OTHM LEVEL 7 DIPLOMA IN DATA SCIENCE

Qualification Number: 610/2153/2

Specification | OCTOBER 2025

TABLE OF CONTENTS

QUALIFICATION OBJECTIVES

The principal objective of the OTHM Level 7 Diploma in Data Science qualification is to develop the knowledge and skills needed to construct the means for extracting and communicating business-focused insights from data. They will develop an understanding of how value and information flows in a business, and the ability to use that understanding to identify business opportunities.

Learners will become competent and reflective data science practitioners, understanding the benefits of their newfound skills in a wide variety of contexts, enabling them to tackle more challenging roles in the future.

OTHM has developed a suite of Level 7 Diploma qualifications. The qualifications provide learners with industry-specific and practical skills, enabling them to successfully apply their knowledge in the workplace, enhance their career prospects, and allow progression to further study.

Key features of OTHM Level 7 qualifications;

- A stimulating and challenging programme of study that will be both engaging and informative for learners.
- Learners can gain the essential subject knowledge needed to progress successfully into further study or the world of work.
- Refreshed content that is closely aligned with employer and higher education needs,
- Assessments that consider cognitive skills along with affective and applied skills.
- Learners will develop knowledge and academic study skills including active research skills, effective writing skills, analytical skills, creative problem-solving, decision-making, and digital literacy.

Upon successfully completing the OTHM Level 7 Diploma in Data Science learners will be able to;

- Develop a strong understanding of data science, including the data analytics pipeline, and descriptive, predictive, and prescriptive analytics.
- Develop sufficient skill in appropriate programming languages, tools, and libraries, to use them to successfully carry out data analysis at an advanced level for a variety of data types and modalities.
- Gain the mathematical and statistical knowledge and understanding required to carry out informed basic and advanced data analysis.
- Develop an understanding of fundamental, classical, and modern data science approaches, and the evolving landscape of data science in the context of Artificial Intelligence.
- Become familiar with and make use of modern tools and techniques for data visualisation.
- Understand current and future ethical, social, and legal issues that concern data scientists.

QUALITY, STANDARDS AND RECOGNITIONS

OTHM Qualifications are approved and regulated by Ofqual (Office of Qualifications and Examinations Regulation). Visit the Register of Regulated Qualifications.

OTHM has progression arrangements with several UK universities that acknowledges the ability of learners after studying level 7 qualifications to be considered for advanced entry into corresponding Master's programmes.

REGULATORY INFORMATION

Qualification Title	OTHM Level 7 Diploma in Data Science
Qualification Ref. Number	610/2153/2
Regulation Start Date	8/2/2023
Operational Start Date	13/2/2023
Duration	1 Year
Total Credit Value	120 Credits
Total Qualification Time (TQT)	1200 Hours
Guided Leaning Hours (GLH)	600 Hours
Sector Subject Area (SSA)	15.3 - Business management
Overall Grading Type	Pass / Fail
Assessment Methods	Coursework
Language of Assessment	English

EQUIVALENCES

OTHM level 7 diplomas are located on the Regulated Qualifications Framework (RQF) and are recognised as being at the same level as Master's degrees. However, they are shorter (120 credits) qualifications which means learners will have to proceed to the dissertation stage (60 credits) with an appropriate university to achieve a full Master's qualification.

QUALIFICATION STRUCTURE

The OTHM Level 7 Diploma in Data Science consists of 6 mandatory units for a combined total of 120 credits, 1200 hours Total Qualification Time (TQT) and 600 Guided Learning Hours (GLH) for the completed qualification.

Unit Ref.	Mandatory units	Credits	GLH	TQT
F/650/5562	Data Science Foundations	20	100	200
H/650/5563	Probability and Statistics for Data Analysis	20	100	200
K/650/5565	Data Analysis and Visualisation	20	100	200
J/650/5564	Advanced Predictive Modelling	20	100	200
J/650/5573	Data Mining, Machine Learning and Artificial Intelligence	20	100	200
L/650/5566	Advanced Computing Research Methods	20	100	200

DEFINITIONS

Total Qualification Time (TQT) is the number of notional hours which represents an estimate of the total amount of time that could reasonably be expected to be required in order for a Learner to achieve and demonstrate the achievement of the level of attainment necessary for the award of a qualification.

Total Qualification Time is comprised of the following two elements –

a) *the number of hours which an awarding organisation has assigned to a qualification for Guided Learning, and*
b) *an estimate of the number of hours a Learner will reasonably be likely to spend in preparation, study or any other form of participation in education or training, including assessment, which takes place as directed by – but, unlike Guided Learning, not under the Immediate Guidance or Supervision of – a lecturer, supervisor, tutor or other appropriate provider of education or training.*

(Ofqual 15/5775 September 2015)

Guided Learning Hours (GLH) is defined as the hours that a teacher, lecturer or other member of staff is available to provide immediate teaching support or supervision to a learner working towards a qualification.

Credit value is defined as being the number of credits that may be awarded to a Learner for the successful achievement of the learning outcomes of a unit. One credit is equal to 10 hours of TQT.

ENTRY REQUIREMENTS

For entry onto the OTHM Level 7 Diploma in Data Science qualification, learners must possess:

- An honours degree in related subject or UK level 6 diploma or an equivalent overseas qualification
- Mature learners with management experience (learners must check with the delivery centre regarding this experience prior to registering for the programme)
- Learner must be 21 years old or older at the beginning of the course

English requirements: If a learner is not from a majority English-speaking country must provide evidence of English language competency. For more information visit English Language Expectations page on our website www.othm.org.uk.

Alternative professional qualifications with at least three years' relevant work experience in the public service field may also be considered. This could be in roles in local or national government, or in non-governmental and inter-governmental organisations, the voluntary and charitable sector, and private sector roles which support or deliver public services.

PROGRESSION

The OTHM Level 7 Diploma in Data Science enables learners to progress into or within employment and/or continue their further study.

As this qualification is approved and regulated by Ofqual (Office of the Qualifications and Examinations Regulation), learners maybe eligible to progress to Master's top-up at many universities in the UK and overseas with advanced standing. For more information visit the University Progressions page on the OTHM website.

