an opportunity to review the key ideas and obtain further guidance and support.

10.3 Timings

Lectures:	2 hours
Private Study:	8 hours
Tutorials:	3 hours

10.4 Lecture Notes

The following is an outline of the material to be covered during the lecture time. Please also refer to the slides.

The structure of this topic is as follows:

- Evolution of Business Intelligence
- What is Business Intelligence?
- Benefits of Business Intelligence
- Types of BI tools and software
- Example of BI case studies

10.5 Tutorial Sessions

The tutorial time allocation for this topic is 3 hours.

- Divide the students into small groups.
- Instruct groups to brainstorm and list potential benefits of implementing BI in the given case study.
- Each group presents their findings and discusses the unique advantages of BI in different sectors.

Case Study: Improving Sales Forecasting with BI Analytics

Industry: Retail

Background:
A large retail chain with numerous stores and a diverse product range was facing challenges with sales forecasting. Inaccurate sales predictions resulted in issues such as overstocking, leading to increased carrying costs, and understocking, causing lost sales and customer dissatisfaction. To address these issues, the company decided to leverage BI analytics to enhance its sales forecasting capabilities.

Objectives:
1. Improve the accuracy of sales forecasts for individual stores and product categories.
2. Reduce overstock and understock situations.
3. Optimize inventory management.
4. Enhance overall supply chain efficiency.

Solution:

Data Integration: The company integrated a variety of data sources, including historical sales data, inventory levels, seasonal patterns, and external factors like economic indicators and weather data. These sources were consolidated into a data warehouse.

Data Analysis and Modelling: BI analysts and data scientists used advanced analytics and machine learning algorithms to analyze the data and create predictive models. These models considered historical sales trends, seasonal fluctuations, local market conditions, and external factors.

Dashboard Creation: The BI team developed user-friendly dashboards that provided real-time insights into sales forecasts. The dashboards displayed:
- Sales forecasts for each store and product category.
- Inventory levels compared to forecasts.
- Alerts for potential understock and overstock situations.

Alerting System: An automated alerting system was implemented to notify inventory managers when forecasts indicated potential issues, allowing for proactive inventory adjustments.

Conclusion:

This BI case study illustrates how the strategic implementation of BI analytics can lead to substantial improvements in sales forecasting, inventory management, and overall supply chain efficiency. By leveraging data-driven insights, the retail company was able to optimize operations, reduce costs, and enhance the customer experience.

Task 1: List potential benefits of implementing BI

Task 2: Share a list of common types of BI tools (e.g., reporting tools, data visualization tools, self-service BI platforms) with brief descriptions. Available in Lecture Slides
Ask students to match each type of tool with a relevant use case or scenario where it is most effective. Facilitate a class discussion to share and compare their answers.

Task 3: Identify the problem the organization faced, the solution implemented, and the outcomes achieved. Encourage a class discussion on the key takeaways from the case study.

10.6 Private Study

The time allocation for private study in this topic is expected to be 8 hours.

This private study aims to deepen your understanding of Business Intelligence (BI) by analysing a real-world case study. You will be required to identify key factors, outcomes, and lessons learned from the BI implementation. It is expected that each student chooses the case study based on the interest and write a report of 1000 words with the details mentioned in the study guide. There is no specific answer to this as it just to provide student with deeper understanding of the topic. Lecturer can ask students to submit the report and provide them with overall feedback during the next session.

Instructions:

Step 1: Case Study Selection
1. Choose one BI case study from a reliable source related to your area of interest (e.g., industry or field).
2. Ensure that the case study is sufficiently detailed and includes information on the problem, solution, and outcomes.

Step 2: Case Study Analysis
Pay close attention to the following aspects:

- Problem or challenge faced by the organization.
- Solution or BI implementation undertaken.
- Outcomes and impact of the BI solution.

Step 3: Prepare a Report (approx. 1000 words)

1. Write a concise report that includes the following sections:
    - Introduction: Briefly introduce the organization and the context of the case study.
    - Problem Statement: Describe the specific challenge or problem the organization faced that led to the need for a BI solution.
    - BI Solution: Explain the BI implementation or solution adopted by the organization.
    - Outcomes and Impact: Discuss the results and outcomes of the BI implementation. Highlight key successes and benefits.
    - Key Takeaways: Identify any lessons learned or best practices that can be derived from the case study.

Additional Notes:

- Ensure that you provide proper citation and references for the case study you analyze.
- Feel free to use external sources to complement your analysis, but the primary focus should be on the chosen case study.

This private study will allow you to delve deeper into a real-world BI case study, analyse the application of BI concepts, and gain insights.

Topic 11: Data Science Ethical and Privacy Issues

11.1 Learning Objectives

This topic begins with the introduction of data ethics and privacy. It will explore topics such as accountability and governance, understanding the ethical principles that guide data-driven actions and the importance of transparency and compliance with regulations. Additionally, the lecture will cover data provenance and aggregation, emphasizing the need to track data sources and ensure that the aggregation process does not compromise privacy or fairness.

