Module Specification

Applied Statistics

Version: 2026-27, v1.0, 19 Sep 2023

Contents

Part 1: Information

Module title: Applied Statistics

Module code: UFCE5A-30-3

Level: Level 6

For implementation from: 2026-27

UWE credit rating: 30

ECTS credit rating: 15

College: College of Arts, Technology and Environment

School: CATE School of Computing and Creative Technologies

Partner institutions: None

Field:

Module type: Module

Pre-requisites: None

Excluded combinations: None

Co-requisites: None

Continuing professional development: No

Professional, statutory or regulatory body requirements: None

Part 2: Description

Overview: The modern data scientist must be keenly aware of modern statistical methods and their applications within data science. This module aims to assist students in their understanding and application of advanced statistical modelling techniques within real-world scenarios, and through the utilization of real-world data. We aim to cover dimensionality reduction, sophisticated data imputation, supervised and unsupervised learning for the application of business data analytics.

Features: Not applicable

Educational aims: The purpose of this module is to draw upon students with a previous understanding of data analysis and statistics and assist students in developing their understanding of further statistical methods. This module focuses on the application of statistical techniques to real-world datasets, which are often messy and contain dirty data.

On completion of the module, students should be able to apply current statistical methodologies frequently used within data science and apply methodologies such as dimensionality reduction and data imputation to allow students to gain insights and perform statistical inference within complex and poorly cleaned datasets

Outline syllabus: Missing Data (Missing at Random, Missing Not at Random, Missing Completely at Random)

Basic Statistical Imputation (Mean, Mode, Median, Last Observation First, Worst Case Imputation)

Advanced Statistical Imputation (Multiple Imputation techniques such as MVNI and MICE)

Decision Tree Modelling

Random Forest Models

Bagging and Bootstrapping within Random Forest Models

Dimensionality Reduction (Principal Component Analysis, Linear Discriminant Analysis)

Cluster Analysis (K-Means, DBSCAN, Hierarchical Clustering)

Application of Statistical Analysis on Real World Data

Data Cleaning and Preprocessing

Part 3: Teaching and learning methods

Teaching and learning methods: Learning and teaching will be provided to students in two forms: lectures and computer labs.

During lectures, theoretical aspects of the module will be provided to students by the teaching staff. Lectures will be supported by presentation published and available to the students on Teams as well as additional materials (publications, videos, etc.).

Computer labs are devoted to practical data analysis using modern software. Jupyter will be used for illustration of techniques and output analysis; students are allowed to use other software packages (Spyder, VS Code, etc.) by a prior agreement with the module instructor. Computer lab classes are reserved for requirement clarifications, problem discussion, and assessment. Students are expected to carry out the work independently outside the classes.

300 hours study time of which 72 hours will represent scheduled learning. Scheduled learning includes lectures, seminars, tutorials, demonstration, practical classes, and workshops; external visits; supervised time in studio/workshops.

Independent learning includes hours engaged with essential reading, case study preparation, assignment preparation and completion. Apprentice study time will be organised each week with a series of both essential and further readings and preparation for practical workshops. It is suggested that preparation for lectures, practical workshops, session delivery and seminars will take 7 hours per week

Module Learning outcomes: On successful completion of this module students will achieve the following learning outcomes.

MO1 Demonstrate an understanding of statistical imputation, using both basic and advanced imputation techniques, and an understanding of data cleaning and sanitization, and the importance of correcting, repairing, or removing invalid data.

MO2 Demonstrate the use of supervised learning through the usage of decision tree and random forest modelling.

MO3 Demonstrate an understanding of Principal Component Analysis and Linear Discriminant Analysis and appropriately apply these techniques to high dimensional data.

MO4 Critically evaluate the metrics that can be used for measuring the accuracy of imputation predictions.

Hours to be allocated: 300

Contact hours:

Independent study/self-guided study = 228 hours

Face-to-face learning = 72 hours

Total = 300

Reading list: The reading list for this module can be accessed at readinglists.uwe.ac.uk via the following link

Part 4: Assessment

Assessment strategy: This module consists of two assessments, a report and a portfolio.

REPORT
The first assessment (2000-word Imputation Analysis Report) will require students to conduct an analysis on a publicly available dataset.

The report will enable students to explain and apply their knowledge in statistical imputation and provide an analysis of their findings. Students will utilise at least two basic statistical imputation techniques and provide a comparison with at least one advanced imputation technique. They will utilise a real-world dataset and contrast the predicted values with the actual values. The report will focus on the difference between the imputation techniques, the importance of data cleaning and sanitization,

and the importance of the metrics chosen and what their values mean. Students will be asked to draw conclusions on the best imputation method, and why it was the most favourable for this dataset.

Students will also be asked to evaluate the difference between PCA and LDA and discuss how their usage has been improved by the imputation of the missing data.

PORTFOLIO
In the final assessment (Practical Portfolio) students will compare the difference between supervised and unsupervised learning and justify their methodology for usage on two different datasets. Their methodology should utilise Random Forest for Supervised, and Clustering for Unsupervised. They will further need to use data visualization and exploration to ensure they can fully justify their choice, before performing an analysis using the dataset. Their practical submission should include the code used for their data exploration and analysis, and should come with an evaluation, which would briefly describe the methodology chosen and the results of the analysis.

Students will conduct the work individually and independently.

The resit opportunities should follow the same format as the first assessment, however consideration should be given to re-work the portfolio data given the complexity of assessment.

Assessment tasks:

Report (First Sit)
Description: 2000 word Imputation Analysis Report
Weighting: 60 %
Final assessment: No
Group work: No
Learning outcomes tested: MO1, MO4

Portfolio (First Sit)

Description: Students will be given two separate datasets and asked to justify the usage of a Supervised and Unsupervised methodology on each dataset.
Weighting: 40 %
Final assessment: Yes
Group work: No
Learning outcomes tested: MO2, MO3

Report (Resit)
Description: 2000 word Imputation Analysis Report
Weighting: 60 %
Final assessment: No
Group work: No
Learning outcomes tested: MO1, MO4

Portfolio (Resit)
Description: Students will be given two separate datasets and asked to justify the usage of a Supervised and Unsupervised methodology on each dataset.
Weighting: 40 %
Final assessment: Yes
Group work: No
Learning outcomes tested: MO3, MO4

Part 5: Contributes towards

This module contributes towards the following programmes of study:

Digital and Technology Solutions (Data Analyst) {Apprenticeship-UCW} [UCW] BSc (Hons) 2023-24