National Unit Specification

General information

Unit title: Machine Learning (SCQF level 6)

Unit code: J2G6 46

Superclass: CB

Publication date: July 2019

Source: Scottish Qualifications Authority

Version: 03 (August 2019)

Unit purpose

The purpose of this unit is to provide a grounding in some of the computational approaches that pertain to machine learning, along with an appreciation of methods to prepare and select data to facilitate model development and use. It will develop skills in fitting and evaluating a predictive model, and introduce strategies to measure and improve model performance.

This is a **specialist** unit, intended for learners who wish to extend a basic understanding of machine learning to encompass the principles that guide model selection and development, and apply machine learning algorithms for supervised learning in multi-class classification and linear regression. It is suitable for learners who have a sound underpinning of mathematics and computational methods.

The unit covers the following knowledge and skill: data scaling and normalisation; feature engineering; model validation; linear regression algorithms; gradient descent; logistic regression for binary classification; interpretation of algorithm outputs; measurement and improvement of model performance.

On completion of this unit, learners will know the two main categories of supervised learning (classification and regression) and the algorithms used to derive models for prediction. They will understand the importance of well-ordered data and suitable choice of features for model development. They will know methods for measuring model performance and some strategies to improve model performance. They will be able to use descriptive analytics to select appropriate features in a given dataset, submit the dataset for processing by a machine learning algorithm, interpret the output, and evaluate the skill of the derived model.

Learners may progress to further study in Machine Learning at SCQF level 7. They can apply their knowledge and skills to a data science project, such as J2GV 46 *Data Science Project* (SCQF level 6).

National Unit Specification: General information (cont)

Unit title: Machine Learning (SCQF level 6)

Outcomes

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

1 Explain the purpose, applications and key features of data preparation and feature selection for machine learning.
2 Explain how machine learning models can be evaluated, and their predictive performance improved by ensemble methods.
3 Describe the application of regression models to problems of prediction and classification.
4 Derive a prediction model from a given dataset using linear regression.

Credit points and level

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

Recommended entry to the unit

It is recommended that learners possess a basic knowledge and understanding of machine learning, its algorithmic basis and its relationship to artificial intelligence, including its role in modern society. This might be evidenced by possession of the SCQF level 5 unit *Machine Learning* or any equivalent qualification. It is desirable that learners have a sound understanding of mathematical methods and computation. Some previous knowledge of statistical methods would also be desirable but is not essential.

Core Skills

Achievement of this Unit gives automatic certification of the following Core Skills components:

Core Skill components	Providing/Creating Information at SCQF level 5
	Critical Thinking at SCQF level 5

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 6 there is overlap with other units within this award (particularly J2G4 46 *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 6)

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

Explain the purpose, applications and key features of data preparation and feature selection for machine learning.

Performance criteria

(a) Explain how machine learning attempts to approximate an unknown mapping function from inputs to outputs
(b) Describe the purpose and methods of data scaling and normalisation in relation to machine learning
(c) Describe the purpose and methods of feature extraction and selection in model development in machine learning
(d) Explain data bias and describe methods to reduce bias

Outcome 2

Explain how machine learning models can be evaluated, and their predictive performance improved by ensemble methods.

Performance criteria

(a) Explain the purpose and process of model validation
(b) Explain the concepts of over-fitting and under-fitting, and bias and variance
(c) Describe the common measures of model performance used in supervised learning
(d) Describe common strategies to address problems with model performance
(e) Explain the role that ensemble methods play in improving performance

Outcome 3

Describe the application of regression models to problems of prediction and classification.

Performance criteria

(a) Describe the purpose of a linear regression model and associated measures of goodness of fit
(b) Describe the common algorithms for fitting a linear regression model
(c) Describe the purpose of logistic regression as applied to binary classification

National Unit Specification: Statement of standards (cont)

Unit title: Machine Learning (SCQF level 6)

Outcome 4

Derive a prediction model from a given dataset using linear regression.

