Level 3 Diploma in Data Science

Qualification Specification

January 2023

Contents

About QUALIFI

QUALIFI is recognised and regulated by Ofqual (Office of Qualifications and Examinations Regulator). Our Ofqual reference number is RN5160. Ofqual regulates qualifications, examinations, and assessments in England.

As an Ofqual recognised Awarding Organisation, QUALIFI is required to carry out external quality assurance to ensure that centres approved for the delivery and assessment of QUALIFI's qualifications meet the required standards.

Why Choose QUALIFI Qualifications?

QUALIFI qualifications aim to support learners to develop the necessary knowledge, skills and understanding to support their professional development within their chosen career and or to provide opportunities for progression to further study.

Our qualifications provide opportunities for learners to:

- apply analytical and evaluative thinking skills
- develop and encourage problem solving and creativity to tackle problems and challenges
- exercise judgement and take responsibility for decisions and actions
- develop the ability to recognise and reflect on personal learning and improve their personal, social, and other transferable skills.

Employer Support for the Qualification Development

During the development of this qualification QUALIFI consults with a range of employers, providers, and existing centres (where applicable) to ensure rigour, validity and demand for the qualification and to ensure that the development considers the potential learner audience for the qualification and assessment methods.

Equality and Diversity

QUALIFI's qualifications are developed to be accessible to all learners who are capable of attaining the required standard. QUALIFI promotes equality and diversity across aspects of the qualification process and centres are required to implement the same standards of equal opportunities and ensure teaching and learning are free from any barriers that may restrict access and progression.

Learners with any specific learning need should discuss this in the first instance with their approved centre who will refer to QUALIFI's Reasonable Adjustment and Special Consideration Policy.

Qualification Title and Accreditation Number

This qualification has been accredited to the Regulated Qualification Framework (RQF) and has its own unique Qualification Accreditation Number (QAN). This number will appear on the learner's final certification document. Each unit within the qualification has its own RQF code. The QAN for this qualification is as follows:

Qualifi Level 3 Diploma in Data Science 610/1950/1

Qualification Aims and Learning Outcomes

Aims of the QUALIFI Level 3 Diploma in Data Science

The aim of the QUALIFI Level 3 Diploma in Data Science is to provide learners with an introduction and understanding of the field of data science.

The Level 3 Diploma provides a contemporary and holistic overview of data science, artificial intelligence, and machine learning, from the birth of artificial intelligence and machine learning in the late 1950s, to the dawn of the "big data" era in the early 2000s, to the current applications of AI and machine learning and the various challenges associated with them. In addition to the standard machine learning models of linear and logistic regression, decision trees and *k*-means clustering, the diploma introduces learners to two new exciting and emerging areas of data science: synthetic data and graph data science.

The Diploma also introduces learners to the data analytical landscape and associated analytical tools, teaching introductory Python so that Learners can analyse, explore, and visualise data, as well as implement a number of basic data science models.

Successful completion of the QUALIFI Level 3 Diploma in Data Science provides learners with the opportunity to progress to further study or employment.

Learning Outcomes of the QUALIFI Level 3 Diploma in Data Science

The overall learning outcomes of the qualification are for learners to:

i) Gain the mathematical and statistical knowledge and understanding required to conduct basic data analysis.

ii) Develop analytical and machine learning skills with Python.

iii) Develop a strong understanding of data and data processes, including data cleaning, data structuring, and preparing data for analysis and visualisation.

iv) Understand the data science landscape and ecosystem, including relational databases, graph databases, programming languages such as Python, visualisation tools, and other analytical tools.

v) Understand the machine learning processes, understanding which algorithms to apply to different problems, and the steps required build, test and verify a model.

vi) Develop an understanding of contemporary and emerging areas of data science, and how they can be applied to modern challenges.

The learning outcomes and assessment criteria for each unit are outlined in the unit specifications.

Progression and Links to other QUALIFI Programmes

Completing the **QUALIFI Level 3 Diploma ii Data Science** will enable learners to:

- Progress to QUALIFI Level 4 Diploma in Data Science.
- Apply for entry to a UK university for an undergraduate degree.
- Progress to employment in an associated profession.

Delivering the Qualification

External Quality Assurance Arrangements

All centres are required to complete an approval process to be recognised as an approved centre. Centres must have the ability to support learners Centres must commit to working with QUALIFI and its team of External Quality Assurers (EQAs). Approved Centres are required to have in place qualified and experienced tutors, all tutors are required to undertake regular continued professional development (CPD).

Approved centres will be monitored by QUALIFI External Quality Assurers (EQAs) to ensure compliance with QUALIFI requirements and to ensure that learners are provided with appropriate learning opportunities, guidance, and formative assessment.

QUALIFI's guidance relating to invigilation, preventing plagiarism and collusion will apply to centres.

QUALIFI, unless otherwise agreed:

- sets all assessments;
- moderates assessments prior to certification;
- awards the final mark and issues certificates.

Learner Induction and Registration

Approved Centres should ensure all learners receive a full induction to their study programme and the requirements of the qualification and its assessment.

All learners should expect to be issued with the course handbook and a timetable and meet with their personal tutor and fellow learners. Centres should assess learners carefully to ensure that they are able to meet the requirements qualification and that, if applicable, appropriate pathways or optional units are selected to meet the learner's progression requirements.

Centres should check the qualification structures and unit combinations carefully when advising learners. Centres will need to ensure that learners have access to a full range of information, advice and guidance to support them in making the necessary qualification and unit choices. During recruitment, approved centres need to provide learners with accurate information on the title and focus of the qualification for which they are studying.

All learners must be registered with QUALIFI within the deadlines outlined in the QUALIFI Registration, Results and Certification Policy and Procedure.

Tutor/trainer requirements

Tutors must be appropriately qualified and occupationally competent in the areas in which they are training. They must

i) Hold a Level 6 qualification or higher in data science or in a related technical subject.
ii) Have a minimum of 5 years' relevant experience in data science or in a related technical subject.
iii) Hold, or be working towards, a Level 3 Award in Education and Training or equivalent.
iv) Demonstrate that they have undertaken Continued Professional Development (CPD) activities relating to data science and analytics to maintain and update their skills and knowledge within each 12-month period.

Internal Verifier/Moderator Requirements

Internal Verifiers must be appropriately qualified and occupationally competent in the areas in which they are moderating. They must

i) Hold a Level 6 qualification or higher in data science or in a related technical subject.
ii) Have a minimum of 3 years' relevant experience in data science or in a related technical subject.

iii) Hold, or be working towards, a Level 4 Award in the Internal Quality Assurance of Assessment Processes and Practice and/or Level 4 Certificate in Leading the Internal Quality Assurance of Assessment Processes and Practice.

iv) Demonstrate that they have undertaken Continuous Professional Development (CPD) activities relating to data science and analytics to maintain and update their skills and knowledge within each 12-month period.

Entry Criteria

Approved Centres are responsible for reviewing and making decisions as to the applicant's ability to complete the learning programme successfully and meet the demands of the qualification. The initial assessment by the centre will need to consider the support that is readily available or can be made available to meet individual learner needs as appropriate.

The qualification has been designed to be accessible without artificial barriers that restrict access. For this qualification, applicants must be aged 18 or over.

Entry to the qualification will be through centre-led registration processes which may include interview or other appropriate processes.

Although there is a significant amount of advanced mathematics and statistics in advanced data science courses, including linear algebra and differential calculus, in this Level 3 Diploma, Learners only need to be comfortable with GCSE level mathematics. All the mathematical and statistical concepts covered in the Diploma require nothing more than standard mathematical operations of addition, multiplication, and division.

Prior to starting the Level 3 Diploma in Data Science, learners are expected to hold at a minimum:

i) GCSE Mathematics at grade B or higher (new level 6 or above); and

ii) GCSE English at grade C or higher (new level 4 or above).

In addition, no prior coding experience is required though learners must be willing and comfortable to learn Python. Python has been specifically chosen as it easy to use and learn.

In certain circumstances, applicants with considerable experience but no formal qualifications may be considered, subject to interview and being able to demonstrate their ability to cope with the demands of the qualification.

Recognition of Prior Learning

Recognition of Prior Learning (RPL) is a method of assessment (leading to the award of credit) 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 so do not need to develop through a course of learning.

QUALIFI encourages centres to recognise learners' previous achievements and experiences whether at work, home or at leisure, as well as in the classroom. RPL provides a route for the recognition of the achievements resulting from continuous learning. RPL enables recognition of achievement from a range of activities using any valid assessment methodology. Provided that the assessment requirements of a given unit or qualification have been met, the use of RPL is acceptable for accrediting a unit, units, or a whole qualification.