DELIVERY OF OTHM QUALIFICATIONS

OTHM do not specify the mode of delivery for its qualifications, therefore OTHM Centres are free to deliver this qualification using any mode of delivery that meets the needs of their Learners. However, OTHM Centres should consider the Learners' complete learning experience when designing the delivery of programmes.

OTHM Centres must ensure that the chosen mode of delivery does not unlawfully or unfairly discriminate, whether directly or indirectly, and that equality of opportunity is promoted. Where it is reasonable and practicable to do so, it will take steps to address identified inequalities or barriers that may arise.

Guided Learning Hours (GLH) which are listed in each unit gives the Centres the number of hours of teacher-supervised or direct study time likely to be required to teach that unit.

The qualification has been designed to take learners on a structured learning pathway. The sequencing of units is likely to encourage proactive engagement due to the nature of the subjects and topics therein, whilst also supporting learners to develop the learning and assessment skills required to be successful at level 7.

CENTRE RESOURCE REQUIREMENTS

Tutor / Assessor Requirements

- Tutors/Assessors must be appropriately qualified and occupationally competent in the areas in which they are training.
- They must hold a Level 6 qualification or equivalent
- They should hold or be working towards a Level 3 qualification in Assessing Vocationally Related Achievement such as the OTHM Level 3 Award in Assessing Vocationally Related Achievement.

Internal Verifier Requirements

- Internal quality assurers or verifiers must be appropriately qualified and occupationally competent in the areas in which they are moderating.
- They must hold or be working towards a Level 4 Award in the Internal Quality Assurance of Assessment Processes and Practice and/or a Level 4 Certificate in Leading the Internal Quality Assurance of Assessment Processes and Practice such as the OTHM Level 4 Certificate in Leading the Internal Quality Assurance of Assessment Processes and Practice.
- They must demonstrate that they have undertaken Continued Professional Development (CPD) activities relating to occupational health and safety or auditing quality assurance to maintain and update their skills and knowledge within the last year.

OTHM will request to see copies of relevant qualifications from assessors and verifiers.

ASSESSMENT AND VERIFICATION

The units in this qualification are internally assessed by the centre and externally verified by OTHM. The qualifications are criterion referenced, based on the achievement of all the specified learning outcomes.

To achieve a 'pass' for a unit, learners must provide evidence to demonstrate that they have fulfilled all the learning outcomes and meet the standards specified by all assessment criteria. Judgement that the learners have successfully fulfilled the assessment criteria is made by the Assessor.

The Assessor should provide an audit trail showing how the judgement of the learners' overall achievement has been arrived at.

Specific assessment guidance and relevant marking criteria for each unit are made available in the Assignment Brief document. These are made available to centres immediately after registration of one or more learners.

OPPORTUNITIES FOR LEARNERS TO PASS

Centres are responsible for managing learners who have not achieved a Pass for the qualification having completed the assessment. However, OTHM expects at a minimum that centres must have in place a clear feedback mechanism to learners by which they can effectively retrain the learner in all the areas required before re-assessing the learner.

RECOGNITION OF PRIOR LEARNING AND ACHIEVEMENT

Recognition of Prior Learning (RPL) is a method of assessment that considers whether learners can demonstrate that they can meet the assessment requirements for a unit through knowledge, understanding or skills they already possess and do not need to develop through a course of learning.

RPL policies and procedures have been developed over time, which has led to the use of a number of terms to describe the process. Among the most common are:

- Accreditation of Prior Learning (APL)
- Accreditation of Prior Experiential Learning (APEL)
- Accreditation of Prior Achievement (APA)
- Accreditation of Prior Learning and Achievement (APLA).

All evidence must be evaluated with reference to the stipulated learning outcomes and assessment criteria against the respective unit(s). The assessor must be satisfied that the evidence produced by the learner meets the assessment standard established by the learning outcome and its related assessment criteria at that particular level.

Most often RPL will be used for units. It is not acceptable to claim for an entire qualification through RPL. Where evidence is assessed to be only sufficient to cover one or more learning outcomes, or to partly meet the need of a learning outcome, then additional assessment methods should be used to generate sufficient evidence to be able to award the learning outcome(s) for the whole unit. This may include a combination of units where applicable.

EQUALITY AND DIVERSITY

OTHM provides equality and diversity training to staff and consultants. This makes clear that staff and consultants must comply with the requirements of the Equality Act 2010, and all other related equality and diversity legislation, in relation to our qualifications.

We develop and revise our qualifications to avoid, where possible, any feature that might disadvantage learners because of their age, disability, gender, pregnancy or maternity, race, religion or belief, and sexual orientation.

If a specific qualification requires a feature that might disadvantage a particular group (e.g. a legal requirement regarding health and safety in the workplace), we will clarify this explicitly in the qualification specification.

UNIT SPECIFICATIONS

DATA SCIENCE FOUNDATIONS

Unit Reference Number	F/650/5562
Unit Title	Data Science Foundations
Unit Level	7
Number of Credits	20
Total Qualification Time (TQT)	200 hours
Guided Learning Hours (GLH)	100 hours
Mandatory / Optional	Mandatory
Sector Subject Area (SSA)	15.3 - Business management
Unit Grading type	Pass / Fail

Unit Aims

Data science makes use of the power of computation, statistical methods, and expert domain knowledge to analyse and gain practical insights from the huge amounts of data produced by organisations in business environments. The aim of this unit is to help learners understand what data science is, the role of data scientists, and the impact that big data has had and continues to have on society. As part of this unit, learners will be introduced to key concepts, tools, advantages, and challenges in the field of data science. Following completion of the unit, they will have acquired an understanding of the breadth of the field of data science, as well the role of modern approaches including machine learning and deep learning in this context.

Learning Outcomes, Assessment Criteria, and Indicative Content

Learning Outcomes – the learner will:	Assessment Criteria – the learner can:	Indicative content
1. Understand the scope of data science and the roles of data scientists.	1.1 Define the landscape of Data Science. 1.2 Evaluate key topics in Data Science. 1.3 Analyse the role of a Data Scientist in comparison to other IT roles.	**Data Science** Discussion of key topics, their roles, terms, and definitions in data science, for example: ● Statistics and probability. ● Data science, data mining, data analytics, and data visualisation.

