College of Contract Management
United Kingdom

Professional Diploma in Data Science

Syllabus

Unit 1: Introduction to Data Science and AI

Unit code:	**DS101**
Level:	**7**
Credit value:	
Guided learning hours:	**20**

Unit aim

To provide participants with a foundational understanding of the fields of Data Science and Artificial Intelligence. By the end of the module, attendees should have a clear grasp of the evolution, significance, and applications of Data Science and AI in various industries. Additionally, they will be introduced to the fundamental differences and intersections between Data Science and AI and become familiar with the critical tools and technologies driving these domains. This foundational knowledge will serve as a platform for more advanced topics in subsequent modules.

Unit introduction

Data Science and AI have become transformative forces in our quickly digitising world, altering sectors, rethinking company strategies, and reimagining customer experiences. The applications span a wide range of industries, from manufacturing to healthcare to finance. But first, it's important to have a fundamental comprehension of what complicated algorithms, predictive models, and neural networks are and how they work.

Learning outcomes and assessment crtieria

To pass this unit, the evidence that the learner presents for assessment needs to demonstrate that they can meet all the learning outcomes for the unit. The assessment criteria determine the standard required to achieve the level of proficiency.

On completion of this unit, a learner should:

Learning outcomes	Assessment criteria
1. Understand the historical context of data science and artificial intelligence.	1.1 Describe the evolution and pivotal moments in the history of Data Science and AI
2. Differentiate between data science and artificial intelligence.	2.1 Clearly distinguish between the core concepts, techniques, and objectives of Data Science and AI 2.2 Identify areas of intersection and collaboration between the two fields
3. Recognise data types and structures.	3.1 Classify various data types and explain their relevance in Data Science and AI projects 3.2 Understand the significance of structured and unstructured data in different industry applications
4. Appreciate real-world applications.	4.1 Enumerate key industries where Data Science and AI have made a significant impact 4.2 Cite real-world case studies showcasing the transformative power of these technologies
5. Identify key tools and technologies.	5.1 List and describe essential software, platforms, and technologies used in Data Science and AI 5.2 Understand the significance of choosing the right tool for specific tasks
6. Acknowledge industry barriers.	5.1 Understand the challenges industries face in adopting Data Science and AI 5.2 Identify potential strategies to mitigate these challenges

Unit content

1. Overview and Evolution

We will begin by taking a retrospective look at how data science and AI have emerged over the decades, understanding their roots, pivotal moments, and trajectory to modern-day significance. The "overview" offers a snapshot of the current state, defining key concepts and illuminating the field's primary applications. In contrast, the "evolution" traces the journey of the discipline from its inception to the present, marking pivotal moments, groundbreaking discoveries, and paradigm shifts. Together, they provide a holistic understanding, painting both a macroscopic picture of the field's purpose and potential as well as a chronological narrative of its development over time.

2. Differentiating Data Science and AI

It's common for many to use data science and 'AI' interchangeably. This module will delineate the unique aspects of each field while also highlighting where they intersect.

Initially, we'll delve into data science, highlighting its primary goal of extracting meaningful insights from vast and varied datasets. Participants will understand its multidisciplinary nature, encompassing statistics, data analysis, and visualisation techniques. Real-world applications, like business analytics and predictive modelling, will underscore its practical relevance.

Transitioning to artificial intelligence (AI), we'll explore its overarching aim: enabling machines to mimic human-like intelligence. This section will introduce subsets of AI, notably machine learning, emphasising how it allows systems to learn from data and make decisions.

Crucially, the module will spotlight intersections between the two and how data science methodologies often serve as the bedrock for training AI models. Yet, it will also emphasise their distinct objectives, with data science primarily concerned with discovering patterns and AI with decision-making. By the module's conclusion, participants should not only grasp the nuances distinguishing these fields but also appreciate how they collaboratively drive modern innovations.

3. Data Structures & Their Importance

Data is often called the 'oil' of the digital age. We'll explore different data types and structures, emphasising why a robust understanding of these is vital for anyone looking to delve into data science or AI.

We'll kick off with an introduction to basic data structures like arrays, linked lists, stacks, and queues, illustrating their characteristics and operations. Progressing to more complex structures, participants will explore trees, graphs, and hash tables, understanding their relevance in various computational problems.

A vital segment will address the distinction between linear and non-linear structures, elaborating on their respective use cases. Additionally, the module will touch upon dynamic data structures, which grow and shrink during runtime, emphasising their flexibility.

Underpinning the theoretical knowledge will be real-world applications showcasing how these structures drive efficiency in algorithms. For instance, how trees, specifically binary search trees, facilitate faster data retrieval or how hash tables optimise storage and lookup operations

The crux of the module is to impress upon participants that choosing the appropriate data structure is paramount. An apt choice can dramatically improve the speed, readability, and overall efficacy of software solutions, whereas a poor choice can lead to inefficient and cumbersome programmes. By the end, attendees should appreciate data structures not just as theoretical constructs but as pivotal tools in effective problem-solving.

4. Real-world Applications

We will take a cursory look at how different industries are leveraging Data Science and AI. These case studies will give you a taste of the transformative power of these technologies.