Evidence of learning must be valid and reliable. For full guidance on RPL please refer to QUALIFI's *Recognition of Prior Learning Policy*.

Data Protection

All personal information obtained from learners and other sources in connection with studies will be held securely and will be used during the course and after they leave the course for a variety of purposes and may be made available to our regulators. These should be all explained during the enrolment process at the commencement of learner studies. If learners or centres would like a more detailed explanation of the partner and QUALIFI policies on the use and disclosure of personal information, please contact QUALIFI via email support@QUALIFI-international.com

Learner Voice

Learners can play an important part in improving the quality through the feedback they give. In addition to the on-going discussion with the course team throughout the year, centres will have a range of mechanisms for learners to feed back about their experience of teaching and learning.

Professional Development and Training for Centres

QUALIFI supports its approved centres with training related to our qualifications. This support is available through a choice of training options offered through publications or through customised training at your centre.

The support we offer focuses on a range of issues including:

- planning for the delivery of a new programme

- planning for assessment and grading
- developing effective assignments
- building your team and teamwork skills
- developing learner-centred learning and teaching approaches
- building in effective and efficient quality assurance systems.

Please contact us for further information.

Qualification Structure and Requirements

Credits and Total Qualification Time (TQT)

The QUALIFI [enter qualification title] is made up of [enter credit value] credits which equates to hours [enter TQT value] of TQT.

Total Qualification Time (TQT) is an estimate of the total amount of time that could reasonably be expected to be required for a learner to achieve and demonstrate the achievement of the level of attainment necessary for the award of a qualification.

Examples of activities that can contribute to Total Qualification Time includes: guided learning, independent and unsupervised research/learning, unsupervised compilation of a portfolio of work experience, unsupervised e-learning, unsupervised e-assessment, unsupervised coursework, watching a prerecorded podcast or webinar, unsupervised work-based learning.

Guided Learning Hours (GLH) are defined as the time when a tutor is present to give specific guidance towards the learning aim being studied on a programme. This definition includes lectures, tutorials and supervised study in, for example, open learning centres and learning workshops, live webinars, telephone tutorials or other forms of e-learning supervised by a tutor in real time. Guided learning includes any supervised assessment activity; this includes invigilated examination and observed assessment and observed work-based practice.

Rules of Combination for QUALIFI Level 3 Diploma in Data Science

The QUALIFI Level 3 Diploma in Data Science comprises fourteen mandatory units. All units cover a number of topics relating to learning outcomes. Al units are mandatory.

Unit Reference	Unit	Level	TQT	Credit	GLH
H/650/4951	The Field of Data Science	3	60	6	45
J/650/4952	Python for Data Science	3	90	9	68
K/650/4953	Creating and Interpreting Visualisations in Data Science	3	30	3	23
L/650/4954	Data and Descriptive Statistics in Data Science	3	60	6	45
M/650/4955	Fundamentals of Data Analytics	3	30	3	23
R/650/4956	Data Analytics with Python	3	30	3	23
T/650/4957	Machine Learning Methods and Models in Data Science	3	30	3	23
Y/650/4958	The Machine Learning Process	3	30	3	23
A/650/4959	Linear Regression in Data Science	3	30	3	23
H/650/4960	Logistic Regression in Data Science	3	30	3	23
J/650/4961	Decision Trees in Data Science	3	30	3	23
K/650/4962	K-means Clustering in Data Science	3	30	3	23
L/650/4963	Synthetic Data for Privacy and Security in Data Science	3	60	6	45
M/650/4964	Graphs and Graph Data Science	3	60	6	45
		Total	**600**	**60**	**455**

Achievement Requirements

Learners must demonstrate they have met all learning outcomes and assessment criteria for all the required units to achieve this qualification. QUALIFI will issue certificates to all successful learners via their registered centres.

Awarding Classification/Grading

This qualification grading is **Pass/Merit/Distinction.**

Fail: 0 – 49%
Pass: 50 – 59%
Merit: 60% - 69%
Distinction: 70% and over.

All units will be internally assessed through written assignment, internally marked by the QUALIFI approved centre and subject to external quality assurance by QUALIFI.

Assessment Strategy and Methods

To demonstrate all learning outcomes and assessment criteria, each unit will be assessed formatively, i.e., assignments focusing on knowledge and understanding of technical skills using sample data.

These tasks will address all learning outcomes and related assessment criteria, all of which must be demonstrated/passed in order to achieve the qualification.

In addition, learners will need to demonstrate their knowledge, understanding, original thought, and problem-solving skills where appropriate. Intellectual rigour will be expected that is appropriate to the level of the qualification.

The summative assignments will contain a question strand for each of the given unit's learning outcomes. The assignment tasks will address the LO (learning outcome) and AC (assessment criteria) requirements. Within assignments, there will always be requirements for learners to engage with important and relevant theory that underpins the subject area.

Evidence of both formative and summative assessment MUST be made available at the time of external quality assurance – EQA.

Unit Specifications

Unit 3DS01: The Field of Data Science

Unit code: H/650/4951

RQF Level: 3

Unit Aims

This unit introduces learners to the field of data science from the birth of artificial intelligence and machine learning in the late 1950s to the dawn of the "big data" era in the early 2000s, to the current applications of AI, machine learning and deep learning and the various challenges associated with them.

Learning Outcomes and Assessment Criteria

Learning Outcomes: **To achieve this unit, the learner must be able to:**	**Assessment Criteria:** **Assessment of these learning outcomes will require a learner to demonstrate that they can:**
1. Understand the core issues of data science.	1.1 Explain what is meant by the terms "data science" and "data scientist". 1.2 Explain how data science relates to other academic fields. 1.3 Analyse the features, uses, benefits and drawbacks of tools and software commonly used by data scientists.
2. Understand the core issues of data and big data.	2.1 Explain what is meant by "big data". 2.2 Analyse the challenges and criticisms of "big data". 2.3 Analyse the successes and two failures of "big data". 2.4 Analyse the features, uses, benefits and drawbacks of the tools and software commonly used to process and analyse "big data".
3. Understand the core issues of artificial intelligence.	3.1 Explain what is meant by the term "artificial intelligence". 3.2 Explain the difference between the terms "artificial narrow intelligence", "artificial general intelligence" and "artificial super intelligence" 3.3 Analyse the challenges in achieving artificial intelligence. 3.4 Analyse the successes and failures of artificial intelligence.
4. Understand the core issues of machine learning.	4.1 Explain what is meant by the term "machine learning".

	4.2 Explain the main types of machine learning: "supervised", "unsupervised" and "reinforcement learning". 4.3 Analyse the uses and limitations of machine learning. 4.4 Explain the difference between artificial intelligence and machine learning.
5. Understand the core issues of deep learning.	5.1 Explain what is meant by the term "deep learning". 5.2 Explain basic deep learning architecture. 5.3 Analyse the uses and limitations of deep learning. 5.4 Analyse current areas of research in deep learning.

Indicative Content

- Data science
- Big Data
- Google Flu Trends
- Python
- Hadoop and Spark
- Artificial intelligence
- Artificial narrow intelligence
- Artificial general intelligence
- Artificial super intelligence
- Machine learning
- Supervised machine learning
- Unsupervised machine learning
- Reinforcement learning
- Deep learning

Recommended Texts and readings

Jiawei Han, Micheline Kamber and Jian Pei, "*Data Mining: Concepts and Techniques*", Third Edition, 2012

Delivery Guidance

The Level 3 Diploma in Data Science can be delivered:

i) Via distance learning, with all learning materials and assessments available to learners on-line. Support to students is provided via email, or via tools such as Zoom for one-to-one feedback and support.

ii) Via a classroom-based environment, typically taught as 6 hours per week over three terms of 10-week semesters, by data science tutors and supported by data science teaching assistants.

In both cases, learners are provided with detailed core learning materials for each of the fourteen units, and supplementary materials as appropriate, including PDF of lecture slides, Question & Answer bank booklets, and sample code.

Assessment Guidance

To demonstrate all learning outcomes and assessment criteria, each unit will be assessed by a single summative assessment (i.e., an assessment taken after the learner has completed the learning and study for the unit) designed to assess learner's technical knowledge and understanding of the unit's learning outcomes.

Each summative assessment will contain a question for each of the given unit's learning outcomes. The assignment tasks will address the LO (learning outcome) and AC (assessment criteria) requirements. Within assessments there will always be requirements for learners to engage with important and relevant theory that underpins the subject area. Learners will also be given data

A sample assessment with model solutions should be made available to learners.