		● Artificial Intelligence, machine learning, and deep learning. ● Data-driven approaches and big data. ● Programming, data structures, scientific computing, and cloud computing. **Role of a Data Scientist** Data scientist roles include: ● Collection and extraction of data from multiple sources. ● Analysis of data from multiple angles to produce insights. Looking for trends that highlight problems or opportunities. ● Mining vast amounts of data for valuable and actionable insights. ● Effective visualisation of data to highlight key results. ● Communication of important information and insights to business and IT leaders to enable effective decision making, operational excellence and business performance. ● Using insights acquired from data analysis to influence how an organisation approaches business challenges. Making data driven recommendations for business strategy.
2. Understand the impact of big data on society.	2.1 Define big data. 2.2 Evaluate the impact of big data on users and organisations for organisational decision making. 2.3 Critically analyse how Big Data is driving digital transformation.	**Big Data** Big data vs. traditional data management: key challenges and differences in approach.

	2.4 Critically analyse how big data and traditional data management differ. 2.5 Evaluate industry-leading tools and software for analysing and visualising data.	Fundamental characteristics of big data e.g., Doug Laney's Three Vs of Big Data (volume, velocity, and variety), as well as an extension of the Vs (variability, veracity, visualisation and value). **Data-Driven Decision-Making** Advantages of data-driven decision-making: ● Continuous improvement and planning. ● Real-time insights and identifying new opportunities. ● Cost reduction. ● Aligning decision making with business strategy. Challenges for data-driven approaches, for example: ● Inconsistent and unstandardised data. ● Bias and discrimination inherent in datasets or sampling approaches. Value that digital transformation projects can bring to business and how they achieve this e.g.: ● Revenue. ● Employee retention. ● Increased productivity. ● Creative performance. ● Brand sentiment. ● Customer satisfaction.

		Data Science Tools Industry leading tools and software solutions to analyse data: ● Programming languages and software e.g. Python, R, Matlab, Microsoft Excel. ● Data analytics tools: Oracle Analytics, Qlik Analytics platform, Google Fusion Tables, Open Refine, Apache Spark/Hadoop, SAS Sentiment Analysis, Node XL. ● Data visualisation tools and software e.g. Power BI, Google Chart, Canvas, Tableau, Oracle Visual Analyzer, SAS Visual Analytics, Matplotlib, Tensorboard. ● Deep learning libraries e.g. Tensorflow, Pytorch. ● Cloud services e.g. AWS, Microsoft Azure. ● Database tools e.g., SQL, MySQL.
3. Understand the legal and ethical responsibilities of data scientists and the challenges they face.	3.1 Explain the legal and ethical roles, responsibilities and challenges faced by data specialists. 3.2 Describe approaches for building ethics into data science. 3.3 Review the different strategies used by data specialists to ensure data compliance.	**Legal and Ethical Considerations for Data Scientists** Data protection, informed consent, and privacy issues for compliance to include: ● Personally identifiable information. ● Sensitive information e.g. protected health information.

		● General Data Protection Regulation (GDPR) rights and obligation, enforcement, and regulatory legal penalties. ● How legal rules for data compliance differ globally, challenges faced by global corporations in collecting and managing data on individuals. ● Ethical and privacy considerations for automated and large scale data collection. **Addressing legal and ethical challenges** Communities and frameworks for good practice in data science e.g. Data for Good Exchange (D4GX); Fairness, Accountability and Transparency in Machine Learning group (FAT/ML); Data Ethics Framework (gov.uk). Industry-leading compliance management software and tools, e.g., Microsoft Compliance Manager, Amazon Web Services (AWS) Compliance, IBM DataOps.

Assessment

To achieve a 'pass' for this unit, learners must provide evidence to demonstrate that they have fulfilled all the learning outcomes and meet the standards specified by all assessment criteria.

Learning Outcomes to be met	Assessment criteria to be covered	Type of assessment	Word count (approx. length)
All LO 1 to 3	All AC under LO 1 to 3	Report	3500 words

Indicative Reading List

Kelleher, J.D. & Tierney, B. (2018). *Data Science (The MIT Press Essential Knowledge Series) (Illustrated edition)*. The MIT Press.

Kotu, V., & Deshpande, B. (2019). *Data Science: Concepts and Practice*. Morgan Kaufmann Publishers, an imprint of Elsevier.

Grus, J. (2019). *Data Science from Scratch: First Principles with Python (2nd edition)*. O'Reilly.

Goodfellow, I. & Bengio, Y. & Courville A. (2016) *Deep Learning*. The MIT Press.

Norvig, P. & Russell, S. (2021). *Artificial Intelligence: A Modern Approach (4th edition)*. Pearson.

Nussbaumer Knaflic C. (2015). *Storytelling with Data: A Data Visualization Guide for Business Professionals*. Wiley.

Hill, D. G. (2019). *Data Protection: Governance, Risk Management, and Compliance*. CRC Press.

PROBABILITY AND STATISTICS FOR DATA ANALYSIS

Unit Reference Number	H/650/5563
Unit Title	Probability and statistics for data analysis
Unit Level	7
Number of Credits	20
Total Qualification Time (TQT)	200 hours
Guided Learning Hours (GLH)	100 hours
Mandatory / Optional	Mandatory
Sector Subject Area (SSA)	15.3 - Business management
Unit Grading type	Pass / Fail

Unit Aims

The goal of this unit is to provide an overview of fundamental concepts in probability and statistics for data science. Statistics is an essential mathematical tool used by data scientists to analyse data, test hypotheses on it, and draw conclusions from it. This unit will introduce fundamental notions of probability and statistical methods used in data science, equipping the learner with the foundational understanding necessary to appreciate the use of statistics in a variety of data science applications.