Participants will embark on a journey exploring a myriad of sectors. In the field of healthcare, researchers and practitioners are exploring how artificial intelligence (AI) might contribute to various aspects of the industry. Specifically, AI has shown promise in supporting diagnostic processes, forecasting patient outcomes, and optimising administrative operations. When exploring the field of finance, individuals will observe the significant impact of artificial intelligence in several areas such as fraud detection, algorithmic trading, and personalised financial services.

The retail industry is expected to demonstrate advancements in inventory management, consumer analytics, and chatbot technology, all powered by artificial intelligence, hence increasing the whole shopping experience. Within the field of transportation, the focal point of discourse often revolves around autonomous vehicles, the optimisation of routes, and the implementation of predictive maintenance strategies.

Each industry-specific exploration will not only highlight the technologies in use but also the transformative impact they've ushered. Case studies will be employed, offering a granular look into successful implementations, challenges encountered, and the solutions devised.

5. Exploration of Tools & Technologies

No professional can function effectively without the right tools. This section introduces the software, platforms, and technologies pivotal in the realms of Data Science and AI.

Initiating with Data Manipulation and Analysis, participants will learn about tools like Pandas and NumPy that offer robust capabilities for handling large datasets. The segment on Visualization will introduce platforms such as Matplotlib and Tableau, illuminating how they transform raw data into insightful visuals.

Diving into Machine Learning, we'll cover frameworks like TensorFlow and Scikit-learn, showcasing their roles in creating and tuning predictive models. The Deep Learning section will touch upon tools such as Keras and PyTorch, elucidating their significance in constructing neural networks.

The module will also shed light on Big Data Technologies like Hadoop and Spark, crucial for processing vast data streams efficiently. Additionally, an overview of Cloud Platforms such as AWS, Azure, and Google Cloud will emphasize their role in scalable AI solutions.

Integral to this module are hands-on demos, ensuring participants not only understand the theoretical aspects but also witness these tools in action. By the module's end, attendees should be well-versed in the technological landscape of AI and Data Science, and equipped with knowledge to select and implement appropriate tools for specific tasks.

6. Barriers to Adoption

Lastly, understanding challenges is as crucial as understanding capabilities. We'll touch upon common barriers industries face while adopting these technologies.

Technical Challenges will headline the discussion, focusing on issues like data quality, integration complexities, and the need for robust infrastructure. The intricacies of migrating from traditional systems to AI-driven ones, scalability concerns, and the ever-evolving tech landscape requiring continuous upskilling will be emphasized.

Under Organisational Barriers, the module will explore resistance to change, misalignment between business and IT units, and the scarcity of skilled professionals in the AI domain. We'll also delve into the high initial costs associated with AI adoption and the struggle to demonstrate immediate ROI.

Ethical and Regulatory Challenges Will examine the presence of biases within AI models, address privacy concerns, and explore the ethical ramifications associated with automation. The evolving nature of AI regulation, compliance issues, and the need for transparent AI will be highlighted.

Lastly, Societal Barriers will examine public perception, fears of job displacement due to automation, and the broader societal implications of widespread AI adoption.

By exploring these hurdles, the module aims to equip professionals with a comprehensive understanding of potential pitfalls, encouraging proactive strategies for smoother AI integration and ensuring ethical, transparent, and beneficial deployments.

Why this module is important

To truly grasp the more intricate aspects of Data Science and AI, one must first have a comprehensive understanding of their foundational elements. This module aims to ensure that every participant, whether a novice or someone with some prior exposure, gains clear, cohesive, and consistent foundational knowledge.

Delivery

Tutors must keep in mind the diverse backgrounds of industry professionals and because this module serves as their gateway into a rapidly evolving field, tutors must ensure the content remains accessible, relatable, and engaging. I have explained these in more detail, and these shall serve as guiding principles during the preparation of any course materials and delivery of lectures.

Relatability

The tutor must anchor theoretical concepts with real-world applications. Case studies can be an invaluable tool in this regard, offering tangible illustrations of abstract ideas.

Interactivity

The tutor must encourage participation throughout this course to participate fully. The brainstorming sessions, interactive visualizations, and case discussions are designed to foster an environment of active learning. Prompting questions and discussions can make abstract concepts more tangible.

Practicality

The hands-on exercises aim to consolidate theoretical knowledge. Tutors will ensure participants get hands-on exposure, even if rudimentary. The act of ‘doing’ often reinforces understanding.

Feedback

Tutors must be proactive in providing feedback, especially during assignments and hands-on sessions. Constructive feedback can bridge gaps in understanding and reinforce concepts.

Adaptability

Tutors must recognize the pace of the class. While the module has a defined structure, feel free to spend more time on concepts that participants find challenging.

When considering learning outcome 1 i.e. “Understand the Historical Context of Data Science and Artificial Intelligence,” it’s essential for learners to the foundational theories and initial advancements that gave birth to AI and Data Science. Recognizing seminal figures like Alan Turing or foundational concepts such as the Turing Test can give learners an appreciation for the field’s roots.

Learners must be able to Identify major milestones in the evolution of AI and Data Science. This could include the advent of neural networks, the AI winter periods, or the resurgence of AI with deep learning. They must also recognize how advancements in computing power, storage, and data availability have driven progress in AI and Data Science and how various fields, including statistics, computer science, cognitive psychology, and even biology, converged to shape AI and Data Science’s trajectory. Learners must be able to grasp the societal impacts i.e., how AI and Data Science have historically influenced industries, economies, and broader society. This encompasses both the positive transformations and the challenges or ethical dilemmas they introduced.