Suggested Resources

Peter Lake and Robert Drake, "*Information Systems Management in the Big Data Era*", Springer, 2015

Unit 3DS02: Python for Data Science

Unit code: J/650/4952

RQF Level: 3

Unit Aims

This unit provides learners with an introduction to Python programming for data science. The unit assumes no prior knowledge of coding or of Python and so starts by explaining the basics of Python, its design philosophy, syntax, naming conventions and coding standards.

The unit then introduces the basic Python data types of integers, floats, strings, complex numbers and booleans and explains how these data types can be created, changed, manipulated, and calculated using standard mathematical functions, logical operators, and Python's built-in methods and functions. The unit also introduces more complex data structures critical to many data analytics and data science tasks, such as "lists", "tuples", "sets", and "dictionaries".

The unit explains how to use control and flow statements such as branching and looping as well as the basics of writing user-defined Python functions – all the ingredients needed to later perform data analysis and to code data science models successfully.

Learning Outcomes and Assessment Criteria

Learning Outcomes: **To achieve this unit, the learner must be able to:**	**Assessment Criteria:** **Assessment of these learning outcomes will require a learner to demonstrate that they can:**
1. Understand the design philosophy and features of Python.	1.1 Explain what is meant by Python being a "*high-level, interpreted, dynamically-typed, general-purpose language*." 1.2 Analyse the features, uses, benefits and drawbacks of programming languages such as C++ and R. 1.3 Explain Python's syntax, indentation, naming conventions and coding standards.
2. Understand Python's basic data types.	2.1 Explain the basic Python data types: strings, integers, floats, complex numbers, and Booleans. 2.2 Use arithmetical operators and standard mathematical functions correctly to perform basic calculations. 2.3 Use the logical, bitwise and identity operators

	correctly to perform logical operations. 2.4 Explain the order of operator precedence. 2.5 Use string methods and functions correctly to create new strings or to retrieve values and properties. 2.6 Obtain string elements correctly by indexing and slicing.
3. Be able to create and manipulate lists and tuples.	3.1 Explain the difference between a “list” and a “tuple”. 3.2 Use list and tuple methods and functions correctly to update or to retrieve values and properties. 3.3 Obtain list and tuples elements correctly by indexing and slicing.
4. Be able to create and manipulate sets and dictionaries.	4.1 Explain the difference between “sets” and “dictionaries”. 4.2 Use set and dictionary methods and functions correctly to update or to retrieve values and properties.
5. Be able to write Python functions and flow statements.	5.1 Construct and use correctly various control flow statements: - Conditional statements. - Transfer statements. - Iterative statements. 5.2 Create “def” and “lambda” functions correctly, passing parameters and returning values. 5.3 Explain the difference between “keyword”, “positional” and “optional” parameters.

Indicative Content

- Python environments
- Anaconda
- Basic data types, i.e., strings, integers, floats, complex number and booleans
- Numerical operations
- Logical operations
- String methods
- String indexing and slicing
- If-Else statements
- For loops
- While loops
- Lists, tuples, sets and dictionaries
- Python def functions and Lambda functions

Recommended Text

Michael Dawson, "*Python Programming for the absolute beginner*", Third Edition, 2005

Delivery Guidance

The Level 3 Diploma in Data Science can be delivered:

i) Via distance learning, with all learning materials and assessments available to learners on-line. Support to students is provided via email, or via tools such as Zoom for one-to-one feedback and support.

ii) Via a classroom-based environment, typically taught as 6 hours per week over three terms of 10-week semesters, by data science tutors and supported by data science teaching assistants.

In both cases, learners are provided with detailed core learning materials for each of the fourteen units, and supplementary materials as appropriate, including PDF of lecture slides, Question & Answer bank booklets, and sample code.

Assessment Guidance

To demonstrate all learning outcomes and assessment criteria, each unit will be assessed by a single summative assessment (i.e., an assessment taken after the learner has completed the learning and study for the unit) designed to assess learner's technical knowledge and understanding of the unit's learning outcomes.

Each summative assessment will contain a question for each of the given unit's learning outcomes. The assignment tasks will address the LO (learning outcome) and AC (assessment criteria) requirements. Within assessments there will always be requirements for learners to engage with important and relevant theory that underpins the subject area. Learners will also be given data

A sample assessment with model solutions should be made available to learners.

Suggested Resources

Luciano Ramalho, "Fluent Python", O'Reilly, 2005

Unit 3DS03: Creating and Interpreting Visualisations in data science

Unit code: K/650/4953

RQF Level: 3

Unit Aims

This unit introduces the learner to basic charts and visualisations and how to create and interpret them. The unit starts by explaining why visualisations are critical when understanding data and what makes a good and a poor visualisation.

The unit introduces learners to a number of basic chart and plot types, explaining their purpose, how to interpret them and explains when they should and should not be used. The unit then focuses on the technology used to produce charts and visualisations in Python, using Seaborn, Matplotlib and other Python libraries.

Learning Outcomes and Assessment Criteria

Learning Outcomes: **To achieve this unit, the learner must be able to:**	**Assessment Criteria:** **Assessment of these learning outcomes will require a learner to demonstrate that they can:**
1. Understand the role and importance of visualising data.	1.1 Explain the role and importance of visualising data before conducting data analysis. 1.2 Explain why poorly created visualisations can be misleading. 1.3 Explain good practices when creating plots and charts.
2. Understand basic plots and charts.	2.1 Define the basic chart and plot types: - Scatter plots - Line charts - Pie charts - Bar and column charts - Histogram and density curves - Box-and-whisker plots 2.2 Explain the advantages and disadvantages of each chart type. 2.3 Explain which chart types should be used for different types of data.
3. Be able to create and interpret plots and charts.	3.1 Analyse the features, uses, benefits and drawback of Python libraries for constructing charts and visualisations. 3.2 Write Python code correctly to construct, format

	and display the charts and plots: - Scatter plots - Line charts - Pie charts - Bar and column charts - Histogram and density curves - Box-and-whisker plots 3.3 Interpret correctly the charts produced from 3.2

Indicative Content

- Anscombe's quartet
- Scatter plots
- Line charts
- Pie charts
- Bar and column charts
- Histogram and density curves
- Box-and-whisker plots
- Matplotlib
- Seaborn

Recommended Text

Igor Milovanovic, Dimitry Foures and Giuseppe Vettigli, "*Python Data Visualization Cookbook - Second Edition*", 2015

Delivery Guidance

The Level 3 Diploma in Data Science can be delivered:

i) Via distance learning, with all learning materials and assessments available to learners on-line. Support to students is provided via email, or via tools such as Zoom for one-to-one feedback and support.

ii) Via a classroom-based environment, typically taught as 6 hours per week over three terms of 10-week semesters, by data science tutors and supported by data science teaching assistants.

In both cases, learners are provided with detailed core learning materials for each of the fourteen units, and supplementary materials as appropriate, including PDF of lecture slides, Question & Answer bank booklets, and sample code.

Assessment Guidance

To demonstrate all learning outcomes and assessment criteria, each unit will be assessed by a single summative assessment (i.e., an assessment taken after the learner has completed the learning and study for the unit) designed to assess learner's technical knowledge and understanding of the unit's learning outcomes.

Each summative assessment will contain a question for each of the given unit's learning outcomes. The assignment tasks will address the LO (learning outcome) and AC (assessment criteria) requirements. Within assessments there will always be requirements for learners to engage with important and relevant theory that underpins the subject area. Learners will also be given data

A sample assessment with model solutions should be made available to learners.

Suggested Resources

Jiawei Han, Micheline Kamber and Jian Pei, "*Data Mining: Concepts and Techniques*", Third Edition, 2012

Unit 3DS04: Data and Descriptive Statistics in Data Science

Unit code: L/650/4954

RQF Level: 3

Unit Aims

With modern software, packages, and programming languages, it is too easy for aspiring data scientists to rely on these tools to calculate descriptive statistics for them. It is critical for the modern data scientist to not only be able to interpret descriptive statistics, but also understand them and know how they are calculated. A lack of knowledge and the inability to interpret statistics correctly often leads to erroneous decisions being made which can have serious negative consequences.

This unit aims to provide learners with an introduction to descriptive statistics and methods which are key for data analysis and data science. This unit introduces different types of data and descriptive statistics from measures of centre, various measures of spread (including range, percentiles, variance and standard deviation), measures of symmetry (skewness and kurtosis) and measures of joint variability (correlation and covariance). The unit also explains which descriptive statistics can be calculated for the data measured on different scales. In this unit, learners will gain first-hand experience and practice of calculating descriptive statistics for small data sets manually.