Performance criteria

(a) Select and use appropriate analytic tools to examine and choose appropriate features in a given dataset with a view to making predictions
(b) Select and use a linear regression algorithm to fit a regression model to a given dataset, and interpret its output in terms of performance
(c) Use an ensemble method to improve the performance of this regression model

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 two common algorithms for fitting a linear regression model must be described. At least two common measures of model performance used in supervised learning must be described. At least two common strategies to address problems with model performance 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 sampling frame must always include Outcome 1, Performance Criterion (d).

The **product evidence** will relate to Outcome 4. It will demonstrate that the learner has the competence and understanding to deal with datasets to prepare them for machine learning, and then select and apply a machine learning algorithm and interpret its results. The product evidence must satisfy the following criteria:

- Select and use appropriate analytic tools to examine and choose appropriate features in a given dataset with a view to making predictions
- Select and use a linear regression algorithm to fit a regression model to a given dataset, and interpret its output in terms of performance
- Use an ensemble method to improve model performance on a 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.

National Unit Specification: Statement of standards (cont)

Unit title: Machine Learning (SCQF level 6)

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 6)

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

The purpose of this unit is to provide a grounding in some of the computational approaches that pertain to machine learning, along with an appreciation of methods to prepare and select data features that facilitate model development and use. It will develop skills in fitting and evaluating a linear predictive model for both regression and classification purposes, and introduce strategies to measure and improve model performance.

There are three outcomes that build knowledge and understanding of the key elements of supervised machine learning, including approaches to optimisation and measuring model performance. The importance of well-behaved data is emphasised, with reference to feature engineering to improve model fitting. The final outcome provides opportunity for the practical application of these concepts to real datasets, so that learners will experience most of the stages in the machine learning workflow.

For those learners undertaking this unit as part of the NPA in Data Science at SCQF level 6, this unit has strong links to the concepts that learners will encounter in *Data Citizenship* (SCQF level 6), and *Data Science* (SCQF level 6).

Learners will recognise the importance of the unit content to modern business, health and science where the use of big data and artificial intelligence methods are revolutionising operations and decision-making. Individuals who successfully complete this unit will have an informed view of any claims made for predictions arising from a machine learning model.

The knowledge and skills developed in this unit all have direct relevance to the practice of machine learning for supervised learning. Their application will be practised through practical problem solving and the use of computational methods.

The emphasis on algorithmic approaches to machine learning, and the use of software tools to address given problems, all contribute to the development of computational thinking. Learners will break down problems into their constituent elements and select and apply computational methods to achieve a required outcome. They will also evaluate the output from computation and measure performance.

Learners may progress to units in Data Science at higher levels, or to HNC in Data Analytics at SCQF level 7. Unit titles include: *Machine Learning* (SCQF level 7); *Artificial Intelligence* (SCQF level 7) and *Big Data* (SCQF level 7).

National Unit Support Notes (cont)

Unit title: Machine Learning (SCQF level 6)

Please note that 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

The machine learning workflow model; define the machine learning task (find a function to map inputs to outputs); understand how the machine 'learns'; supervised (prediction) vs unsupervised (investigation); cost functions and their optimisation.

Model assumptions regarding distribution form of data; normalisation vs data scaling; data requirements for specified algorithms.

Role of domain knowledge and heuristics in feature engineering; representation using one-hot encoding; identifying highly correlated features; examining performance of particular features with a view to their selection (eg, use of feature importance charts).

Outcome 2

Metrics for future predictive accuracy; mitigation of risk in using learned model; 'held-out' data (training/testing datasets). Over-fitting and its causes; under-fitting and its causes; problems arising from over/under-fitting; concept of bias and variance; bias/variance trade-off. Training error; testing error; RSQ (adjusted); resampling; cross-validation (leave one out, k-fold, stratified k-fold); Ensemble methods; concept of random sampling with replacement (bootstrapping); bagging (bootstrap aggregation); role of bagging (the 'average' model).