Tutors must provide the right perspective and industry-based examples so that learners develop an appreciation for how past challenges, failures, and successes have informed current best practices, ethical considerations, and the direction of future research.

For learners, grasping these elements not only provides a comprehensive understanding of where AI and Data Science have been but also offers context on their current state and future potential.

Learning plan

Topic and suggested assignments/activities and/assessment
Introduction to the unit and programme of learning
Tutor-led lecture on **Overview of Data Science** supplemented by brainstorming session on real-world applications of Data Science
Tutor-led lecture on the **Basics of Artificial Intelligence** followed by an assignment on “Research and summarise a pivotal moment in AI’s evolution”
Tutor-led lecture on **How Data Science fuels AI and the synergy between them** followed by a case study discussion on a real-world example using both AI and Data Science
Tutor-led lecture on **Introduction to data sources, data quality, and preprocessing techniques** followed by an interactive session using simple tools to visualize basic algorithms in action
Tutor-led lecture on **Basic Algorithms and Models** followed by a hands-on exercise on data cleaning using sample dataset
Tutor-led lecture on the **Exploration of industries revolutionized by Data Science and AI** followed by an assignment to pick a real-time use case or industry and detail its AI transformation in the last few years
Towards the end of the module, there will be a **Group Project** in which participants will team up to identify a real-world problem and propose a solution utilizing the basics of Data Science and AI, culminating in a ***PowerPoint presentation***
A multiple choice exam of 30 minutes covering foundational concepts of the module.
Participants will write a brief ~200 words **reflection** on the importance of Data Science and AI in modern society, based on their learnings from the module

Learning expectations & assessment

For learners, differentiating between Data Science and AI ensures they can accurately navigate discussions, make informed decisions in their respective roles, and appreciate where interdisciplinary collaboration can be most impactful. It also helps in avoiding common misconceptions and ensures clarity when approaching projects, research, or further studies in either domain

Upon completing the “**Overview and Evolution**” section, learners are expected to grasp the chronological progression of Data Science and AI, identifying key milestones and figures. They should be able to understand pivotal technological and conceptual shifts that have shaped the fields. And that they recognize the impact of AI and Data Science on society and industries throughout history.

For the section: “**Differentiating Data Science and AI**”, learners are expected to discern the distinct characteristics, methodologies, and objectives of both Data Science and AI. They should be equipped to identify the primary tools and applications unique to each field, while also recognizing their overlapping domains, notably in areas like machine learning. Furthermore, learners should be able to engage in informed discussions, confidently differentiating between the two disciplines and understanding their interconnected roles in the broader landscape of technology and data-driven decision-making.

While working on the section “**Data Structures & Their Importance**”, learners will recognize the significance of selecting the appropriate data structure for various tasks and understanding the implications for efficiency, data storage, and algorithm performance. With this foundation, learners are positioned to better analyse, store, and manipulate data, optimizing the efficacy of subsequent data-driven processes and AI computations.

Upon completing the “**Real-world Applications**”, learners are expected to understand the tangible impacts and implementations of Data Science and AI in diverse industries and societal contexts. They should be able to identify specific instances where these technologies have brought about transformative changes, addressing real-world challenges and enhancing various facets of daily life.

With the “**Exploration of Tools & Technologies**” section, learners will have a good view of key tools, platforms, and technologies that underpin the worlds of Data Science and AI. They should possess a foundational knowledge of how to select, deploy, and navigate these tools effectively in various contexts, understanding their strengths, limitations, and appropriate use cases.

Upon finishing the section: “**Barriers to Adoption**”, learners are expected to recognize and understand the challenges and obstacles hindering the widespread adoption of Data Science and AI in various sectors. This encompasses technical constraints, ethical considerations, regulatory concerns, and societal apprehensions.

Programme of suggested assignments

The assessment criteria directly reference the learning outcomes and expectations from each section of the module. The assignments are crafted to align closely with these criteria, ensuring learners have a comprehensive understanding and ability to apply their knowledge in real-world contexts. This is for guidance only and learners might not be expected to produce all of these.

Assessment criteria covered	Assignment title	Scenario	Assessment method
Historical progression of Data Science and AI	Historical Timeline Project	As a data scientist in the making, trace the history of AI and Data Science using a timeline.	Visual Presentation
Characteristics and methodologies of Data Science and AI	Compare and Contrast	Draft a report distinguishing between AI and Data Science based on a chosen industry scenario.	Written Report
Fundamental data structures & their roles	Data Structures Workbook	Utilise an array, list, tree, and graph to solve specific data challenges in a company setup.	Practical Workbook/ Code
Impacts and implementations in diverse industries	Real-world Case Study Analysis	Analyse a real-life industry scenario where AI and Data Science have been pivotal.	Written Report & Discussion
Key tools, platforms, and technologies	Tool Exploration Lab	Simulate a business problem and apply a solution using a chosen Data Science/AI tool.	Hands-on Lab & Reflection
Challenges in Adopting Data Science & AI	Barrier Analysis Presentation	Discuss the potential barriers to AI integration in a traditional industry.	Oral Presentation & Q/A

Resources

Text Books

“Data Science for Business” by Foster Provost and Tom Fawcett
ISBN: 978-1449361322
This book introduces the core principles of Data Science and its implications for business scenarios.