Learning Outcomes and Assessment Criteria

Learning Outcomes: **To achieve this unit, the learner must be able to:**	**Assessment Criteria:** **Assessment of these learning outcomes will require a learner to demonstrate that they can:**
1. Understand the different types of data and their characteristics.	1.1 Explain the differences between data measured on "nominal", "ordinal", "interval" and "ratio" scales. 1.2 Explain the difference between "discrete" and "continuous" data.
2. Understand measures of centre.	2.1 Define the mathematical formulas for: - The arithmetic, geometric and harmonic means. - The mode. - The median. 2.2 Explain the relationship between the arithmetic, geometric and harmonic means. 2.3 Calculate the measures of centre correctly for a dataset. 2.4 Interpret calculated measures of centre and draw reasoned conclusions. 2.5 Explain which measures of centre apply to different

	types of data.
3. Understand measures of spread.	3.1 Define the mathematical formulas for: - The range. - Percentiles, deciles, and quartiles. - The Interquartile range. - The variance and the standard deviation. - The coefficient of variation. 3.2 Calculate the measures of spread correctly for a dataset. 3.3 Interpret calculated measures of spread and draw reasoned conclusions. 3.4 Explain which measures of spread apply to different types of data.
4. Understand measures of symmetry and peakness.	4.1 Define the mathematical formulas for: - Skewness - Kurtosis 4.2 Explain the terms positively skewed and negatively skewed. 4.3 Calculate the skewness and kurtosis correctly for a given dataset. 4.4 Interpret the calculated skewness and kurtosis and draw reasoned conclusions.
5. Understand measures of joint variability and linear relation.	5.1 Explain the definitions and mathematical formulas for: - The covariance between two numerical variables. - The Pearson correlation coefficient between two numerical variables. 5.2 Calculate the covariance and correlation coefficient between two numeric variables correctly. 5.3 Interpret the calculated covariance and correlation coefficient and draw reasoned conclusions. 5.4 Explain the difference between correlation and causality.

Indicative Content

- Nominal data
- Ordinal data
- Interval data
- Ratio data
- Discrete and continuous data
- Arithmetic, Geometric and Harmonic means.
- Mode
- Median
- Range
- Percentiles, deciles, and quartiles
- Interquartile range
- Variance

- Standard deviation
- Coefficient of variation
- Skewness
- Kurtosis
- Covariance
- Correlation coefficient
- Correlation and causality

Recommended Text

Jiawei Han, Micheline Kamber and Jian Pei, "*Data Mining: Concepts and Techniques*", Third Edition, 2012

Delivery Guidance

The Level 3 Diploma in Data Science can be delivered:

i) Via distance learning, with all learning materials and assessments available to learners on-line. Support to students is provided via email, or via tools such as Zoom for one-to-one feedback and support.

ii) Via a classroom-based environment, typically taught as 6 hours per week over three terms of 10-week semesters, by data science tutors and supported by data science teaching assistants.

In both cases, learners are provided with detailed core learning materials for each of the fourteen units, and supplementary materials as appropriate, including PDF of lecture slides, Question & Answer bank booklets, and sample code.

Assessment Guidance

To demonstrate all learning outcomes and assessment criteria, each unit will be assessed by a single summative assessment (i.e., an assessment taken after the learner has completed the learning and study for the unit) designed to assess learner's technical knowledge and understanding of the unit's learning outcomes.

Each summative assessment will contain a question for each of the given unit's learning outcomes. The assignment tasks will address the LO (learning outcome) and AC (assessment criteria) requirements. Within assessments there will always be requirements for learners to engage with important and relevant theory that underpins the subject area. Learners will also be given data

A sample assessment with model solutions should be made available to learners.

Suggested Resources

Peter Bruce, Andrew Bruce, and Peter Gedeck, "*Practical Statistics for Data Scientists*", 2020.

Unit 3DS05: Fundamentals of Data Analytics

Unit code: M/650/4955

RQF Level: 3

Unit Aims

This unit serves as the introduction to the core concepts of data analytics.

The unit will help learners to differentiate between the roles of a Data Analyst, Data Scientist and Data Engineer. Learners will also be able to summarize the data ecosystem such as databases and data warehouses and learn about major vendors within the data ecosystem and explore the various tools.

The unit also introduces learners to the fundamental tasks and processes in the data discovery process such as data cleaning, methods for dealing with data quality and methods for standardising data ready for analysis.

Learning Outcomes and Assessment Criteria

Learning Outcomes: **To achieve this unit, the learner must be able to:**	**Assessment Criteria:** **Assessment of these learning outcomes will require a learner to demonstrate that they can:**
1. Understand the processes and types of data analytics.	1.1 Explain the "Knowledge Discovery from Data" process. 1.2 Explain the different types of data analytics: "descriptive", "predictive", and "prescriptive". 1.3 Explain the differences between the roles: "data engineer", "data analyst", "data scientist" and "business intelligence analyst".
2. Understand the data analytics ecosystem.	2.1 Analyse the features, uses, benefits and drawbacks of different types of data format: CSV, JavaScript Object Notation (JSON), Excel, text, audio, and images. 2.2 Explain the difference between relational and non-relational databases. 2.3 Analyse the features, uses, benefits and drawbacks of common software tools used for data analytics.
3. Understand the issues and methods for dealing with data quality issues.	3.1 Explain the strategies for identifying and dealing with: - Missing data. - Duplicate data. - Inconsistent data. - Outliers.

	3.2 Analyse the features, uses, benefits and drawbacks of the “mean”, “median”, and “mode” strategies for data imputation.
4. Understand the issues and methods of basic data transformations.	4.1 Explain the purpose of data transformation strategies: smoothing; feature engineering; aggregation; normalization; discretization. 4.2 Explain the definitions and mathematical formulas for: - Min-max normalization - Z-score normalization 4.3 Explain the difference between min-max and z-score normalization. 4.4 Explain how binning can be used to smooth data and to discretize data.

Indicative Content

- Knowledge Discovery from Data process
- Descriptive analytics
- Predictive analytics
- Prescriptive analytics
- Data formats
- Relational databases
- Non-relational databases
- Data quality
- Data imputation
- Data transformations
- Data aggregation
- Feature engineering
- Min-max normalisation
- Z-score normalisation

Recommended Text

Jiawei Han, Micheline Kamber and Jian Pei, “*Data Mining: Concepts and Techniques*”, Third Edition, 2012

Delivery Guidance

The Level 3 Diploma in Data Science can be delivered:

i) Via distance learning, with all learning materials and assessments available to learners on-line. Support to students is provided via email, or via tools such as Zoom for one-to-one feedback and support.

ii) Via a classroom-based environment, typically taught as 6 hours per week over three terms of 10-week semesters, by data science tutors and supported by data science teaching assistants.

In both cases, learners are provided with detailed core learning materials for each of the fourteen units, and supplementary materials as appropriate, including PDF of lecture slides, Question & Answer bank booklets, and sample code.

Assessment Guidance

To demonstrate all learning outcomes and assessment criteria, each unit will be assessed by a single summative assessment (i.e., an assessment taken after the learner has completed the learning and study for the unit) designed to assess learner's technical knowledge and understanding of the unit's learning outcomes.

Each summative assessment will contain a question for each of the given unit's learning outcomes. The assignment tasks will address the LO (learning outcome) and AC (assessment criteria) requirements. Within assessments there will always be requirements for learners to engage with important and relevant theory that underpins the subject area. Learners will also be given data

A sample assessment with model solutions should be made available to learners.

Suggested Resources

1. Charu C. Aggarwal, "*Data Mining: The Textbook*", 2015

2. S. Sumathi and S.N. Sivanandam, "*Introduction to Data Mining and its Applications*", Springer Science & Business Media, 2006.

Unit 3DS06: Data Analysis with Python

Unit code: R/650/4956

RQF Level: 3

Unit Aims

This unit introduces basic data analysis with Python. Learners are introduced to core concepts such as Pandas DataFrames and Series, merging and joining data.

This unit also builds on previous units by teaching how to import data, using Python to create descriptive statistics for analysis and interpretation. The unit also teaches learners how to use Python when preparing data for machine learning models by improving data quality and standardising data.

