National Unit Specification

General information

Unit title: Machine Learning (SCQF level 5)

Unit code: J2G6 45

Superclass: CB

Publication date: July 2019

Source: Scottish Qualifications Authority

Version: 02, August 2019

Unit purpose

The purpose of this unit is to provide a straightforward introduction to machine learning and its applications.

This is a **non-specialist** unit, intended for learners with a vocational interest in STEM subjects and the use of data in a modern context. It is particularly suitable for learners who have experience of other data science topics and wish to broaden their understanding of the discipline. However, it will be of interest to all learners who wish to understand the basic principles of machine learning.

No previous knowledge or experience of machine learning is required. However, it is recommended that learners possess computational and numerical skills before undertaking this unit.

This introductory unit covers the a wide range of knowledge and skills including: the broad purpose of machine learning and its applications in business, health and science; the machine learning workflow; supervised and unsupervised learning; the role of algorithms; training and test datasets; fitting a classifier model and interpreting the results; under-fitting and over-fitting; and the ethical implications of machine learning.

On completion of this unit, learners will know the purpose of machine learning and the role of algorithms in learning from data. They will know the importance of good quality data and the role of feature selection in the learning process. They will be able to distinguish between supervised and unsupervised learning and use a classifier algorithm to produce a model for a given dataset. They will be able to judge the performance of a model, and identify the role of training and test datasets in this process.

Learners may progress to further study of Machine Learning by taking a unit such as J2G7 46 *Data Science: Machine Learning* at SCQF level 6.

National Unit Specification: General information

Unit title: Machine Learning (SCQF level 5)

Outcomes

On successful completion of the unit the learner will be able to:

1 Describe the purpose, applications and key features of machine learning.
2 Explain the importance of good quality data and feature selection in effective machine learning.
3 Apply standard machine learning techniques to create and evaluate a classification model.

Credit points and level

1 National Unit credit at SCQF level 5: (6 SCQF credit points at SCQF level 5)

Recommended entry to the unit

While entry is at the discretion of the centre, it would be beneficial if learners possessed.an understanding of the basics of computing and mathematics. This may be evidenced by possession of a mathematics-related qualification at SCQF level 4 or above, and a computing-related qualification at SCQF level 4 or above. Some previous knowledge of statistical methods (perhaps as part of a mathematics unit) is also desirable but not essential.

Core Skills

Achievement of this Unit gives automatic certification of the following Core Skill component:

Core Skill component Critical Thinking at SCQF level 4

There are also opportunities to develop aspects of Core Skills which are highlighted in the Support Notes of this Unit specification.

Context for delivery

If this unit is delivered as part of a group award, it is recommended that it should be taught and assessed within the subject area of the group award to which it contributes. For example, if this unit is delivered as part of the National Progression Award in Data Science at SCQF level 5 there is overlap with other units within this award (particularly J2G3 45 *Data Science*) and there will be opportunities to contextualise and integrate teaching, learning and assessment across component units.

Equality and inclusion

This unit specification has been designed to ensure that there are no unnecessary barriers to learning or assessment. The individual needs of learners should be taken into account when planning learning experiences, selecting assessment methods or considering alternative evidence.

Further advice can be found on our website **www.sqa.org.uk/assessmentarrangements**.

National Unit Specification: Statement of standards

Unit title: Machine Learning (SCQF level 5)

Acceptable performance in this unit will be the satisfactory achievement of the standards set out in this part of the unit specification. All sections of the statement of standards are mandatory and cannot be altered without reference to SQA.

Where evidence for outcomes is assessed on a sample basis, the whole of the content listed in the knowledge and/or skills section must be taught and available for assessment. Learners should not know in advance the items on which they will be assessed and different items should be sampled on each assessment occasion.

Outcome 1

Describe the purpose, applications and key features of machine learning.

Performance criteria

(a) Describe the advances in technology that have given rise to the growth in machine learning as a facet of artificial intelligence
(b) Describe real-world uses of machine learning, identifying any ethical issues with its use
(c) Explain the role that algorithms play in machine learning
(d) Describe the difference between supervised and unsupervised machine learning problems
(e) Describe the key elements required for supervised learning in machine learning

Outcome 2

Explain the importance of good quality data and feature selection in effective machine learning.

Performance criteria

(a) Describe the characteristics of good quality data and approaches to data collection that help secure it
(b) Explain techniques for identifying and reducing bias in datasets
(c) Explain the role of training and test datasets in developing and evaluating a machine learning model
(d) State the role of feature identification in model development and describe methods in common use
(e) Manipulate datasets with software tools to generate descriptive analytics with a view to feature selection

Outcome 3

Apply standard machine learning techniques to create and evaluate a classification model.