“Artificial Intelligence: A Modern Approach” by Stuart Russell and Peter Norvig
ISBN: 978-0136042594
A comprehensive textbook on AI, discussing its history, methodologies, and prospects.

“Python for Data Analysis” by Wes McKinney
ISBN: 978-1491957660
A guide to using Python for data wrangling and analysis, crucial for many Data Science tasks.

“The Hundred-Page Machine Learning Book” by Andriy Burkov
ISBN: 978-1790368384
A concise and beginner-friendly overview of machine learning principles and techniques.

“Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy” by Cathy O’Neil
ISBN: 978-0553418811
A discussion on the ethical and societal challenges posed by unchecked algorithmic decisions.

Articles (with citations)

“Artificial intelligence: A powerful paradigm for scientific research” on National Library of Medicine, Yongjun Xu. (2021)

“The Role of ChatGPT in Data Science”, Hossein Hassani, (2023)
How AI-Assisted Conversational Interfaces Are Revolutionizing the Field

“Real-world applications of artificial intelligence (AI) in business” on IBM’s blog, Watson, J. (2018). Real-world applications of AI in business. IBM Blogs.

“The Barriers To Using AI Effectively” on Harvard Business Review, Bughin, J., & Hazan, E. (2017). The Barriers to Using AI Effectively. Harvard Business Review.

Unit 2: Data Management and Preprocessing

Unit code:	**DS102**
Level:	**7**
Credit value:	
Guided learning hours:	**20**

Unit aim

To equip participants with the essential skills and knowledge required for effective data management and preprocessing. By the end of this module, attendees should be proficient in various data acquisition techniques, understand the importance of data cleaning, and be able to transform and preprocess data for further analysis. This unit will serve as a foundation for subsequent modules that delve deeper into data analysis and machine learning.

Unit introduction

Data management and preprocessing are the foundational steps in any data science or AI project. Before any meaningful analysis can be conducted, data must be gathered, cleaned, and transformed into a suitable format. This unit delves into the various techniques and tools used in these crucial initial stages, ensuring that participants are well-equipped to handle real-world data challenges.

Learning outcomes and assessment crtieria

To successfully complete this unit, learners must demonstrate that they can meet all the learning outcomes for the unit. The assessment criteria determine the standard required to achieve each outcome.

On completion of this unit, a learner should:

Learning outcomes	Assessment criteria
1. Understand various data acquisition techniques.	1.1 Describe different sources of data, including databases, APIs, and web scraping 1.2 Understand the challenges associated with gathering data from various sources 1.3 Discuss the importance of data quality and its impact on analysis 1.4 Evaluate the ethical considerations when acquiring data 1.5 Understand Master data management and impact on organisations
2. Master data cleaning and preprocessing techniques.	2.1 Identifying common issues in datasets, such as missing values, outliers, and inconsistencies 2.2 Apply various techniques to handle missing data 2.3 Understand the importance of data consistency and standardisation 3.4 Use tools and software to automate the data cleaning process
3. Grasp the art of feature engineering.	3.1 Define what features are and their significance in data analysis 3.2 Extract meaning features from raw data 3.3 Transform features to enhance their predictive power 3.4 Evaluate the importance of feature selection in model performance
4. Understand data transformation techniques.	3.1 Differentiate between scaling, normalisation, and standardisation 3.2 Apply various transformation techniques to prepare data for machine learning 3.3 Evaluate the impact of data transformation on model performance

5. Master data visualisation and Exploratory Data Analysis (EDA).	5.1 Describe the importance of EDA in understanding datasets 5.2 Use various visualisation tools to explore data distributions, correlations, and patterns 5.3 Interpret visualisations to derive meaningful insights from the data 5.4 Apply statistical techniques to further understand datasets
6. Data Governance	6.1 Define and explain the core principles and objectives of data governance 6.2 Design and implement effectice data governance strategies tailored to specific organisational needs 6.3 Navigate the ethical considerations and regulatory requirements associated with data governance 6.4 Recognise the role of data governance in the broader landscape of data management

Unit content

1. Overview of Data Management

We'll dive into the world of data management, exploring its key aspects and understanding its importance. At its core, data management is about organizing, storing, and taking care of data. In today's digital age, where we deal with vast amounts of information, having a system to manage all this data is crucial. This "overview" will shed light on the main ideas, tools, and practices that make data management work. From databases that store information to best practices that ensure data is safe and easy to access, we'll get a clear picture of how it all comes together to help businesses, researchers, and everyday users make the most of their data.

2. Data Acquisition Techniques

Data acquisition is like gathering puzzle pieces for a bigger picture. It's all about collecting the right information from various sources. While businesses traditionally relied on manual data entry or spreadsheets, technology has introduced more efficient methods. Web scraping, for example, extracts specific details from websites, providing insights into online brand mentions. APIs, or Application Programming Interfaces, allow seamless data retrieval from digital platforms, offering real-time updates like weather or stock market trends. Regardless of the method, it's vital to prioritize data quality and ethical collection. This ensures that the resulting insights are both accurate and trustworthy, paving the way for informed business decisions.

Participants will gain an understanding of the diverse methods used to collect data in today's digital age. They'll learn the significance of web scraping and the power of APIs in real-time data retrieval. Moreover, they'll grasp the importance of data quality, ensuring that the information they work with is both accurate and ethically sourced.