Outcome 3

Assumptions of a linear model (linear relationship, normal distribution; feature independence, homogeneity of variance); linear model parameters; predictive performance; MSE; RSQ and RSQ (adjusted). Cost functions (L1, L2, log loss); principle of gradient descent; gradient descent applied to linear regression (algorithm).

Fitted models: OLS; learned model: Stochastic Gradient Descent (SGD); regression trees; single variable; multi-variable. Logistic model (log odds); estimation of model parameters; predicting probability of a class label; decision boundary.

Outcome 4

Opportunity to introduce machine learning workbenches, tools for exploratory data analysis and visualisation, program code (Python and R libraries). Interpret the output from a linear regression algorithm. Interpretation of the several statistics reported by an SGD algorithm in terms of model performance. Application of a bagging approach to improve the fitted model.

National Unit Support Notes (cont)

Unit title: Machine Learning (SCQF level 6)

Guidance on approaches to delivery of this unit

Staff delivering this unit should have competence in applying the machine learning workflow to datasets with a view to prediction (linear regression) or classification (logistic regression). They should possess good data skills and have familiarity with the range of software environments and products that support machine learning.

The delivery of this unit will require access to computing power sufficient for the application of machine learning algorithms. A significant proportion of the delivery of this unit should be in an ICT-enabled classroom that supports access to machine learning services such as supplied by Amazon, Microsoft, IBM and others.

It is expected that the approach to delivery will be learner-centred, participative and practical. Concepts should be presented and illustrated through short case studies, using examples with a manageable number of features (explanatory variables). Where possible, learners should be encouraged to practice data manipulation skills as the unit progresses. In this regard, Outcome 4 should be experienced in a holistic way alongside the presentation of concepts and methods in Outcomes 1–3.

The datasets used by the tutor should have well-understood characteristics to aid the learning of the methods and approaches, and correspond well to the concepts being learned. There are many sources for datasets of this kind, and these should be assembled and checked by the tutor prior to the start of the course to ensure that they do not present problems that are beyond the scope of SCQF level 6.

Centres should be able to offer learners free access to online machine learning services (such as Amazon, IBM Watson, Weka) to generate models and measure performance. Alternatively, where the learner cohort permits, use may be made of coding environments such as Python and R Studio, along with appropriate libraries for machine learning and data analysis (eg, scikit-learn; matplotlib; caret; ggplot2). In this case, learners should have sufficient familiarity with coding to implement these libraries.

Centres are also encouraged to give learners the opportunity to hear from industry experts — webinars are useful for making this easy to arrange.

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

- Outcome 1: 6 hours
- Outcome 2: 10 hours
- Outcome 3: 10 hours
- Outcome 4: 14 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.

National Unit Support Notes (cont)

Unit title: Machine Learning (SCQF level 6)

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.

Knowledge evidence for this unit could be assessed by methods such as:

1 A selected response test consisting of four options (one key) with a pass mark of 60%. Given that some performance criteria relate to explanations (rather than descriptions), there may need to be some scenario type questions to assess the learner's competency. The test could consist of a relatively high number of questions (30 or 40 for example), lasting an hour, which would span all of the outcomes and sample all of the knowledge statements (including at least one question for each statement).

Or

2 A constructed response test comprising a number of short answer questions, marked and assessed traditionally. For example, the test may comprise of 10 questions, requiring a response comprising no more than one or two paragraphs, selected across Outcomes 1, 2 and 3, each worth five marks, with the learner responses marked out of 50 and a pass mark of 25. If this approach is taken, it is recommended that some (or all) of the questions combine the knowledge and understanding within and across outcomes. This test would be taken, sight-unseen, in controlled and timed conditions without reference to teaching materials. A suitable duration could be 60 minutes.

Or

3 A report or presentation.

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 descriptions and explanations) and product evidence (using, for example, screenshots from package runs). The blog should be assessed using defined criteria to permit a correct judgement about the quality of the evidence. In this scenario, every performance 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.