Learning Outcomes, Assessment Criteria, and Indicative Content

Learning Outcomes – the learner will:	Assessment Criteria – the learner can:	Indicative content
1. Understand the fundamentals of probability and statistics.	1.1 Define probability and statistics, and explain the difference between them. 1.2 Explain the notion of a probability distribution and give examples of well-known distributions. 1.3 Understand the role of stochastic processes in modelling sequences of random events. 1.4 Describe key statistics used to describe sets of events.	**Introduction to Probability and Statistics** Probability: the mathematical study of the likelihood of events to occur. Statistics: the analysis of the frequency of occurrence of past events. Understanding the difference between an underlying and an empirical distribution. Key concepts in probability and statistics:

	1.5 Understand the role of hypothesis testing in using data to answer questions. 1.6 Describe methods for probability distribution fitting with examples.	● Random variables and observed values. ● Independence and dependence. ● Sampling. ● Cumulative distribution function vs. probability density function. ● Discrete and continuous probability distributions. Common probability distributions – the Bernoulli distribution, Binomial distribution, and Normal distribution Stochastic processes for modelling sequences of random events over time. Examples of stochastic processes - Bernoulli process, Wiener process, Poisson process, Markov chains, Random Walk. **Statistics and Inference** Essential statistics for probability distributions: ● Mean, median, and mode. ● Variance and standard deviation. ● Understanding what each variable tells us about a set of data samples and/or the true distribution. Key concepts in hypothesis testing: p-values, null and alternative hypotheses, confidence intervals. Approaches to distribution fitting, e.g.: ● Maximum likelihood ● Method of moments

		● Maximum spacing estimation ● Method of L-moments
2. Be able to use statistics to test hypotheses on datasets.	2.1 Analyse how inferential statistical analysis differs from descriptive statistics. 2.2 Evaluate the need for different degrees of parameterisation in statistical models. 2.3 Explain the distinction between linear, generalised linear, and non-linear statistical models. 2.4 Use normal hypothesis tests to evaluate whether a given claim is true. 2.5 Understand how statistics plays a role in computational and machine learning models.	**Statistics** Descriptive statistics: describing and summarise the characteristics of a dataset e.g. through the use of calculating the mean and variance. Inferential statistics: using the dataset to make inferences (i.e. drawing conclusions or making predictions). Types of statistical models: ● Fully parametric, Semi-parametric, non-parametric. Advantages and disadvantages of each (i.e. level of assumption/bias, level of variance/uncertainty following model fitting). ● Linear, Generalised Linear (with a link function), and Non-Linear statistical models. **Hypothesis Testing** Hypothesis testing using an assumption of normality: defining the null and alternative hypotheses, computing p-values for the claim using look up tables, verifying whether or not we reject the null hypothesis based on the p-value. **Statistics and Data Science** The role of statistics in computational data science and machine learning: e.g. the

		bias/variance trade-off, parameterisation/over-parameterisation of models, model complexity, dataset size.
3. Understand Bayesian statistics.	3.1 Describe the difference between frequentist and Bayesian views of statistics. 3.1 Explain Bayes' theorem and its importance and use in statistics. 3.2 Evaluate Bayesian experimental design with appropriate examples. 3.3 Discuss Markov Chain Monte Carlo (MCMC) methods and MCMC simulations with examples.	**Frequentist vs. Bayesian viewpoint** Two approaches to estimation of statistical parameters from samples. Frequentist: assuming that probability is based on the frequency of events over time. Leads to the need to infer via point estimates (estimates of the probability at a single value on the distribution support), confidence intervals, and hypothesis testing (p-values). Bayesian: assumes that probability is based on degree of belief in parameters (priors). Inference updates beliefs Core difference: consideration of the parameters as random variables or as fixed. Comparison of frequentist and Bayesian approaches: ● Treatment of probability as objective (frequentist) vs. degree of belief (Bayesian). ● Parameters are fixed values (frequentist) vs. random variables (Bayesian). ● Incorporating prior information/knowledge (Bayesian) vs. not accounting for this (frequentist).

		Bayes' Theorem and Bayesian Inference Introduction to Bayesian inference and Bayes' Theorem. Core components: ● Prior: Encodes beliefs about parameters before seeing data. ● Likelihood: The probability of the observed data given the prior on the parameters. ● Posterior: Beliefs about parameters after seeing data. ● Marginal likelihood/evidence: The probability of the observed data. Using Bayes' theorem for updating the belief, the significance of conjugate priors for obtaining closed-form solutions to the posteriors. **Markov Chain Monte Carlo** A computational Bayesian method for approximating the posterior distribution empirically, rather than analytically. Key for enabling Bayesian inferefence where there is no closed form solution for the posterior distribution. MCMC sampling methods: ● Gibbs sampling ● Metropolis–Hastings algorithm ● Slice sampling

		● Hamiltonian (or Hybrid) Monte Carlo (HMC) MCMC will not work for model comparisons, where an explicit marginal likelihood (evidence) is needed.
4. Be able to perform linear statistical modelling.	4.1 Discuss simple and multiple linear regression models with examples. 4.2 Discuss and compare linear least squares methods with examples. 4.3 Define heteroscedasticity, analyse the need for it, and explain how it is used in the weighted least squares algorithm. 4.4 Compare traditional statistical and machine learning approaches to linear regression	**Linear Modelling** Distinction between simple linear regression, multiple regression in terms of number of independent variables. Multivariate regression: multiple dependent variables. Linear least squares methods: Ordinary least squares (error variances are all the same), weighted least squares (for heteroscedastic data, assuming variances of errors differ), generalised least squares (assumes an arbitrary covariance for errors). Regression in statistics vs. machine learning: ● Goal: in machine learning, the goal is to minimise prediction error using cost functions and gradient descent vs. in statistics, where the goal is to form estimates of statistical parameters. ● Solution: Statistics typically has a closed form Ordinary Least Squares expression for linear regression vs. machine learning methods rely on iterative optimisation e.g. (stochastic) gradient descent. ● Analysis and evaluation: Analysis of results in statistics involves e.g. significance testing and confidence

		intervals or residual analysis, whereas in machine learning it involves comparison of training/validation/test error. ● Use cases: Machine learning methods can be useful for higher-dimensional data, where a closed form statistical expression is computationally complex.

Assessment

To achieve a 'pass' for this unit, learners must provide evidence to demonstrate that they have fulfilled all the learning outcomes and meet the standards specified by all assessment criteria.