Learning Outcomes and Assessment Criteria

Learning Outcomes: **To achieve this unit, the learner must be able to:**	**Assessment Criteria:** **Assessment of these learning outcomes will require a learner to demonstrate that they can:**
1. Be able to load and save data	1.1 Create Pandas DataFrames and Series correctly from different data sources and file types. 1.2 Save a Pandas DataFrame correctly as a CSV, Excel, or JSON file.
2. Be able to perform basic data wrangling and exploratory analysis.	2.1 Use in-built Python functions or user defined functions to: - Drop and reorder unwanted columns in a DataFrame - Create new rows or columns in a DataFrame - Rename column headings in a DataFrame - Select subsets of data from DataFrames based on conditions. - Merge and concatenate multiple DataFrames. - Create descriptive statistics for a dataset. - Create visualisations that are appropriate for the given data type
3. Be able to perform basic data cleaning tasks.	3.1 Improve a dataset's quality by identifying and dealing with: - Missing values. - Duplicate data. - Inconsistent values. - Outliers.

4. Be able to perform basic data transformation tasks.	4.1 Create new features correctly from existing data. 4.2 Discretize data correctly by applying equal-width and equal-frequency binning. 4.3 Normalize data correctly using: - Min-max normalization. - Z-score normalization.

Indicative Content

- Pandas
- Pandas DataFrame
- Pandas Series
- Numpy
- Numpy arrays
- Loading and saving data from files
- Data wrangling
- Data cleaning
- Seaborn
- Matplotlib

Recommended Text

Jiawei Han, Micheline Kamber and Jian Pei, "*Data Mining: Concepts and Techniques*", Third Edition, 2012

Delivery Guidance

The Level 3 Diploma in Data Science can be delivered:

i) Via distance learning, with all learning materials and assessments available to learners on-line. Support to students is provided via email, or via tools such as Zoom for one-to-one feedback and support.

ii) Via a classroom-based environment, typically taught as 6 hours per week over three terms of 10-week semesters, by data science tutors and supported by data science teaching assistants.

In both cases, learners are provided with detailed core learning materials for each of the fourteen units, and supplementary materials as appropriate, including PDF of lecture slides, Question & Answer bank booklets, and sample code.

Assessment Guidance

To demonstrate all learning outcomes and assessment criteria, each unit will be assessed by a single summative assessment (i.e., an assessment taken after the learner has completed the learning and study for the unit) designed to assess learner's technical knowledge and understanding of the unit's learning outcomes.

Each summative assessment will contain a question for each of the given unit's learning outcomes. The assignment tasks will address the LO (learning outcome) and AC (assessment criteria) requirements. Within assessments there will always be requirements for learners to engage with important and relevant theory that underpins the subject area. Learners will also be given data

A sample assessment with model solutions should be made available to learners.

Suggested Resources

1. Wes Mckinney, "*Python for Data Analysis, 2e: Data Wrangling with Pandas, Numpy, and Ipython*", 2017

2. A.J. Henley and Dave Wolf, "*Learn Data Analysis with Python*", Apress, 2018

Unit 3DS07: Machine Learning Methods and Models in Data Science

Unit code: T/650/4957

RQF Level: 3

Unit Aims

This unit provides a high-level overview (rather than a deep dive) of the three main types of machine learning: supervised learning, unsupervised learning, and reinforcement learning. The unit discusses the use-cases and real-world problems the various methods can be applied to, summarises the key-features of the different methods, as well as the challenges of each method.

Learning Outcomes and Assessment Criteria

Learning Outcomes: **To achieve this unit, the learner must be able to:**	**Assessment Criteria:** **Assessment of these learning outcomes will require a learner to demonstrate that they can:**
1. Understand the concepts of basic supervised machine learning models	1.1 Explain the features and objectives of: - Linear regression - Logistic regression - Decision trees and random forests - Support Vector Machines - K-Nearest Neighbour 1.2 Analyse the challenges of supervised models. 1.3 Explain the types of use-cases to which supervised Machine Learning models can be applied.
2. Understand the concepts of basic unsupervised machine learning models	2.1 Explain the features and objectives of: - Clustering - Association rules - Dimensionality reduction 2.2 Analyse the challenges of unsupervised models. 2.3 Explain the types of use-cases to which unsupervised Machine Learning models can be applied.
3. Understand the concepts of basic reinforcement learning.	3.1 Explain the features and objective of reinforcement learning. 3.2 Analyse the challenges of reinforcement learning. 3.3 Explain a use-case to which reinforcement learning can be applied.

Indicative Content

- Supervised machine learning
- Linear regression
- Logistic regression
- Decision trees and random forests
- Support Vector Machines
- K-Nearest Neighbour
- Unsupervised machine learning
- Clustering
- Association rules
- Dimensionality reduction
- Reinforcement learning

Recommended Texts

Jiawei Han, Micheline Kamber and Jian Pei, "*Data Mining: Concepts and Techniques*", Third Edition, 2012

Delivery Guidance

The Level 3 Diploma in Data Science can be delivered:

i) Via distance learning, with all learning materials and assessments available to learners on-line. Support to students is provided via email, or via tools such as Zoom for one-to-one feedback and support.

ii) Via a classroom-based environment, typically taught as 6 hours per week over three terms of 10-week semesters, by data science tutors and supported by data science teaching assistants.

In both cases, learners are provided with detailed core learning materials for each of the fourteen units, and supplementary materials as appropriate, including PDF of lecture slides, Question & Answer bank booklets, and sample code.

Assessment Guidance

To demonstrate all learning outcomes and assessment criteria, each unit will be assessed by a single summative assessment (i.e., an assessment taken after the learner has completed the learning and study for the unit) designed to assess learner's technical knowledge and understanding of the unit's learning outcomes.

Each summative assessment will contain a question for each of the given unit's learning outcomes. The assignment tasks will address the LO (learning outcome) and AC (assessment criteria) requirements. Within assessments there will always be requirements for learners to engage with important and relevant theory that underpins the subject area. Learners will also be given data

A sample assessment with model solutions should be made available to learners.

Suggested Resources

1. Peter A. Flach, "*Machine Learning - The Art and Science of Algorithms that Make Sense of Data*", 2012

2. John D. Kelleher, Brian Mac Namee and Aoife D'Arcy, "*Fundamentals of Machine Learning for Predictive Data Analytics*", Second Edition, 2020

3. Phil Winder, "*Reinforcement Learning: Industrial Applications of Intelligent Agents*", 2020

Unit 3DS08: The Machine Learning Process

Unit code: Y/650/4958

RQF Level: 3

Unit Aims

This unit introduces the many steps and processes involved when building and evaluating machine learning models.

The unit explains the core elements of the machine learning process from how to prepare data to selecting the correct machine learning algorithm to the importance of splitting data into training, test, and validation datasets to avoid the pitfalls of under and overfitting. The unit also covers how to identify and correct class imbalance and discusses when such approaches are needed.

Many of the machine learning models that are encountered are supervised classification models and so the unit introduces the common performance metrics as well as how to interpret them. Finally, the unit discusses briefly how to deal with model bias and variance.

Learning Outcomes and Assessment Criteria

Learning Outcomes: **To achieve this unit, the learner must be able to:**	**Assessment Criteria:** **Assessment of these learning outcomes will require a learner to demonstrate that they can:**
1. Understand the machine learning process.	1.1 Analyse the components of the machine learning process: data collection; data preparation; selecting the machine learning algorithm; training and testing models; parameter tuning; deploying a model. 1.2 Explain the difficulties and solutions for each component of the machine learning process.
2. Understand the data preparation process for machine learning models.	2.1 Analyse the data requirements for different machine learning models. 2.2 Explain how to convert categorical data to numerical values. 2.3 Explain why “class imbalance” can be dangerous for models. 2.4 Analyse the features, uses, benefits and drawbacks of the strategies for balancing classes: - Over sample the minority class - Under sample the majority class

	2.5 Explain the purpose of splitting data into training, test, and validation subsets.
3. Understand how to evaluate machine learning models.	3.1 Explain what is meant by “a confusion matrix”. 3.2 Define the classification metrics: “Precision”, “Accuracy”, “Recall”, “Support”, and “F1”. 3.3 Explain what is meant by a “Receiver Operating Characteristic Curve (ROC)”, and the “Area under the ROC curve” (AUC). 3.4 Explain the difficulties with assessing unsupervised machine learning models.
4. Be able to evaluate classification models.	4.1 Calculate the classification metrics correctly from a confusion matrix. 4.2 Interpret a ROC curve and AUC and make reasoned conclusions.
5. Understand the issues of bias and variance in models.	5.1 Explain what is meant by “overfitting” and “underfitting”. 5.2 Analyse the features, uses, benefits and drawbacks of the methods to prevent overfitting: cross validation; removing features; bagging; boosting; early stopping.