National Unit Support Notes (cont)

Unit title: Machine Learning (SCQF level 6)

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

Outcome 4 presents opportunities for learners to demonstrate the Core Skills of: *Numeracy* (handling data); *Communication* (presenting a report); *Information and Communication Technology (ICT)* (use of software to manipulate data and create reports) and *Problem Solving* (making decisions about model form and model adequacy, feature engineering).

This unit also develops computational thinking, such as skills in abstraction (model building), decomposition (machine learning workflow), pattern recognition (feature engineering and model selection), and generalisation (using predictive data). The skills developed are those in demand by employers seeking learners with good data skills. The broad understanding of the applications and limitations of machine learned modelling will serve to enhance citizenships skills (data citizenship).

The Providing/Creating Information component of Information and Communication Technology at SCQF level 5 is embedded in this unit and the Critical Thinking component of Problem Solving at SCQF level 5 is embedded in this unit. When a learner achieves these units, their Core Skills profile will also be updated to include these components.

History of changes to unit

Version	Description of change	Date
03	Coding changed due to hierarchy	23/08/19
02	Core Skills Component Providing/Creating Information at SCQF level 5 embedded. Core Skills Component Critical Thinking at SCQF level 5 embedded.	16/08/19

General information for learners

Unit title: Machine Learning (SCQF level 6)

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 is a **specialist** unit that will extend a basic understanding of machine learning to develop knowledge of the principles that guide model selection and development. You will learn to apply machine learning algorithms for supervised learning using linear models for prediction. To succeed in this unit you should have a sound underpinning of mathematics and computational methods.

You will acquire knowledge and skills in the following aspects of machine learning: data scaling and normalisation; feature engineering; model validation; linear regression algorithms; gradient descent; logistic regression for binary classification; interpretation of algorithm outputs; measurement and improvement of model performance.

You will gain an understanding of the importance of well-ordered data and how to choose features for model development. You will be able to use descriptive analytics to select appropriate features in a given dataset, submit the dataset for processing by a machine learning algorithm, interpret the output, and evaluate the skill of the derived model. Finally, you will learn about an ensemble method (bagging) to improve model performance.

Your learning experience will be a combination of presentations of concepts, along with the examination of case studies and participating in practical workshops. In the practical sessions you will gain skills in using online machine learning services provided by the likes of Microsoft and Amazon, or in some cases, skills in coding solutions using a programming environment.

The knowledge elements of this unit may be assessed by a knowledge test. This may be an online test, or a classroom test. The assessment of your skills in machine learning will comprise the analysis of given datasets. You will demonstrate skills in analysing data, using a linear regression algorithm and reporting on the results from your analysis and model building. Finally, you will perform an ensemble method (bagging) to a given problem.

The learning experiences in this unit will afford opportunity for you to develop the Core Skills of *Numeracy* (handling data); *Communication* (presenting a report); *Information and Communication Technology (ICT)* (use of software to manipulate data and create reports) and *Problem Solving* (making decisions about model form and model adequacy, feature engineering). You will develop computational thinking, such as skills in abstraction (model building), decomposition (machine learning workflow), pattern recognition (feature engineering and model selection), and generalisation (using predictive data).

The Providing/Creating Information component of Information and Communication Technology and the Critical Thinking component of Problem Solving at SCQF level 5 are embedded in this unit.

These are skills developed that are in demand by employers seeking learners with good data skills. The broad understanding of the applications and limitations of machine learned modelling will serve to enhance your citizenships skills (data citizenship).

You may progress to units in Data Science at higher levels, such as: J1CN 47 *Machine Learning* (SCQF level 7); J1CD *47 Artificial Intelligence* (SCQF level 7) and HR9T 47 *Big Data* (SCQF level 7). You may also consider the HH7X 34 HNC in Data Analytics at SCQF level 7.