Performance criteria

(a) Identify the main types of algorithm used for classification
(b) Explain how a decision tree works to develop a classifier model
(c) Produce a classifier model by submitting pre-prepared data to a machine learning classifier algorithm
(d) Evaluate the measures of model fit for this classifier
(e) Describe the characteristics of under-fitting and over-fitting in a classifier model
(f) Describe approaches to model improvement in a classifier problem

National Unit Specification: Statement of standards (cont)

Unit title: Machine Learning (SCQF level 5)

Evidence requirements for this unit

Learners will need to provide evidence to demonstrate the performance criteria across all outcomes. The evidence requirements for this unit will take two forms.

1 Knowledge evidence
2 Product evidence

The knowledge evidence will relate to Outcome 1, Outcome 2 and Outcome 3. The knowledge evidence may be written or oral or a combination of these. Evidence may be captured, stored and presented in a range of media (including audio and video) and formats (analogue and digital). The amount of evidence may be the minimum required to infer competence across all outcomes. At least five real-world examples of machine learning and their ethical implications must be described. At least three characteristics of good quality data must be identified and at least one approach to data collection that would secure it must be described. At least two types of algorithm that can be used for classification must be identified. At least three measures of model fit in a classifier problem must be defined. At least two approaches to model improvement must be described.

The knowledge evidence may be sampled when testing is used. Testing must be carried out under supervised conditions and it must be controlled in terms of location and time. Access to reference material is not permitted. The sampling frame, on all occasions, must include Outcome 1, Outcome 2 and Outcome 3 (but not every performance criterion within each outcome).

The **product evidence** will relate to performance criteria in Outcome 2 and Outcome 3. It will demonstrate that the learner can select and use appropriate software tools in regular use for machine learning. In particular, the evidence must demonstrate that the learner can:

- Correctly select and use two statistical tools to generate descriptive statistics from two different datasets for the purpose of feature selection
- Correctly select a classifier algorithm, load a dataset and produce a classifier model
- Correctly identify three features of the classifier model and write a brief report that interprets them in the context of the given dataset

This evidence may be produced over the life of the unit, under loosely controlled conditions (including access to reference materials). Authentication will be necessary (see below). The datasets will be provided by the centre, chosen to match the performance criteria being evidenced.

The SCQF level of this unit (level 5) provides additional context on the nature of the required evidence and the associated standards. Appropriate level descriptors should be used when making judgements about the evidence.

When evidence is produced in loosely controlled conditions it must be authenticated. The guide to assessment provides further advice on methods of authentication.

The support notes section of this specification provides specific examples of instruments of assessment that will generate the required evidence.

National Unit Support Notes

Unit title: Machine Learning (SCQF level 5)

Unit support notes are offered as guidance and are not mandatory.

While the exact time allocated to this unit is at the discretion of the centre, the notional design length is 40 hours.

Guidance on the content and context for this unit

This unit may be delivered as a stand-alone unit or in combination with other units as part of a group of units making up an award, such as NPA in Data Science at SCQF level 5. This unit is suitable for learners who have no previous experience of machine learning or who wish to enhance their knowledge and further develop their skills in the area of data science.

This unit provides an introduction to the basic principles that underpin machine learning, along with an appreciation of its power in supporting decision making in areas such as business, health and science. Learners will also be made aware of the impact on the individual of decisions made on the basis of big data and algorithmic learning.

The overall aim of the unit is to equip learners with the knowledge and understanding to appreciate the opportunities and challenges of machine learning and the skill to apply a machine learning algorithm to a problem dataset. The unit focuses on supervised learning in the context of classifier models. It requires access to cloud services that provide machine learning algorithms, and also to statistical software for data analytics. The choice of exemplar datasets to provide a context for the unit should take into account the interests of the learner cohort.

The following guidance, relating to specific outcomes, does not seek to explain each performance criterion, but to clarify the statement of standards where it is potentially ambiguous. It also focuses on non-apparent teaching and learning issues that may be over-looked, or not emphasised, during unit delivery. As such, it is not representative of the relative importance of each outcome or performance criterion.

Outcome 1: In this outcome, learners will learn how machine learning relates to other artificial intelligence methods and understand the basic principles of how a machine learns though an algorithmic process. The key enablers of machine learning should be covered: big data, data storage, data unification and manipulation, processing power and algorithm development. Real-world applications in fields such as health, science and business should be used to illustrate the benefits and risks of using models derived from machine learning, with particular reference to ethical considerations. Learners should be made familiar with the workflow that enables effective machine learning to take place and understand the importance of good quality data. The concept of supervised and unsupervised learning should be introduced, along with the different purposes of model development in each case. Particular focus should be given to supervised learning and its associated algorithms.