Learning Outcomes to be met	Assessment criteria to be covered	Type of assessment	Word count (approx. length)
All LO 1 to 4	All AC under LO 1 to 4	Report + Source Code	3500 words

Indicative Reading List

Kaptein, M., & den, H. E. van. (2022). *Statistics for data scientists: An introduction to probability, statistics, and data analysis*. Springer.

Rigdon, S. E., & Fricker, R. D., & Montgomery D. C. (2025). *Introduction to Probability and Statistics for Data Science with R*. Cambridge University Press

Nield, T. (2022) *Essential Math for Data Science*. O'Reilly.

DATA ANALYSIS AND VISUALISATION

Unit Reference Number	K/650/5565
Unit Title	Data Analysis and Visualisation
Unit Level	7
Number of Credits	20
Total Qualification Time (TQT)	200 hours
Guided Learning Hours (GLH)	100 hours
Mandatory / Optional	Mandatory
Sector Subject Area (SSA)	15.3 - Business management
Unit Grading type	Pass / Fail

Unit Aims

This unit introduces learners to the fundamentals of the data analysis pipeline. These including the process of gathering, cleaning and analysing data and communicating key insights derived visually. Learners will acquire an understanding of common tools and software for data analysis and visualisation, and will gain practical experience applying these techniques to clean, prepare, visualise, and communicate a variety of types of data, keeping intended audience in mind.

Learning Outcomes, Assessment Criteria, and Indicative Content

Learning Outcomes – the learner will:	Assessment Criteria – the learner can:	Indicative content
1. Understand the foundational principles of data analytics.	1.1 List examples of the three types of data analytics and evaluate their use in industry. 1.2 Describe the key stages of data analysis 1.3 Demonstrate the ability to prepare datasets for analysis and visualisation.	**Data Analytics** Three types of data analytics: ● Descriptive: Understanding a dataset. ● Predictive: Making predictions for future outcomes based on current data. ● Prescriptive: Providing recommendations for future courses of action based on data.

		Examples: ● Descriptive: data visualisation, cluster analysis, factor analysis, univariate and bivariate analysis. ● Predictive: regression analysis, time series analysis e.g. ARIMA, sentiment analysis. ● Prescriptive analytics: supply chain optimisation, pricing strategies, cohort analysis. **Data Analytics Pipeline** Stages of data analysis: ● Data curation and preprocessing, including: sampling, cleaning/transformation (e.g. outlier identification, missing values treatment, normalisation), integration of data from multiple sources, and dimensionality reduction. ● For predictive and prescriptive purposes: fitting models to data following preprocessing. ● Visualising results and communicating them to stakeholders in a way which is comprehensible and effective for the intended audience.
2. Be able to apply prescriptive analytics to inform decision making.	2.1 Understand how data can be used to inform decision making. 2.2 Evaluate the advantages of predictive and prescriptive analytics for organisations and	**Predictive and Prescriptive Analytics** Advantages of data analytics for business, e.g.:

	businesses. 2.3 Understand different approaches to data-driven decision making. 2.3 Demonstrate the use of prescriptive analytics using appropriate software or tools.	● Anticipation of the future using predictive analytics, identifying trends, and forecasting demand. ● Recommending optimal actions. ● Improving the quality of decision making by using data to support arguments (justifiability). ● Establishing the defensibility of decisions through data-driven decision making in the context of prescriptive analytics. Approaches to decision making using data, e.g.: ● Naïve approach: using simple assumptions or basing predictions on recent history. ● Average approach: basing predictions on the mean of past values. ● Advanced approaches: times series, regression, or machine learning for forecasting. ● Qualitative vs. quantitative approaches to decision making. Many predictive methods are quantitative, but qualitative approaches can be used where some data is missing. Techniques for prescriptive analytics: ● Classical optimisation. ● Linear and non-linear programming. ● Dynamic programming. ● Simulation to explore how a model behaves for different decisions and under different assumptions. ● Decision analysis: systematic evaluation of alternatives, taking into account probability, cost, and benefit.

		Using R or Python for forecasting modelling and associated libraries or functions (e.g. Python's SciPy).
3. Be able to visualise and present data in an audience-appropriate way.	3.1 Understand the need for effective data visualisation. 3.2 Understand how to tailor the presentation of data to a given audience. 3.3 Use appropriate tools and software to visualise data. 3.4 Critically evaluate data visualisation approaches.	**Data Visualisation** Why do we need to visualise data? Examples of discussion points: ● To enable information to be conveyed more concisely and rapidly to a broader audience. ● To tailor the presentation of information to different audiences. ● To present information in a way that guides insight and understanding. ● Identification of patterns: trends, outliers. ● To form compelling visual arguments to propose, justify, and defend decisions. Considerations for data visualisation: clarity and simplicity, accuracy, functionality, storytelling, aesthetics, best practices and rules of thumb. Visualisation approaches for different data modalities: ● Categorical and qualitative data: bar charts, pie chart etc. ● Continuous and quantitative data: line charts, box plots, histograms, scatter plots etc. ● Time series data: line charts, area charts. ● Relationships between features: scatter plots, heatmaps, bubble charts etc. Software for data visualisation, e.g. ● Python and associated libraries (Matplotlib, Seaborn).

		● R for statistical visualisation (using ggplot2). ● Interactive dashboards, e.g. Tableau, Power BI, Plotly (Python). ● Excel and Google Sheets for basic visualisations. Case studies on data visualisation, analysing them and critically evaluating their effectiveness in conveying information in a clear and concise way in a manner appropriate for the intended audience.

Assessment

To achieve a 'pass' for this unit, learners must provide evidence to demonstrate that they have fulfilled all the learning outcomes and meet the standards specified by all assessment criteria.

Learning Outcomes to be met	Assessment criteria to be covered	Type of assessment	Word count (approx. length)
All LO 1 to 3	All AC under LO 1 to 3	Report + Source Code	3500 words

Indicative Reading List

Runkler, T. A. (2020). *Data analytics: Models and algorithms for Intelligent Data Analysis*. Springer Vieweg.

Chen, Y. D. (2022). *Pandas for Everyone: Python Data Analysis (2nd edition)*. Addison-Wesley Professional.

Delen, D. (2019). *Prescriptive Analytics: The Final Frontier for Evidence-Based Management and Optimal Decision Making*. FT Press.