3. Data Acquisition Techniques

In the vast realm of data, it's often the case that not every piece of information is immediately ready for analysis. This is where the essential practices of data cleaning and preprocessing come into play. On the other hand, preprocessing sets the stage for insightful analysis, transforming data into a format that algorithms and models can easily interpret. Participants will gain a deep understanding of the nuances of data refinement. They'll learn the art of spotting inconsistencies, the science behind data transformations, and the importance of ensuring that data is in its prime state for analysis.

For hands-on experience, we'll delve into key tools and libraries that have become industry standards in this domain. Pandas, a Python library, stands out as a versatile tool for a myriad of data manipulations, from handling gaps in data to intricate transformations. OpenRefine offers a dedicated environment for data cleaning, allowing users to dive deep into large datasets, pinpoint inconsistencies, and apply batch corrections. Scikit-learn, beyond its renowned machine learning capabilities, is a treasure trove of utilities for data preprocessing, be it scaling, normalization, or encoding. And, of course, Excel remains a staple, showcasing that even familiar spreadsheet tools can play a pivotal role in data refinement.

4. Data Transformation

Data transformation is like language translation. Data is transformed to be suited for analysis or a different data format, just as we translate French to English for a different audience. This step is crucial to data management and preprocessing, ensuring data is in the correct format, scale, and structure for analysis, visualisation, and modelling.

Data transformation involves formatting raw data for analytical methods. This may require scaling numerical data, encoding category variables, or rearranging the data. A dataset may be converted from a wide format, where each variable has a column, to a long format, where variables are stacked vertically. Data preparation for some analysis or visualisations requires such changes.

Participants will learn about data transformation and associated approaches. Normalisation and standardisation, which give data a mean of zero and a standard deviation of one, will be covered. Logarithmic and power transformations, which stabilise variances and normalise data, will also be covered.

Tools and libraries that enable these transformations will be explored hands-on. Scikit-learn in Python provides powerful data scaling and encoding tools. In data management, SQL's query language may modify data. Tableau and Power BI allow on-the-fly data modifications during visualisation.

5. Exploration of Tools & Technologies

Data transformation is like language translation. Data is transformed to be suited for analysis or a different data format, just as we translate French to English for a different audience. This step is crucial to data management and preprocessing, ensuring data is in the correct format, scale, and structure for analysis, visualisation, and modelling.

Data transformation involves formatting Tools and technology guide professionals through the complex processes of data handling, transformation, and analysis in data management and preprocessing. This section discusses the several tools' capabilities, strengths, and best uses.

Specialised data tools have proliferated to solve specific problems. They range from powerful databases like SQL and NoSQL that store massive amounts of data to data wrangling tools like Pandas and OpenRefine that clean and transform data to visualisation platforms like Tableau and PowerBI that visualise data through charts and graphs.

Participants will explore these tools and learn their practical uses. They'll learn SQL and relational databases. Python modules like Pandas, Matplotlib, and Seaborn will be used for data manipulation and visualisation. Hadoop and Spark, which process massive datasets, will be discussed.
Beyond the tools, this section emphasises job-specific tool selection. Understanding tool strengths and

weaknesses is essential because not all tools are fit for all tasks. Participants will assess their data needs, project size, and difficulties to determine the best tools and technologies. By the end of this section, learners will have a good knowledge to use these technologies for data management and preparation.

6. Data Governance

Data governance protects an organization's data quality, integrity, and security. Data management concepts, practises, and methods guarantee data is valued, protected, and used properly. In this section, we'll examine data governance's importance, components, and issues.

Data governance involves defining data management roles, policies, and standards. It involves keeping data accurate, consistent, and accessible while preventing misuse. Define who can access, use, and maintain data quality. Enabling data-driven decision-making and meeting regulations and ethics is a delicate balance.
Data governance principles lik
e data stewardship, quality, and security will be covered. They'll learn how data stewards and custodians manage data quality and use. Data catalogues that centralise data assets and data lineage tools that trace data's path through systems will be covered.

Data ethics and regulation will be a focus. Understanding compliance is vital with GDPR and CCPA creating strict data protection rules. Participants will learn about these regulations' problems and how organisations comply.

Creating a data-driven culture will also be stressed. Data governance must be integrated into the company's culture and valued by all stakeholders. This part will empower learners to champion data governance in their organisations and ensure that data is not only handled but also valued as a vital asset.

Why this module is important

The digital age has ushered in an era where data is often referred to as the 'new oil.' Every day, vast amounts of data are generated, collected, and stored. However, raw data, in its unprocessed form, can be messy, inconsistent, and riddled with errors. Before any meaningful insights can be derived from this data, it must undergo a series of preprocessing steps to ensure its quality, consistency, and relevance.

Foundation for Decision Making

The integrity of any data-driven decision or analysis hinges on the quality of the data it's based upon. Data management and preprocessing ensure that the data fed into analytical models is clean, relevant, and devoid of errors. Without these foundational steps, subsequent analyses, no matter how sophisticated, risk being flawed or misleading.

Efficiency and Cost Saving

Poor data management can lead to inefficiencies, duplicated efforts, and costly mistakes. By ensuring data is well-managed and pre-processed, organizations can streamline their operations, reduce redundancies, and avoid costly errors that arise from basing decisions on poor-quality data.