On completion of the topic, students will be able to:

- Understand the ethical challenges and concerns in data science.
- Explain strategies and techniques to mitigate ethical risks in data science.
- Explore practical solutions for implementing responsible data practices.
- Recognize the importance of fostering a culture of data ethics within an organization.

11.2 Pedagogic Approach

Information will be transmitted to the students during the lectures. They will then practise the skills during the laboratory sessions and extend their understanding during private study time. The tutorial will then provide an opportunity to review the key ideas and obtain further guidance and support.

11.3 Timings

Lectures:	2 hours
Private Study:	8 hours
Tutorials:	3 hours

11.4 Lecture Notes

The following is an outline of the material to be covered during the lecture time. Please also refer to the slides.

The structure of this topic is as follows:

- What is Data Science Ethics?
- Accountability and Governance
- Data Provenance and Aggregation

11.5 Tutorial Sessions

The time allowance for tutorials in this topic is 3 hours.

Q1. What is Data Science Ethics, and why is it important in the field of data science? Provide examples of ethical dilemmas in data science.

Q2. Discuss the concept of data privacy in data science. What are the key principles and considerations when handling personal data? Provide a real-world example where privacy concerns were significant.

Q3. How can data scientists ensure fairness in machine learning algorithms? Explain the challenges and methods to mitigate bias in algorithmic decision-making.

Q4. What is transparency in data science, and why is it crucial? How can organizations promote transparency in their data practices? Provide an example of an organization that benefited from transparent data handling.

11.6 Private Study

The time allocation for private study in this topic is expected to be 8 hours.

Accountability and Governance:

1. Define accountability in data science and explain why it is vital. Provide an example of a situation where clear accountability would have made a difference.

2. What is data governance, and what are its primary objectives? Discuss the role of data stewards in data governance.

3. Explain how data governance frameworks, like DAMA-DMBOK, help organizations establish best practices for data governance. Provide an overview of the key components of such frameworks.

4. How can an organization ensure regulatory compliance in data handling, such as with HIPAA in healthcare? Discuss the responsibilities of a Chief Data Officer (CDO) in maintaining governance and compliance.

Data Provenance and Aggregation:

1. Define data provenance and explain why it is essential in data science. How does tracking data origins contribute to data quality and reliability?

2. Discuss the challenges associated with data aggregation, such as data granularity and accuracy. Provide an example where data aggregation led to inaccurate results.

3. How do tools and techniques like metadata help in tracking data provenance? Describe a scenario where metadata played a crucial role in maintaining data integrity.

4. Provide a real-world use case where data provenance and aggregation are critical, such as in scientific research or financial reporting. Explain how they ensure the reliability of the data.

Answer 4: In scientific research, genomic data analysis relies on data provenance to track the sources and processing of genetic information. This ensures the accuracy of research findings and supports reproducibility. In financial reporting, stock market data aggregation requires precise tracking of sources and transformations to produce reliable market analysis reports, impacting investment decisions.

Topic 12: Unit Summary

12.1 Learning Objectives

This topic provides the summary of this Introduction to Data Science and Big Data unit. It recaps all the important concepts cover in this unit.

On completion of the topic, students will be able to:

- Demonstrate a systematic understanding of the concept of Data Science and Big Data
- Demonstrate proficiency in data collection, design, and modelling for data processing.
- Select appropriate techniques and tools for data pre-processing.
- Understand the analytical techniques and software tools to effectively generate insights from data.
- Understand the suitability of different visualization methods for different types of Data.
- Demonstrate a critical understanding of Data Science Ethics

12.2 Pedagogic Approach

Information and theory of the topic will be presented to the students during lectures. They will then practise the skills during the tutorial sessions. Students are expected to undertake their own private study to understand the theory fully and put the lectures in context.

12.3 Timings

Lectures:	3 hours
Tutorials:	2 hours
Private Study:	9 hours

12.4 Lecture Notes

The following is an outline of the material to be covered during the lecture time and should be read in conjunction with the slides provided.

The structure of this topic is as follows:

- Data Science and Big Data Fundamentals
- Introduction to Data
- Understanding Data & Exploration
- Data Pre-Processing
- Data Processing
- Model Selection and Evaluation
- Data Visualization
- Business Intelligence and Tools
- Data Science Ethical and Privacy Issues

12.5 Tutorial Sessions

The time allowance for tutorials in this topic is 2 hours.

This session is for support to student to address any questions related to lab or tutorial session and can also introduce assessment or provide some guidance on exam.

12.6 Private Study

The time allocation for private study in this topic is expected to be 9 hours.

The private study time is used for the preparation of exams or to complete any lab or tutorial session.