Meyer, P. I. (2023). *The 6 Pillars of Decision Making*. Mind Mentor.

Nussbaumer Knaflic, C. (2015). *Storytelling with Data: A Data Visualization Guide for Business Professionals*. Wiley.

McCandless, D. (2014) *Knowledge is Beautiful*. William Collins.

McCandless, D. (2012) *Information is Beautiful*. Collins.

ADVANCE PREDICTIVE MODELLING

Unit Reference Number	J/650/5564
Unit Title	Advanced Predictive Modelling
Unit Level	7
Number of Credits	20
Total Qualification Time (TQT)	200 hours
Guided Learning Hours (GLH)	100 hours
Mandatory / Optional	Mandatory
Sector Subject Area (SSA)	15.3 - Business management
Unit Grading type	Pass / Fail

Unit Aims

This unit introduces learners to some of the most widely used predictive modelling techniques and their core principles. Through this unit, learners will build a solid understanding of predictive analytics, which refers to tools and techniques for building statistical or machine learning models to make predictions based on data. Learners will develop and apply the basic understanding of statistical models they have developed previously and implement practical schemes for data analysis and prediction on real datasets.

Learning Outcomes, Assessment Criteria, and Indicative Content

Learning Outcomes – the learner will:	Assessment Criteria – the learner can:	Indicative content
1. Understand a range of generalised linear models.	1.1 Understand the distinction between simple and generalised linear models. 1.2 Identify the key components of a generalised linear model. 1.3 Describe a range of common generalised linear models and their associated link functions. 1.4 Describe the characteristics of dependent variables that lead to different types of generalised linear modelling.	**Linear and Generalised Linear Models** Key distinction: generalised linear models include a link function which allows the output to follow a different distribution to the input. Components of generalised linear models: target distribution (from the exponential family), linear predictor, and link function.

	1.5 Evaluate the difference between regression, ordinal regression, and multinomial regression.	Examples of generalised linear models with associated link functions: ● Linear model: identity link function for e.g. regression to a normal distribution. ● Log linear model: logarithm of the response variable varies linearly (logarithmic link function) i.e. log(p) for e.g. Poisson regression. ● Logistic models (log odds): logarithm of the odds varies linearly (logarithmic link function with odds) i.e. log(p/(1-p)) for e.g. binomial regression (binary classification) or multinomial regression (multiple classes). **Characteristics of dependent variables** Data types and associated generalised linear models: ● Continuous vs. categorical data: regression vs. classification. ● Ordinal vs. multinomial data: ordered vs. unordered categorical data. ● The range of data: infinite, bounded, positive etc.
2. Be able to implement regression methods for continuous, ordered and unordered data.	2.1 Identify appropriate regression modelling approaches to apply to different types of data. 2.2 Develop simple linear models using	**Generalised Linear Modelling** Examples of model types and their use cases:

	suitable software, e.g. R or Python.
	2.3 Explain the notion of maximum entropy classification in the context of multinomial logistic regression.
	2.4 Develop a range of generalised linear models for continuous, nominal, and ordinal data.
	2.5 Apply the Poisson regression model and discuss and address overdispersion and zero inflation.
		● Linear regression: dependent variable is a linear function of the independent variables.
		● Logistic regression: suitable for a binomial distribution.
		● Poisson regression (for count data): makes use of the log linear link function. Approaches for addressing the problems with Poisson regression:
		  ○ Overdispersion (e.g. use negative binomial regression).
		  ○ Zero inflation (e.g. use mixture models such as zero-inflated Poisson).
		● Multinomial logistic regression (multiclass classification): for unordered categorical data with a multinomial logit link (inverse is the softmax function). Notion of maximum entropy classification.
		● Ordinal regression: intermediate between regression and classification – discrete classes which have an order – use the ordered logistic link function, or ordered probit.
		
		Implementing regression for a variety of models (including simple linear and logistic regression) using a popular programming language e.g. R or Python, with appropriate libraries or functions (e.g. scipy in Python).

3. Be able to develop survival analysis models.	3.1 Define survival analysis. 3.2 Compare the hazard function to the survival function and detail how they are	**Survival analysis**

	related. 3.3 Describe use cases for survival analysis. 3.4 Describe and implement the Cox proportional hazards model for survival analysis.	Survival analysis involves time-to-event data: analysing how long it takes until an event occurs. Hazard function: the rate (risk) of an event occurring at a time t. Survival function: the probability of no event occurring up to a time t. The survival function is the exponential of the negative cumulative hazard function. The cumulative hazard gives you the expected number of events over a given interval, and the survival function is the probability of no events occurring over such an interval. Use cases for survival analysis: e.g. modelling time to failure of electrical components, time for a patient to recover from an illness (for modelling hospital capacity), time for customer retention (e.g. cancelling a subscription). Using Cox proportional hazards for survival analysis: ● Cox proportional hazards estimates the hazard-ratios (relative hazards between groups), which are assumed constant over time. ● Predicted hazard ratios allow for estimation of the relative hazard between groups and how these are affected by the independent variables. ● Semi-parametric method: explicitly models effect of independent variables, but not the baseline hazard.

		● The model is linear in log of the hazard.
4. Be able to use neural networks for predictive modelling.	4.1 Evaluate appropriate settings for the application of deep learning methods to predictive analysis. 4.2 Describe key architectures used for predictive modelling using neural networks in different settings. 4.3 Use neural networks to solve predictive modelling problems.	**Neural Predictive Modelling** Neural network architectures: ● Basic and general architectures: fully connected layers (multi-layer perceptron), transformers. ● Time series prediction: convolutional architectures, LSTMs, transformer architectures, autoregressive models. **Developing and Training Neural Predictive Models** Training and testing pipeline, loss function definition, evaluation metric specification, baseline implementation. Choosing networks of appropriate architecture given the dataset and problem setting (e.g. time series predictive modelling vs. regression vs. classification). Choosing the number of parameters/complexity based on dataset size, and including inductive biases through choice of architecture (e.g. number of parameters/using different types of models e.g. convolutional).

Assessment

To achieve a 'pass' for this unit, learners must provide evidence to demonstrate that they have fulfilled all the learning outcomes and meet the standards specified by all assessment criteria.