National Unit Support Notes (cont)

Unit title: Machine Learning (SCQF level 5)

Outcome 2: This outcome provides opportunity to expand on the concept of data quality and introduce approaches that might help to achieve good quality data at the data collection stage (completeness, consistency, accuracy, validity, and timeliness). An introduction to various ways of describing data using analytical tools will underpin the concepts of data quality and help understand the importance of feature selection in model development. The concept of training and test datasets and their use in machine learning for model development and evaluation will be explained. Learners will use appropriate software to examine datasets and identify noteworthy features relevant to model development through machine learning.

Outcome 3: This outcome focuses on supervised learning for the purpose of classification (binary only). The following algorithms should be introduced and briefly described: Decision trees, nearest neighbour, linear classifier (Naïve Bayes). The operation of a decision tree model should be covered in more depth. The measures of model fit derived from the classifier confusion matrix should be introduced and explained. This will lead into considerations of over-fitting and under-fitting, particularly in the case of the decision tree model. The strategies for model improvement should cover leaf pruning, minimum leaf size as well as improvements to data and re-sampling. Learners should use software to submit a given dataset to a classifier algorithm and interpret the model results.

Guidance on approaches to delivery of this unit

This unit is an optional unit for the National Progression Award in Data Science at SCQF level 5. As such, it may be delivered alongside the mandatory units: *Data Citizenship* (SCQF level 5) and *Data Science* (SCQF level 5). In this circumstance, teaching, learning and assessment may be integrated across the units.

In Outcome 1, machine learning should be set in the context of other artificial intelligence methods, and its relationship to these methods explained. This will provide opportunity to explain the particular characteristics of machine learning and its relationship to big data, and give exposure to some readily available practical examples of its use. Initiatives such as CS4Fun, IBM Watson and Apps for Good have developed exemplars that are easy to access and implement.

Outcome 2 should provide opportunity for the exploration of sample datasets that illustrate typical problems with incomplete or inaccurate data, and a discussion of typical sources of data in the real world — such as governments and their agencies. Opportunity should be taken to provide multiple illustrations of feature selection through data analytics, involving learners in practical use of software tools for this purpose. The problem of imbalance in classification models should be highlighted, along with possible solutions, which links to model improvement in Outcome 3.

National Unit Support Notes (cont)

Unit title: Machine Learning (SCQF level 5)

In delivering Outcome 3, the main types of classification algorithm can be covered briefly, but decision trees should be explained in some depth, along with some worked-through practical illustrations. The confusion matrix should be introduced along with the various measures of performance that it gives rise to. Reference should be made to the concept of the zero algorithm as a benchmark. The features of over-fitting and under-fitting should be covered in the context of classification, and strategies such as leaf pruning and minimal leaf size explained. In this outcome learners are given practical hands-on experience of submitting a dataset to a classifier algorithm and interpreting the output. This can take the form of using a cloud service with pre-prepared algorithms (eg, Amazon, IBM Watson, Weka) or a programme provided by the centre (eg, R, Python).

A suggested distribution of time, across the outcomes, is:

- Outcome 1: 10 hours
- Outcome 2: 12 hours
- Outcome 3: 18 hours

Summative assessment may be carried out at any time. However, when testing is used (see evidence requirements) it is recommended that this is carried out towards the end of the unit (but with sufficient time for remediation and re-assessment). When continuous assessment is used (such as the use of a web log), this could commence early in the life of the unit and be carried out throughout the life of the unit.

There are opportunities to carry out formative assessment at various stages in the unit. For example, formative assessment could be carried out on the completion of each outcome to ensure that learners have grasped the knowledge contained within it. This would provide assessors with an opportunity to diagnose misconceptions, and intervene to remedy them, before progressing to the next outcome.

Guidance on approaches to assessment of this unit

Evidence can be generated using different types of assessment. The following are suggestions only. There may be other methods that would be more suitable to learners.

Centres are reminded that prior verification of centre-devised assessments would help to ensure that the national standard is being met. Where learners experience a range of assessment methods, this helps them to develop different skills that should be transferable to work or further and higher education.

Most of the assessment evidence in this unit relates to knowledge and understanding. The tradition approach to assessment would comprise a test, taken towards the end of the unit. The test could consist of a number of selected response questions, chosen from all of the outcomes and performance criteria. Not every performance criterion should be tested but every outcome could be tested by a number of questions.