Indicative Content

- The machine learning process
- Git and version control
- Class imbalance and balancing classes via over and under sampling
- Confusion matrix
- Precision
- Accuracy
- Recall
- Support
- F1
- Receiver Operating Characteristic Curve (ROC)
- Area under the ROC curve (AUC)
- Overfitting and underfitting
- Model bias.
- Bagging
- Boosting

Recommended Text

Jiawei Han, Micheline Kamber and Jian Pei, “*Data Mining: Concepts and Techniques*”, Third Edition, 2012.

Delivery Guidance

The Level 3 Diploma in Data Science can be delivered:

i) Via distance learning, with all learning materials and assessments available to learners on-line. Support to students is provided via email, or via tools such as Zoom for one-to-one feedback and support.

ii) Via a classroom-based environment, typically taught as 6 hours per week over three terms of 10-week semesters, by data science tutors and supported by data science teaching assistants.

In both cases, learners are provided with detailed core learning materials for each of the fourteen units, and supplementary materials as appropriate, including PDF of lecture slides, Question & Answer bank booklets, and sample code.

Assessment Guidance

To demonstrate all learning outcomes and assessment criteria, each unit will be assessed by a single summative assessment (i.e., an assessment taken after the learner has completed the learning and study for the unit) designed to assess learner's technical knowledge and understanding of the unit's learning outcomes.

Each summative assessment will contain a question for each of the given unit's learning outcomes. The assignment tasks will address the LO (learning outcome) and AC (assessment criteria) requirements. Within assessments there will always be requirements for learners to engage with important and relevant theory that underpins the subject area. Learners will also be given data

A sample assessment with model solutions should be made available to learners.

Suggested Resources

1. Peter A. Flach, "*Machine Learning - The Art and Science of Algorithms that Make Sense of Data*", 2012

2. John D. Kelleher, Brian Mac Namee and Aoife D'Arcy, "*Fundamentals of Machine Learning for Predictive Data Analytics*", Second Edition, 2020

3. Sebastian Raschka, "*Python Machine Learning*", PACKT, 2015

Unit 3DS09: Linear Regression in Data Science

Unit code: A/650/4959

RQF Level: 3

Unit Aims

This unit introduces the basic theory of simple linear regression models that are critical to the ability to predict the value of one continuous variable based on the value of another. Learners will be able to estimate the line of best-fit by calculating the regression parameters and understand the accuracy of the line of best-fit.

The unit also introduces extensions to simple linear regression by introducing multiple and polynomial regression models to examine relationships between multiple variables. The unit explains how to build simple, multiple, and polynomial linear regression models using Python and libraries such as scikit-learn.

Learning Outcomes and Assessment Criteria

Learning Outcomes: **To achieve this unit, the learner must be able to:**	**Assessment Criteria:** **Assessment of these learning outcomes will require a learner to demonstrate that they can:**
1. Understand the basic theory of linear regression.	1.1 Explain what is meant by simple, multiple, and polynomial linear regression. 1.2 Analyse the assumptions of linear regression. 1.3 Explain the Ordinary Least Squares method for estimating the parameters in simple linear regression. 1.4 State the formulas used to calculate the intercept and slope coefficient in simple linear regression. 1.5 Explain the use-cases for linear regression models. 1.6 Analyse the benefits and limitations of regression models.
2. Understand regression metrics and how to evaluate a regression model.	2.1 Explain the regression metrics: - The Total Sum of Squares (TSS) - The Residual Sum of Squares (RSS) - The Explained Sum of Squares (ESS) - The Mean Squared Error (MSE) - The Root Mean Square Error (RMSE) - The coefficient of determination (R^2). - The Adjusted R^2 2.2 Explain how to interpret each of the regression

	metrics listed in 2.1.
3. Be able to perform regression calculations and analysis.	3.1 Calculate correctly the intercept and slope coefficient in simple linear regression. 3.2 Calculate correctly the regression metrics in a linear regression model. 3.3 Interpret the calculated metrics and draw reasoned conclusions.
4. Be able to create linear regression models.	4.1 Use Python to build accurate simple linear regression and multiple linear regression models for given datasets. 4.2 Use Python to evaluate the accuracy of the models built in 4.1. and analyse the results.

Indicative Content

- Simple linear regression
- Multiple and polynomial regression
- Ordinary Least Squares
- The Total Sum of Squares (TSS)
- The Residual Sum of Squares (RSS)
- The Explained Sum of Squares (ESS)
- The Mean Squared Error (MSE)
- The Root Mean Square Error (RMSE)
- The coefficient of determination (R^2).
- The Adjusted R^2
- Python
- Sklearn linear_model.LinearRegression
- Seaborn
- Scatter plot
- Line chart

Recommended Text

Giuseppe Bonaccorso, "*Machine Learning Algorithms: Popular algorithms for data science and machine learning, 2nd Edition*", Packt, 2018

Delivery Guidance

The Level 3 Diploma in Data Science can be delivered:

i) Via distance learning, with all learning materials and assessments available to learners on-line. Support to students is provided via email, or via tools such as Zoom for one-to-one feedback and support.

ii) Via a classroom-based environment, typically taught as 6 hours per week over three terms of 10-week semesters, by data science tutors and supported by data science teaching assistants.

In both cases, learners are provided with detailed core learning materials for each of the fourteen units, and supplementary materials as appropriate, including PDF of lecture slides, Question & Answer bank booklets, and sample code.

Assessment Guidance

To demonstrate all learning outcomes and assessment criteria, each unit will be assessed by a single summative assessment (i.e., an assessment taken after the learner has completed the learning and study for the unit) designed to assess learner's technical knowledge and understanding of the unit's learning outcomes.

Each summative assessment will contain a question for each of the given unit's learning outcomes. The assignment tasks will address the LO (learning outcome) and AC (assessment criteria) requirements. Within assessments there will always be requirements for learners to engage with important and relevant theory that underpins the subject area. Learners will also be given data

A sample assessment with model solutions should be made available to learners.

Suggested Resources

1. Jiawei Han, Micheline Kamber and Jian Pei, "*Data Mining: Concepts and Techniques*", Third Edition, 2012

2. John D. Kelleher, Brian Mac Namee and Aoife D'Arcy, "*Fundamentals of Machine Learning for Predictive Data Analytics*", Second Edition, 2020

3. Sebastian Raschka, "*Python Machine Learning*", PACKT, 2015

Unit 3DS10: Logistic Regression in Data Science

Unit code: H/650/4960

RQF Level: 3

Unit Aims

This unit introduces logistic regression and its application as a classification algorithm. The unit explores the basics of binary logistic regression via the logistic function, the Odds ratio, and the Logit function. The unit also explains the differences between linear and logistic regression. Learners will learn how to build and visualise a logistic regression model using Python.

The unit will teach learners when it is relevant to choose logistic regression over linear regression, how to interpret the results of logistic regression correctly and how to choose the best logistic model that describes the relationship under question.

Learning Outcomes and Assessment Criteria

Learning Outcomes: **To achieve this unit, the learner must be able to:**	**Assessment Criteria:** **Assessment of these learning outcomes will require a learner to demonstrate that they can:**
1. Understand the basic theory of logistic regression.	1.1 Explain what is meant by binary logistic regression and the difference between linear and logistic regression. 1.2 Analyse the assumptions for logistic regression. 1.3 Define the Logistic function, the Odds ratio, and the Logit function. 1.4 State basic characteristics and properties of the Logistic function, the Odds ratio and Logit function. 1.5 Explain how to interpret the Odds ratio. 1.6 Analyse the benefits and limitations of logistic regression. 1.7 Explain how Logistic regression can be applied to multiple-class problems.
2. Be able to perform logistic regression calculations.	2.1 Calculate correctly the probability values of inputs belonging to classes using the Logistic function. 2.2 Calculate correctly the Odds-ratio. 2.3 Calculate correctly relevant classification evaluation metrics for logistic regression model outputs.
3. Be able to create logistic regression models.	3.1 Use Python to build an accurate logistic regression model for datasets.

	3.2 Use Python to evaluate the accuracy of the model built in 3.1. and analyse the results.

Indicative Content

- Binary logistic regression
- Logistic function
- Odds ratio
- Logit Function
- Python
- Sklearn linear_model.
- LogisticRegression
- Seaborn

Recommended Text

Giuseppe Bonaccorso, "*Machine Learning Algorithms: Popular algorithms for data science and machine learning, 2nd Edition*", Packt, 2018

Delivery Guidance

The Level 3 Diploma in Data Science can be delivered:

i) Via distance learning, with all learning materials and assessments available to learners on-line. Support to students is provided via email, or via tools such as Zoom for one-to-one feedback and support.

ii) Via a classroom-based environment, typically taught as 6 hours per week over three terms of 10-week semesters, by data science tutors and supported by data science teaching assistants.