Learning Outcomes to be met	Assessment criteria to be covered	Type of assessment	Word count (approx. length)
All LO 1 to 4	All AC under LO 1 to 4	Report + Source Code	3500 words

Indicative Reading List

Kuhn, M., & Johnson, K. (2016). *Applied predictive modeling*. Springer.

Gelman, A. & Hill, J. (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Agresti, A. (2012) *Categorical Data Analysis (3rd edition)*. Wiley.

Agresti, A. (2015) *Foundations of Linear and Generalized Linear Models*. Wiley.

Chollet, F. (2021) *Deep Learning with Python (2nd edition)*. Manning Publications.

DATA MINING, MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Unit Reference Number	J/650/5573
Unit Title	Data Mining, Machine Learning and Artificial Intelligence
Unit Level	7
Number of Credits	20
Total Qualification Time (TQT)	200 hours
Guided Learning Hours (GLH)	100 hours
Mandatory / Optional	Mandatory
Sector Subject Area (SSA)	15.3 - Business management
Unit Grading type	Pass / Fail

Unit Aims

This unit introduces Artificial Intelligence (AI) as the use of data-driven methods to solve real-world problems with minimal hand design, and its role in the field of data science. Learners will be introduced to models and methods popular today, the impact of the future growth of AI and philosophical debates concerning it, and the ethical issues surrounding the use of AI for data mining. As part of this unit, students will gain the knowledge, understanding and practical experience necessary to apply machine learning techniques to a variety of challenging, real-world problems.

Learning Outcomes, Assessment Criteria, and Indicative Content

Learning Outcomes – the learner will:	Assessment Criteria – the learner can:	Indicative content
1. Understand the meaning of Artificial Intelligence and its role in data science.	1.1 Define Artificial Intelligence. 1.2 Compare the scope of the fields of machine learning, deep learning, and Artificial Intelligence as a whole. 1.3 Differentiate between ANI, AGI and ASI. 1.4 Evaluate the impact of deep learning on the field of data science. 1.5 Describe key technologies used in modern Artificial Intelligence.	**Artificial Intelligence** Artificial Intelligence (AI) is emulation of tasks by computers traditionally possible only for humans to perform, e.g. image classification, natural language processing, automated data clustering/labelling, reasoning, and planning. Philosophical debates surrounding ambitions of

	1.6 Describe a range of real-world applications of artificial intelligence in disparate domains.	simulating human intelligence, consciousness etc. Appreciating the difference between AI and its subfields, e.g., symbolic AI, machine Learning, deep learning and related interdisciplinary research areas such as robotics. The connections of AI and its use in diverse fields, e.g. computer science, mathematics and statistics, robotics, neuroscience, computer vision, and linguistics. Understanding the terms Artificial Narrow Intelligence (ANI), Artificial General Intelligence (AGI), and Artificial Super-intelligence (ASI). Data-driven approaches in AI: e.g. machine learning, including traditional methods (e.g. random forests, k-means clustering, support vector machines), and modern methods in deep learning (e.g. generative models, large language models, vision models, deep neural networks). Key technologies used in modern deep learning: transformers, diffusion models, reinforcement learning, and unsupervised learning for learning in an automatic way from large amounts of unlabelled data e.g. text on the internet. **Applications of AI** Business and e-commerce, e.g., chatbots, visual searches, intelligent virtual assistants.

		Engineering, e.g., Computer Aided Design (CAD), automation in factories. Healthcare, e.g., care of the elderly, heart beat analysis, computer-aided interpretation of medical images, drug discovery.
2. Be able to use machine learning methods to address real world problems.	2.1 Describe and compare machine learning and modern deep learning methods for solving different problems. 2.2 List common tools and software used to implement machine learning models. 2.3 Describe and implement processes for preparing data for training and evaluation. 2.4 Use appropriate machine learning methods to solve real world problems in data science. 2.5 Analyse and explore technical options to tune and enhance the performance of machine learning-based systems. 2.6 Use libraries and online resources to implement existing deep learning models.	**Machine Learning Methods** Classical machine learning methods: Linear regression, logistic regression, decision tree, Support Vector Machine (SVM), Naïve Bayes, K-Nearest Neighbor(s) (KNN), k-means, gradient boosting. Introduction to neural networks, deep neural networks, and frontier models (e.g. large language models). Common architectures and training methods: fully connected layers, convolutional layers, transformers, stochastic gradient descent using different optimisers e.g. Adam, RMSprop. Software frameworks and libraries for deep learning: PyTorch, Numpy, Pandas, Scikit-Learn, TensorFlow/Keras, JAX. Tools for machine learning and deep learning, e.g.: Azure ML, Google Colab, Hugging Face, AWS for ML. **Practical Implementation of Machine Learning Methods** Dataset preparation:

		● Selecting appropriate data/datasets, addressing issues such as missing values, class imbalance. Could include common dataset downloads using e.g. PyTorch Datasets. ● Evaluating the reliability of data. ● Feature selection and transformation (e.g. one hot encoding for categorical features). ● Normalisation. ● Efficient loading during training/evaluation. Building neural networks: ● Common software and tools to build neural networks (e.g. PyTorch). ● Choosing dataset hyperparameters (e.g. batch size, output range, train-test split. ● Choosing appropriate loss functions and training techniques for a given dataset and data modality (e.g. autoencoders, mean squared error, cross entropy loss, contrastive learning e.g. Contrastive Learning Image Pretraining (CLIP)). ● Defining architectures using e.g. layerwise definitions, e.g. using the PyTorch nn package, or TensorFlow's Keras. ● Evaluating and visualising training results, comparing to baselines. Implementing existing deep learning models:

		● Using libraries to implement popular models (e.g. Huggingface, Keras, Pytorch models). ● Using online repositories e.g. GitHub to download, reproduce, and extend a wide variety of models from open source projects.
3. Understand technical, ethical, social and legal considerations for AI.	3.1 Analyse technical challenges in using machine learning models for data science at scale. 3.2 Evaluate the societal benefits of AI. 3.3 Evaluate key ethical and legal challenges for modern AI. 3.4 Describe best practices for AI model development, from a technical, ethical, and legal perspective.	**Ethical, Social, and Legal Issues in AI** Technical challenges and limitations of AI, e.g.: ● Hallucinations in Large Language Models. ● Limited training data. ● High energy expenditure and hardware requirements for training frontier models. ● Distributed and parallelised training. Societal benefits of AI, now and in the future, e.g.: ● Education, access to knowledge, and tailored tutoring. ● Improvements to health and medicine and automated drug discovery. ● Addressing environmental challenges and monitoring climate change. Ethical considerations challenges e.g.: ● Ethical concerns relating to training/using large language models trained on large

		datasets e.g. copyright and ownership of training data. ● Use of deep learning in recruiting new employees (automated employee recommendations/CV scanning) and potential for bias. ● Harmful content generation and deepfakes, and their potential negative impact on individuals and society. ● Adversarial examples and poisoned training data. ● Bias, interpretability, and alignment in large language models. ● Potential to widen socio-economic inequality due to unequal access to technologies, unemployment caused by AI. ● Climate change driven by AI's environmental footprint. ● Algorithmic quantitative trading and the impact on global financial markets. Good practice in AI model development, e.g.: ● Accurate and clear documentation. ● Role of statistic testing and review in early defect detection. ● Following specific industry standards (e.g. GDPR).

Assessment

To achieve a ‘pass’ for this unit, learners must provide evidence to demonstrate that they have fulfilled all the learning outcomes and meet the standards specified by all assessment criteria.

Learning Outcomes to be met	Assessment criteria to be covered	Type of assessment	Word count (approx. length)
All LO 1 to 3	All AC under LO 1 to 3	Report + Source Code	3500 words

Indicative Reading List

Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). *Mathematics for Machine Learning*. Cambridge University Press.

Russell, S. J., & Norvig, P. (2022). *Artificial Intelligence: A modern approach*. Pearson.

Bishop, C., & Bishop, H. (2023). *Deep Learning: Foundations and Concepts*. Springer.

Goodfellow, I., & Bengio, Y., & Courville, A. (2023). *Deep Learning*. Alanna Maldonado.

Coeckelbergh, M. & Gerrard, L. (2020). *AI Ethics*. MIT Press Essential Knowledge.

Olson, P. (2024). *Supremacy: AI, ChatGPT and the Race That Will Change the World*. Macmillan Business.

ADVANCED COMPUTING RESEARCH METHODS

Unit Reference Number	L/650/5566
Unit Title	Advanced Computing Research Methods
Unit Level	7
Number of Credits	20
Total Qualification Time (TQT)	200 hours
Guided Learning Hours (GLH)	100 hours
Mandatory / Optional	Mandatory
Sector Subject Area (SSA)	15.3 - Business management
Unit Grading type	Pass / Fail

Unit Aims

The aim of this unit is to develop learners' ability to prepare for various types of academically based computing research through the development and design of a research proposal. Learners will develop a critical understanding of the philosophical, practical, and ethical concepts of research within the context of the computing discipline.

Learning Outcomes, Assessment Criteria, and Indicative Content

Learning Outcomes – the learner will:	Assessment Criteria – the learner can:	Indicative content
1. Be able to evaluate research approaches in the computing discipline.	1.1 Appraise appropriate research problems in your chosen area. 1.2 Develop and justify appropriate research aims and objectives within a defined scope and timeframe. 1.3 Critically explore, select and justify research approaches.	**Proposing a Research Project** Qualitative and quantitative approaches to Computing research The strengths and weaknesses of different approaches to public sector research. SMART objectives; terms of reference; rationale for selection, public sector confidence. GANTT charts, Key milestones, project goals

	1.4 Produce a SMART research plan using a suitable software.	
2. Be able to critically review literature on a relevant research topic.	2.1 Evaluate different literature sources to find the most appropriate literature for the chosen research topic. 2.2 Critically analyse different theoretical approaches to the research problem.	**Literature Review** Conceptualisation of the research problem or hypothesis. The importance of positioning a research project in the context of existing knowledge. Significance and means of providing benchmarks by which data can be judged. Key theoretical frameworks for research. Advantages and limitations of qualitative and quantitative research approaches and methods
3. Be able to design research methodologies for a computing research problem.	3.1 Critically evaluate relevant research methodologies to reflect the research objectives. 3.2 Design an appropriate methodology in terms of the research objectives for a defined population. 3.3 Justify the methodology selected in terms of the research objectives within the bounds of agreed ethical guidelines. 3.4 Propose suitable techniques to use with quantitative and qualitative data collection and analysis.	**Research Methodologies** Research methods e.g., survey, questionnaire, observations; ways to test sufficiency, reliability and validity; definitions of data e.g., primary and secondary sources, qualitative and quantitative; literature search and review – its credibility, use and acceptance; ways to reference sources. Size and sufficiency of data, assessment of the reliability and validity of information gathered.
4. Be able to develop a research proposal.	4.1 Create a research question, literature review, and methodology.	**Research Proposal Writing** Report structure e.g., title, acknowledgements, contents page, introduction, summary of literature review, research methods used,

	4.2 Propose techniques for use with quantitative or qualitative data collection and analysis.	findings, recommendations, references, bibliography, appendices e.g., questionnaires, surveys. Referencing e.g., Harvard system.

Assessment

To achieve a 'pass' for this unit, learners must provide evidence to demonstrate that they have fulfilled all the learning outcomes and meet the standards specified by all assessment criteria.

Learning Outcomes to be met	Assessment criteria to be covered	Type of assessment	Word count (approx. length)
All LO 1 to 4	All AC under LO 1 to 4	Report	3500 words

Indicative Reading List

Lauro, N. C. (2018). *Data Science and Social Research: Epistemology, methods, technology and applications*. Springer.

Additional Resources

Mastering Predictive Analytics with R - Second Edition James D. Miller, Rui Miguel Forte Publisher Packt Publication date: August 2017

IMPORTANT NOTE

Whilst we make every effort to keep the information contained in programme specification up to date, some changes to procedures, regulations, fees matter, timetables, etc may occur during the course of your studies. You should, therefore, recognise that this booklet serves only as a useful guide to your learning experience.

For updated information please visit our website www.othm.org.uk