Enhancing Data's Predictive Power

Properly pre-processed data can significantly enhance the predictive power of machine learning and AI models. Features engineered from clean data can capture intricate patterns, leading to more accurate predictions and insights.

Regulatory and Ethical Compliance

With the rise of data privacy regulations such as GDPR and CCPA, proper data management is not just a best practice—it's a legal requirement. Ensuring data is correctly acquired, stored, and processed is crucial for regulatory compliance. Moreover, ethical considerations, such as fairness and bias in AI, are intrinsically linked to how data is managed and pre-processed.

Building Trust

In an era where data breaches and misuse are common headlines, proper data management can help organizations build trust with their stakeholders. Ensuring data integrity and transparency in its processing can foster trust among customers, partners, and the public at large.

Future-Proofing

As the volume, variety, and velocity of data continue to grow, the challenges associated with managing and preprocessing this data will only intensify. Mastering the skills covered in this module ensures that participants are well-equipped to handle the data challenges of today and the future.

In essence, this module is not just about technical skills; it's about understanding the pivotal role that data management plays in the broader data science and AI landscape. Proper data management and preprocessing are the unsung heroes behind every successful data-driven initiative, ensuring that the insights derived are accurate, meaningful, and actionable.

When considering the learning outcome for Unit 2, i.e., "Understand the Fundamentals of Data Management and Preprocessing," it's paramount for learners to comprehend the bedrock principles and practices that underpin effective data handling. Recognizing the significance of clean, organized data or foundational techniques such as data normalization can give learners a deep appreciation for the meticulous processes behind robust data analysis.

Learners must be equipped to identify key stages in the data management pipeline, from acquisition and cleaning to transformation and storage. This includes understanding the challenges of dealing with missing or inconsistent data, the importance of data integrity, and the tools available for preprocessing tasks. They should also be aware of how advancements in technology, from cloud storage solutions to automated data cleaning tools, have revolutionized the way we manage and preprocess data. Furthermore, the interdisciplinary nature of data management, drawing from fields like computer science, statistics, and business intelligence, should be highlighted, illustrating how these diverse areas converge to shape best practices in data management.

Tutors have the responsibility to offer real-world, industry-relevant examples, ensuring that learners not only grasp the theoretical aspects but also understand their practical implications. By showcasing how poor data management can lead to flawed insights or how effective preprocessing can enhance the accuracy of predictive models, tutors can instil in learners the critical importance of this foundational step in the data lifecycle.

For learners, mastering these concepts is not just about understanding the mechanics of data management and preprocessing. It's about recognizing their pivotal role in ensuring that data-driven decisions, whether in business, research, or technology, are based on reliable, high-quality data. This understanding not only equips them with the skills to manage and preprocess data effectively but also provides a lens through which they can view the broader implications of these processes on the outcomes of data-driven projects.

Learning expectations & assessment

This “Learning Expectations & Assessment” section for Unit 2 provides learners with a clear roadmap of the knowledge and skills they are expected to acquire throughout the module. It ensures a comprehensive understanding of the intricacies of data management and preprocessing, setting the stage for subsequent units in the course.

Upon diving into **Overview of Data Management**, learners are expected to grasp the foundational principles of data management. They should understand the importance of organized, clean data and recognize the challenges and solutions associated with handling vast amounts of information in various formats.

After learning about the **Data Acquisition Techniques**, learners should be familiar with various methods of data collection, from traditional databases to real-time data streams. They should appreciate the nuances of different data sources and understand the challenges and benefits associated with each.

By completing the section; **Data Cleaning and Preprocessing**, learners are expected to recognize the significance of refining raw data. They should understand common data imperfections, from missing values to outliers, and be equipped with techniques to address these issues. The importance of transforming data into a format suitable for analysis, ensuring its quality and integrity, should be clear to them.

Data Transformation techniques will teach learners to understand the processes involved in scaling, normalizing, and encoding data. They should appreciate the importance of these transformations in ensuring data is in the optimal format for analysis and modeling.

After learning the key **Tools & Technologies for Data Management**, learners will be able to use a range of tools and platforms pivotal in data management and preprocessing. They should be able to navigate these tools effectively, understanding when and how to deploy them based on the task at hand.

Upon concluding this segment; **Challenges in Data Management**, learners should be aware of the potential pitfalls and challenges in data management, from technical constraints to ethical dilemmas. They should be equipped to navigate these challenges, ensuring data integrity and ethical considerations are always at the forefront.

Programme of suggested assignments

The assessment criteria are directly tied to the learning outcomes and expectations from each section of the module. These assignments are designed to align closely with these criteria, ensuring that learners not only understand the concepts but can also apply them in real-world contexts. While this is a suggested list, learners might not be expected to complete all of these assignments.