In both cases, learners are provided with detailed core learning materials for each of the fourteen units, and supplementary materials as appropriate, including PDF of lecture slides, Question & Answer bank booklets, and sample code.

Assessment Guidance

To demonstrate all learning outcomes and assessment criteria, each unit will be assessed by a single summative assessment (i.e., an assessment taken after the learner has completed the learning and study for the unit) designed to assess learner's technical knowledge and understanding of the unit's learning outcomes.

Each summative assessment will contain a question for each of the given unit's learning outcomes. The assignment tasks will address the LO (learning outcome) and AC (assessment criteria) requirements. Within assessments there will always be requirements for learners to engage with important and relevant theory that underpins the subject area. Learners will also be given data

A sample assessment with model solutions should be made available to learners.

Suggested Resources

1. Jiawei Han, Micheline Kamber and Jian Pei, "*Data Mining: Concepts and Techniques*", Third Edition, 2012

2. John D. Kelleher, Brian Mac Namee and Aoife D'Arcy, "*Fundamentals of Machine Learning for Predictive Data Analytics*", Second Edition, 2020

3. Sebastian Raschka, "*Python Machine Learning*", PACKT, 2015

Unit 3DS11: Decision Trees in Data Science

Unit code: J/650/4961

RQF Level: 3

Unit Aims

This unit introduces the basic theory and application of decision trees. The unit explains how basic classification trees using the standard ID3 decision-tree construction algorithm are built and how nodes are split based on information theory concepts such has Entropy and Information Gain. The learner will also build and evaluate decision tree models in Python.

Learning Outcomes and Assessment Criteria

Learning Outcomes: **To achieve this unit, the learner must be able to:**	**Assessment Criteria:** **Assessment of these learning outcomes will require a learner to demonstrate that they can:**
1. Understand what a decision tree is in data science.	1.1 Explain what a decision tree is, defining the terms "root node", "parent nodes", "child nodes", "edges" and "leaf nodes". 1.2 Explain the use-cases of decision trees models. 1.3 Analyse the advantages and disadvantages of decision trees.
2. Understand how to construct a decision tree in data science.	2.1 Explain what is meant by splitting and pruning a decision tree. 2.2 Define: - Entropy - Information Gain. 2.3 Explain the key steps in the ID3 (Iterative Dichotomiser) algorithm. 2.4 Analyse improvements and extensions to the ID3 algorithm.
3. Be able to perform calculations using decision tree metrics in data science.	3.1 Calculate correctly Entropy values for a dataset. 3.2 Calculate correctly Information Gain values for a dataset. 3.3 Create accurate visualisations of the Entropy function.
4. Be able to build a decision tree model in data science.	4.1 Use Python to build a decision tree model that is appropriate for a given dataset. 4.2 Use Python to create visualisations that are appropriate for a decision tree.

Indicative Content

- Decision trees
- Root node
- Parent node
- Child node
- Edges
- ID3
- Entropy
- Information Gain
- ID3 algorithm
- Python
- Sklearn tree.DecisionTreeClaissifier
- Tree depth
- Splitting
- Tree pruning

Recommended Text

Giuseppe Bonaccorso, "*Machine Learning Algorithms: Popular algorithms for data science and machine learning, 2nd Edition*", Packt, 2018

Delivery Guidance

The Level 3 Diploma in Data Science can be delivered:

i) Via distance learning, with all learning materials and assessments available to learners on-line. Support to students is provided via email, or via tools such as Zoom for one-to-one feedback and support.

ii) Via a classroom-based environment, typically taught as 6 hours per week over three terms of 10-week semesters, by data science tutors and supported by data science teaching assistants.

In both cases, learners are provided with detailed core learning materials for each of the fourteen units, and supplementary materials as appropriate, including PDF of lecture slides, Question & Answer bank booklets, and sample code.

Assessment Guidance

To demonstrate all learning outcomes and assessment criteria, each unit will be assessed by a single summative assessment (i.e., an assessment taken after the learner has completed the learning and study for the unit) designed to assess learner's technical knowledge and understanding of the unit's learning outcomes.

Each summative assessment will contain a question for each of the given unit's learning outcomes. The assignment tasks will address the LO (learning outcome) and AC (assessment criteria) requirements. Within assessments there will always be requirements for learners to engage with important and relevant theory that underpins the subject area. Learners will also be given data

A sample assessment with model solutions should be made available to learners.

Suggested Resources

1. Jiawei Han, Micheline Kamber and Jian Pei, "*Data Mining: Concepts and Techniques*", Third Edition, 2012

2. John D. Kelleher, Brian Mac Namee and Aoife D'Arcy, "*Fundamentals of Machine Learning for Predictive Data Analytics*", Second Edition, 2020

3. Sebastian Raschka, "*Python Machine Learning*", PACKT, 2015

Unit 3DS12: k-means Clustering in Data Science

Unit code: K/650/4962

RQF Level: 3

Unit Aims

This unit introduces an unsupervised machine learning algorithm: *k*-means clustering. The unit aims to provide learners with the intuition behind *k*-means clustering algorithm and how to find the optimal number of clusters. Finally, the learner will also build and evaluate k-means methods in Python and will learn how visualise the clusters.

Learning Outcomes and Assessment Criteria

Learning Outcomes: **To achieve this unit, the learner must be able to:**	**Assessment Criteria:** **Assessment of these learning outcomes will require a learner to demonstrate that they can:**
1. Understand the theory of *k*-means clustering.	1.1 Explain what is meant by k-means clustering, explaining the terms “cluster” and “centroid”. 1.2 Analyse the steps in the k-means clustering algorithm. 1.3 Explain how to determine the optimal number of clusters “k” by using the elbow method. 1.4 Explain how to interpret: - Sum of Squared Error (SSE) - The Within-Cluster-Sum of Squared Errors (WSS) 1.5 Analyse the limitations of the elbow method. 1.6 Explain the types of use-cases k-means clustering can be applied to. 1.7 Analyse the benefits and limitations of k-means clustering.
2. Understand how to evaluate *k*-means clusters	2.1 Define: - Inertia - Silhouette Score 2.2 Explain how to interpret Inertia and Silhouette score
3. Be able to create and evaluate a *k*-means model.	3.1 Use Python to build an accurate *k*-means model. 3.2 Use Python to create accurate visualisations of the clusters generated by the *k*-means clustering algorithm. 3.3 Use Python to evaluate the accuracy of a *k*-means model.

Indicative Content

- K-means clustering
- Hierarchical clustering
- Density-based clustering
- Clusters and centroids
- The elbow method
- Sum of Squared Error
- Within-Cluster Sum of Squared Errors
- Inertia
- Silhouette Score

Recommended Text

Giuseppe Bonaccorso, "*Machine Learning Algorithms: Popular algorithms for data science and machine learning, 2nd Edition*", Packt, 2018

Delivery Guidance

The Level 3 Diploma in Data Science can be delivered:

i) Via distance learning, with all learning materials and assessments available to learners on-line. Support to students is provided via email, or via tools such as Zoom for one-to-one feedback and support.

ii) Via a classroom-based environment, typically taught as 6 hours per week over three terms of 10-week semesters, by data science tutors and supported by data science teaching assistants.

In both cases, learners are provided with detailed core learning materials for each of the fourteen units, and supplementary materials as appropriate, including PDF of lecture slides, Question & Answer bank booklets, and sample code.

Assessment Guidance

To demonstrate all learning outcomes and assessment criteria, each unit will be assessed by a single summative assessment (i.e., an assessment taken after the learner has completed the learning and study for the unit) designed to assess learner's technical knowledge and understanding of the unit's learning outcomes.

Each summative assessment will contain a question for each of the given unit's learning outcomes. The assignment tasks will address the LO (learning outcome) and AC (assessment criteria) requirements. Within assessments there will always be requirements for learners to engage with important and relevant theory that underpins the subject area. Learners will also be given data

A sample assessment with model solutions should be made available to learners.