It is suggested that some elements of knowledge lend themselves to a form of selected response (multiple choice) questions. Where explanations are required, the form of the question could be constructed response, scenario-based where appropriate. The whole test would be timed and carried out under controlled conditions, without access to reference material. Where re-assessment is required it should contain a significantly different sample selected from the range of mandatory content.

National Unit Support Notes (cont)

Unit title: Machine Learning (SCQF level 5)

In the case of the product evidence, learners should be supplied with datasets that have been pre-prepared by the centre to allow the learner to demonstrate that they can select and use appropriate software tools to achieve the set tasks and report on the outcomes. An assessor observation checklist could be used to record that the assessment tasks have been undertaken successfully by the learner.

A more contemporary approach to assessment would involve the use of a web log (blog) to record learning (and the associated activities) throughout the life of the unit. The blog would provide knowledge evidence (in the form of descriptions and explanations) and product evidence (in the form of screenshots and output reports). The blog should be assessed using defined criteria to permit a correct judgement about the quality of the evidence. In this scenario, every performance criteria must be evidenced; sampling would not be appropriate.

Formative assessment could be used to assess learners' knowledge at various stages throughout the life of the unit. An ideal time to gauge their knowledge would be at the end of each outcome. This assessment could be delivered through an item bank of selected response questions, providing diagnostic feedback to learners.

If a blog is used for summative assessment, it would also facilitate formative assessment since learning (including misconceptions) would be apparent from the blog, and intervention could take place to correct misunderstandings on an on-going basis.

Opportunities for e-assessment

E-assessment may be appropriate for some assessments in this unit. By e-assessment we mean assessment which is supported by Information and Communication Technology (ICT), such as e-testing or the use of e-portfolios or social software.

Centres which wish to use e-assessment must ensure that the national standard is applied to all learner evidence and that conditions of assessment as specified in the evidence requirements are met, regardless of the mode of gathering evidence. The most up-to-date guidance on the use of e-assessment to support SQA's qualifications is available at **www.sqa.org.uk/e-assessment**.

Opportunities for developing Core and other essential skills

The Core Skills that may be developed include: *Numeracy* (SCQF level 5); *Information and Communication Technology* (SCQF level 5). General skills that may be developed include data literacy and data citizenship.

The Critical Thinking component of Problem Solving at SCQF level 4 is embedded in this unit. When a learner achieves this unit, their Core Skills profile will also be updated to include this component.

History of changes to unit

Version	Description of change	Date
02	Core Skills Component Critical Thinking at SCQF level 4 embedded.	16/08/19

General information for learners

Unit title: Machine Learning (SCQF level 5)

This section will help you decide whether this is the unit for you by explaining what the unit is about, what you should know or be able to do before you start, what you will need to do during the unit and opportunities for further learning and employment.

This unit provides a straightforward introduction to machine learning and its applications. It assumes no prior knowledge of machine learning, but requires that you have a good foundational knowledge in mathematical topics, including some statistics, and can understand how computers process information.

The unit will explain the broad purpose of machine learning and its relationship to other artificial intelligence methods, and introduce you to typical applications in business, science and health. You will appreciate that machine learning offers opportunities but also risks to the individual. The unit will cover the role of good quality data in the machine learning workflow; how algorithms function to learn a model; and the uses to which those models can be put.

The fundamentals of machine learning include: supervised and unsupervised learning; training and test datasets; measures of model fit; under-fitting/over-fitting; and feature identification. You will learn how to develop and evaluate a classification model using a given dataset and the possible strategies to improve its performance. This practical work will use pre-prepared programs or software packages.

The knowledge and understanding components of the unit may be assessed through a knowledge test as determined by your centre, in controlled conditions. Typically, this will involve a blend of multiple-choice questions, short response questions and scenario-based questions. The practical elements of the unit will be assessed by a performance checklist completed by the assessor.

The unit will provide an opportunity for you to progress in the Core Skills of *Numeracy*, *Problem Solving* and *Information and Communication Technology (ICT)*. You will also progress in computational thinking through exposure to algorithms and their operation. Your understanding of the implications of machine learning for society and the individual will contribute to your data citizenship competences.

The Critical Thinking component of Problem Solving at SCQF level 4 is embedded in this unit. When a learner achieves this unit, their Core Skills profile will also be updated to include this component.

On completion of this unit you may progress to J2G7 46 *Data Science: Machine Learning* SCQF level 6. You will also have the knowledge and skills to undertake a project in Machine Learning with the unit J2GT 45 *Data Science Project* at SCQF level 5.