Assessment criteria covered	Assignment title	Scenario	Assessment method
Overview of data management	Data Management Blueprint	As an aspiring data manager, create a blueprint outlining the key components of effective data management.	Visual Presentation
Data acquisition techniques	Data Source Exploration	Research and document various data acquisition techniques, highlighting their pros and cons in different scenarios.	Written Report
Data cleaning and preprocessing	Assignment Title: Data Cleaning Workshop	Scenario: Given a raw dataset, apply various cleaning and preprocessing techniques to prepare it for analysis.	Practical Workbook/ Code
Feature engineering	Feature Crafting Lab	Using a sample dataset, design and implement new features to enhance the data's potential for modeling.	Hands-on Lab & Reflection
Data transformation	Transformation Technique Analysis	Discuss and demonstrate various data transformation techniques, emphasising their importance in different data analysis scenarios.	Written Report & Discussion
Tools & Technologies for Data Management	Tool Dive-In Session	Explore a chosen data management tool, demonstrating its functionalities and applications in a simulated business problem.	Hands-on Lab & Reflection

Resources

Text Books

“Data Wrangling with Python: Tips and Tools to Make Your Life Easier” by Jacqueline Kazil and Katharine Jarmul
ISBN: 978-1491948811
Description: This book offers a comprehensive guide to the process of converting raw data into a format suitable for analysis, using Python.

“Data Management for Researchers: Organize, Maintain and Share Your Data for Research Success” by Kristin Briney
ISBN: 978-1784270117
Description: A practical guide for researchers on how to manage, organize, and clean their data, ensuring its integrity and usability.

“Data Preprocessing in Data Mining” by Salvador García, Julián Luengo, and Francisco Herrera
ISBN: 978-3319102460
Description: This book delves into the techniques and methods used in data preprocessing, a crucial step in the data mining process

“Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists” by Alice Zheng and Amanda Casari
ISBN: 978-1491953242
Description: A guide to the art of crafting data features, and enhancing the performance of machine learning models.

“Data-Intensive Text Processing with MapReduce” by Jimmy Lin and Chris Dyer
ISBN: 978-1608453429
Description: This book introduces the concepts of data-intensive processing using the MapReduce framework, essential for handling vast datasets.

“Data Cleaning: A Practical Perspective” by Ihab Ilyas and Xu Chu
ISBN: 978-0367331583
Description: A discussion on the challenges of data cleaning and the methodologies to address them, ensuring data quality and reliability.

Articles (with citations)

“Data Preprocessing and its Significance in Data Mining”
This article delves into the process of transforming raw data into an understandable format. It emphasizes the importance of data preprocessing as a crucial step in data mining
Published on Jan 24, 2023
https://www.tableau.com/learn/articles/what-is-data-cleaning

“The Art and Science of Data Wrangling”
Source: Towards Data Science, Author: Smith, L., (2022)
Description: An in-depth exploration of the challenges and techniques associated with transforming raw data into a usable format.
https://www.cc.gatech.edu/classes/AY2020/cs7643_spring/slides/L11_DataWranglingGATechLectureFeb2020.pdf

“Data Acquisition Systems - Current and future trends”
E.T. Subramaniam, B.P. Ajith Kumar, and R.K. Bhowmik, (2010)
Description: A comprehensive look at the various methods of data collection and the nuances of sourcing data in today’s digital landscape.
http://www.sympnp.org/proceedings/55/I17.pdf

"Data Transformation Techniques"
Dimitris Vogiatzis, Coupler, (2022)
In this guide, we'll go through all the different data transformation techniques you can use to ensure your dataset is clean and ready to be analyzed.
https://blog.coupler.io/data-transformation-techniques/

Unit 3: Practical Data Science

Unit code:	**DS103**
Level:	**7**
Credit value:	
Guided learning hours:	**16**

Unit aim

This module aims to give students practical experience of applying the data science techniques learned in previous modules to a real-world project.

Unit introduction

In previous modules, students will have learned about a variety of data science techniques and algorithms. In this unit they will apply what they have learned, undertaking a practical data science project relevant to their work with real-world data.

The students will gain experience of all stages of a data science project, from initial analysis of the data set, through data preparation, model selection and assessment to final delivery. At each stage of the project, they will be guided by a thorough understanding of the data and the business goals of the project.

The unit will be taught tutorial-style, with the lecturer guiding students though their projects, discussing any issues that they have encountered, and encouraging discussion amongst them.

Learning outcomes and assessment crtieria

In order to pass this unit, the evidence that the learner presents for assessment needs to demonstrate that they can meet all the learning outcomes for the unit. The assessment criteria determine the standard required to achieve the unit.

On completion of this unit, a learner should:

Learning outcomes	Assessment criteria
1. Be able to analyse a dataset and understand the issues involved in using it to solve the chosen problem.	1.1 Understand the statistical properties of the variables and the correlations between them 1.2 Identify any missing or corrupt data and apply appropriate remedies 1.3 Identify any potential sources of bias or other ethical issues in the data and apply appropriate remedies 1.4 Report on how these factors will affect the choice of model
2. Be able to select an appropriate model for the problem, based on the properties of the dataset used and the business requirements.	2.1 Identify a set of potentially suitable candidate models 2.2 Identify important hyperparameters of candidate models which may affect performance of models 2.3 Identify a suitable set of metrics and other criteria with which to assess cadidate models 2.4 Identify the most suitable model and hyperparameters
3. Be able to produce a working demonstration of the chosen model.	3.1 Identify best methods of presenting chosen model's outcomes to users 3.2 Create a demonstration of the chosen system 3.3 Implement a working model 3.4 Report on the performance of the demonstration system and the steps necessary to develop a production model

Unit content

As this is a project-based module, it will be taught tutorial-style. The project will be divided into 3 phases.