Suggested Resources

1. Jiawei Han, Micheline Kamber and Jian Pei, "*Data Mining: Concepts and Techniques*", Third Edition, 2012

2. Sebastian Raschka, "*Python Machine Learning*", PACKT, 2015

Unit 3DS13: Synthetic Data for Privacy and Security in Data Science

Unit code: L/650/4963

RQF Level: 3

Unit Aims

This unit aims to provide learners with an introduction into an emerging area of data science – synthetic data and its application to data privacy and security.

Data collected by companies (such as Google, Facebook, Twitter) as well as governments, are a key resource in today's information age. However, the leaking and inadvertent disclosure of data poses a serious threat to individual privacy.

The unit introduces data privacy, the need for privacy and the legislative landscape. The unit explores traditional means of providing data privacy from anonymisation and encryption, before introducing the learner to the concept of differential privacy and the fundamental challenges of balancing data privacy with data utility.

Learning Outcomes and Assessment Criteria

Learning Outcomes: **To achieve this unit, the learner must be able to:**	**Assessment Criteria:** **Assessment of these learning outcomes will require a learner to demonstrate that they can:**
1. Understand the core issues of data privacy and security.	1.1 Summarise the issues affecting data privacy, data security and data science. 1.2 Analyse standard anonymisation methods for data: shuffling, substitution, masking, binning, deletion. 1.3 Analyse the shortcomings of anonymisation methods. 1.4 Analyse types of attacks on privacy and anonymised data: Linkage attacks, Differencing attacks, Reconstruction attack. 1.5 Analyse high-profile data attacks and breaches and how they occurred.
2. Understand the basics of differential privacy.	2.1 Explain the concept of differential privacy. 2.2 Explain how to interpret the epsilon parameter in differential privacy. 2.3 Analyse the trade-off between data privacy and data utility. 2.4 Analyse the challenges and limitations of

	differential privacy.
3. Understand the core issues of synthetic data.	3.1 Explain the concept of synthetic data 3.2 Analyse techniques to create synthetic data. 3.3 Analyse the features, uses, benefits and drawbacks of synthetic data to anonymisation methods. 3.4 Explain the difference between fake and synthetic data. 3.5 Analyse the benefits of synthetic data over real data. 3.6 Analyse use-cases for synthetic data.
4. Understand the synthetic data ecosystem.	4.1 Analyse the features, uses, benefits and drawbacks of the Python libraries for creating fake data and differential privacy. 4.2 Analyse the features, uses, benefits and drawbacks of other tools for creating fake and synthetic data.
5. Be able to create anonymised or fake data.	5.1 Use Python to create accurate anonymised data from a real dataset. 5.2 Use Python to create fake data with particular attributes in accordance with the specification.

Indicative Content

- Data privacy
- Encryption
- Pseudonoymisation
- Synthetic data
- Shuffling
- Substitution
- Masking
- Binning
- Deletion
- Differential privacy
- Data privacy vs data utility
- Fake data

Recommended Text

Ninghui Li, Min Lyu, Dong Su, and Weining Yang, "*Differential Privacy: From Theory to Practice (Synthesis Lectures on Information Security, Privacy, and Trust)*", Springer, 2016

Delivery Guidance

The Level 3 Diploma in Data Science can be delivered:

i) Via distance learning, with all learning materials and assessments available to learners on-line. Support to students is provided via email, or via tools such as Zoom for one-to-one feedback and support.

ii) Via a classroom-based environment, typically taught as 6 hours per week over three terms of 10-week semesters, by data science tutors and supported by data science teaching assistants.

In both cases, learners are provided with detailed core learning materials for each of the fourteen units, and supplementary materials as appropriate, including PDF of lecture slides, Question & Answer bank booklets, and sample code.

Assessment Guidance

To demonstrate all learning outcomes and assessment criteria, each unit will be assessed by a single summative assessment (i.e., an assessment taken after the learner has completed the learning and study for the unit) designed to assess learner's technical knowledge and understanding of the unit's learning outcomes.

Each summative assessment will contain a question for each of the given unit's learning outcomes. The assignment tasks will address the LO (learning outcome) and AC (assessment criteria) requirements. Within assessments there will always be requirements for learners to engage with important and relevant theory that underpins the subject area. Learners will also be given data

A sample assessment with model solutions should be made available to learners.

Suggested Resources

Khaled El Emam, Lucy Mosquera and Richard Hoptroff, "*Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data*"

Unit 3DS14: Graphs and Graph Data Science

Unit code: M/650/4964

RQF Level: 3

Unit Aims

This unit aims to provide learners with an introduction into another emerging area of data science – graphs and graph data science.

This unit provides a gentle introduction to the field of graph theory which underpins all modern graph databases and graph analytics.

The unit also covers the graph ecosystem, introducing Knowledge Graphs, Labelled Property Graphs and RDF graphs for data storages and processing. The unit introduces graph algorithms which are used to model, store, retrieve and analyse graph-structured data.

Learning Outcomes and Assessment Criteria

Learning Outcomes: **To achieve this unit, the learner must be able to:**	**Assessment Criteria:** **Assessment of these learning outcomes will require a learner to demonstrate that they can:**
1. Understand different types of graphs and their properties.	1.1 Explain what is meant by a “graph”, a “vertex”, a “node” and an “edge”. 1.2 Explain the Bridges of Konigsberg problem and its solution. 1.3 Define the following types of graphs providing examples: - Connected and unconnected graphs - Weighted and unweighted graphs - Directed and undirected graphs - Acyclic and cyclic graphs - Monopartite, Bipartite, and k-partite graphs - Directed Acyclic Graph (DAG)
2. Understand the core types of graph data models.	2.1 Explain what is meant by a “Knowledge graph”. 2.2 Explain what is meant by a “Labelled Property Graph” (LPG). 2.3 Explain what is meant by a “Resource Description Framework” (RDF) graph.
3. Understand the graph ecosystem.	3.1 Outline the graph ecosystem from graph databases, graph languages to graph visualisation tools. 3.2 Analyse the features, uses, benefits and drawbacks of LPG databases, RDF databases and relational databases.

	3.3 Analyse the use-cases and applications for LPG and RDF graph databases. 3.4 Analyse Python graph libraries and their features.
4. Understand the types of graph data science and graph algorithms.	4.1 Explain what is meant by "graph data science". 4.2 Analyse the types of "graph algorithms": search and pathfinding, centrality, and community detection. 4.3 Analyse the types of problems and use-cases that can be tackled by graph data science.

Indicative Content

- Graphs
- Bridges of Konigsberg problem
- Connected and unconnected graphs
- Weighted and unweighted graphs
- Directed and undirected graphs
- Acyclic and cyclic graphs
- Knowledge graphs
- Ontologies
- Labelled property graphs
- Resource Description Framework
- Graph data science
- Pathfinding algorithms
- Centrality algorithm
- Community detection algorithm

Recommended Text

Dr. Alicia Frame and Zach Blumenfeld, "*Graph Data Science for dummies, Second Edition*", Wiley, 2022

Delivery Guidance

The Level 3 Diploma in Data Science can be delivered:

i) Via distance learning, with all learning materials and assessments available to learners on-line. Support to students is provided via email, or via tools such as Zoom for one-to-one feedback and support.

ii) Via a classroom-based environment, typically taught as 6 hours per week over three terms of 10-week semesters, by data science tutors and supported by data science teaching assistants.

In both cases, learners are provided with detailed core learning materials for each of the fourteen units, and supplementary materials as appropriate, including PDF of lecture slides, Question & Answer bank booklets, and sample code.

Assessment Guidance

To demonstrate all learning outcomes and assessment criteria, each unit will be assessed by a single summative assessment (i.e., an assessment taken after the learner has completed the learning and study for the unit) designed to assess learner's technical knowledge and understanding of the unit's learning outcomes.

Each summative assessment will contain a question for each of the given unit's learning outcomes. The assignment tasks will address the LO (learning outcome) and AC (assessment criteria) requirements. Within assessments there will always be requirements for learners to engage with important and relevant theory that underpins the subject area. Learners will also be given data

A sample assessment with model solutions should be made available to learners.

Suggested Resources

1. https://www.tigergraph.com/graph-data-science-library/

2. Victor Lee, Phuc Kien Nguyen, and Xinyu Change, "*Graph-Powered Analytics and Machine Learning with TigerGraph*", O'Reilly, 2022

3. Dr. Jim Webber and Rik Van Bruggen, "*Graph Databases for dummies*", Wiley, 2020

4. Ian Robinson, Jim Webber, and Emil Eifrem, "*Graph databases*", O'Reilly, 2022

Contact Details

Customer service number: +44 (0) 1158882323

Email: support@QUALIFI-international.com

Website: www.QUALIFI.net www.QUALIFI-international.com