The first phase will concern data analysis and preparation, and will consist of two lectures, followed by a report.
In the first lecture, students will introduce themselves, and the lecturer will give a general outline of the course. The students will then outline their proposed projects. The lecturer will then brief the students on the requirements of the data analysis and preparation phase, before leading a discussion of these requirements in the context of the students' projects.
The second lecture will be a guided discussion of the progress the students have made in their analysis and preparation of the data and the issues they have encountered. They will be encouraged to raise questions with the lecturer and suggest ideas to each other. After this, the students will submit their first report, detailing the key features of their data, any data quality issues they have discovered and the remedies applied, and the consequences of their findings for the next phase of the project.

The second phase will concern model selection, and will consist of three lectures, followed by a report.
In the first lecture, the lecturer will outline key points to take into account when selecting and assessing models. The students will discuss these points in the context of their project requirements and their findings from phase one.
In the second lecture, the students will discuss their candidate models and assessment strategies with the lecturer and each other. Further advice will be given on testing strategies.
The third lecture will be a guided discussion of the progress the students have made in their assessment of their candidate models and the issues they have encountered so far. After this, the students will submit their second report, detailing their findings from this phase of the work.

If, at any stage in the first two phases, a student discovers that their initial project goals are likely to be intractable, they should revise their project goals accordingly, raise the matter in the lecture if possible, and document the revision to their project goals in the report for that phase.

The third phase will concern the development of a demonstration system, and will consist of three lectures, followed by the final report.
In the first lecture, the lecturer will discuss factors to take into account when building the demonstration system, and options for how to implement the system. The students will discuss these points in the context of their projects. While full-scale productionization of the model is outside the scope of the module, the lecturer will make students aware of the considerations this may involve, and they should address these in their final reports.
The second lecture will be a guided discussion of the progress students have made in the implementation of their demonstrations and any issues they have encountered. They will be encouraged to raise questions with the lecturer and suggest ideas to each other.
The third lecture will be a show-and-tell session, where the students will present their demos and discuss their findings and any remaining issues. After this, they will submit their final reports, detailing the decisions made in creating their demonstration, the performance of the demonstration, and the issues that would need to be addressed to develop a full production version. If a student is unable to produce a working demonstration at the end of the project, credit will be given for showing understanding of the reasons why it was not possible, and how they could be addressed in future projects.

Essential guidance for tutors

Since this module is project-based, the tutor's role is to guide and facilitate the students' learning. The tutor will introduce each phase of the project, ensure the students understand the learning goals, and lead the students' discussion of their work. The tutor will answer questions raised by the students but also encourage them to discuss

Outline learning plan

The outline learning plan has been included in this unit as guidance and can be used in conjunction with the programme of suggested assignments.

The outline learning plan demonstrates one way in planning the delivery and assessment of this unit.

Topic and suggested assignments/activities and/assessment
Introduction to unit and programme of learning. Discussion of project proposals
Tutor-led discussion on requirements for data analysis and preparation
Guided discussion of issues encountered in data analysis and preparation
Assignment 1: Data analysis and preparation
Tutor-led discussion of requirements for model selection
Tutor-led discussion of candidate models and assessment strategies
Tutor-led discussion of issues encountered when testing candidate models
Assignment 2: Model Selection
Tutor-led discussion of requirements for demonstration systems
Tutor-led discussion of issues encountered in developing demonstration systems
Student demonstrations and discussions
Assignment 3: Demonstration system

Assessment

Students should wherever possible select a project and a dataset relevant to their own work. However, if for any reason they cannot do this, an acceptable alternative would be to base their project on a publicly-available dataset. For each assignment, students should submit a written report and the code they have written in the course of the assignment.

For Learning Outcome 1, students should show that they have gained a thorough understanding of their dataset, and have identified any data quality issues and applied appropriate remedies where necessary. They should be able to use this knowledge to inform the selection of models in phase 2.

For Learning Outcome 2, students should show that they can identify suitable candidate models for their problem, taking into account their findings from phase 1. They should understand the hyperparameters of the candidate models, and be able to identify suitable metrics and other criteria to assess them. They should be able to rigorously select the most suitable model based on these criteria.

For Learning Outcome 3, students should be able to produce a working demonstration of their chosen system. They should take into account the needs of potential users in designing the demonstration. They should be able to quantify the performance of the demonstration system, and show an understanding of the issues that would be encountered in developing it further as a production system. If students are unable to produce a working demonstration, they should show a thorough understanding of why this was not possible.

Programme of suggested assignments

The table below shows a programme of suggested assignments that cover the criteria in the assessment grid. This is for guidance only and it is recommended that centres either write their own assignments or adapt Pearson assignments to meet local needs and resources.

Assessment criteria covered	Assignment title	Scenario	Assessment method
AC 1.1, 1.2, 1.3, 1.4	Data Analysis and Preparation	Students understand their datasets, have addressed any data quality issues, and can use their understanding to inform further work.	Report and code
AC 2.1, 2.2, 2.3, 2.4	Model Selection	Students select the best model for their task based on appropriate criteria.	Report and code
AC 3.1, 3.2, 3.3, 3.4	Demonstration System	Students demonstrate a working data science system and understand its performance and the issues that would be encountered in producing a production system.	Demonstration, report and code

Resources

The students should use open-source Python data science libraries in writing the code necessary for this module, and should consult the online documentation for these libraries.