MODULE SPECIFICATION

This module is a compulsory module of the new MSc Health Data Science. The module specification has been provisionally agreed as part of the validation process for the MSc. Module Organisers are currently developing the details of the teaching to ensure the best possible learning experience and therefore some changes may still be made. We anticipate that the final module specifications will be published by the end of the summer.

1. Overview

Academic Year (student cohort covered by specification)	2020-21			
Module Code	2490			
Module Title	Machine Learning			
Module Organiser(s)	Alex Lewin and Karla Diaz-Ordaz			
Faculty	Epidemiology & Population Health			
FHEQ Level	Level 7			
Credit Value	CATS	15	ECTS	7.5
HECoS Code	101030			
Term of Delivery	Term 2			
Mode of Delivery	Face-to-face			
Mode of Study	Full-time			
Language of Study	English			
Pre-Requisites	Students are required to have completed the module Statistics for Health Data Science, or equivalent. Students must be able to demonstrate basic grounding in probability, hypothesis testing, least squares, maximum likelihood and familiarity with linear regression models and generalised linear models (GLMs). Students must also be able to demonstrate familiarity with R software and be able to perform data manipulation in that software and to run packages.			
Accreditation by Professional	None			

Statutory and Regulatory Body	
Module Cap (Maximum number of students)	In first year of delivery: max 20 students.
Target Audience	This is a compulsory module for the programme MSc Health Data Science. It is intended for students who wish to be able to apply modern machine learning methods in health data research. The focus will be on critical understanding and applied analysis.
Module Description	This module provides an introduction to statistical and machine learning, with application to health data science. The module will cover a range of widely-used machine learning techniques used in health data research. The principles of learning algorithms will be covered, and the data analysis tasks will use existing packages in the freely-available R software.
Duration	5 weeks at 2 days per week
Timetabling slot	Slot C1
Last Revised (e.g. year changes approved)	December 2019

2. Programme(s) that this module is part of

Programme This module is linked to the following programme(s)	**Status**
MSc Health Data Science	Compulsory

3. Module Aim and Intended Learning Outcomes

Overall aim of the module
The overall module aim is to: • introduce the concept of statistical and machine learning, and cover a range of supervised learning methods.

Module Intended Learning Outcomes
Upon successful completion of the module a student will be able to: 1. contrast the principles behind a range of statistical and machine learning methods;   2. examine the application of a range of machine learning techniques to address health data science questions;   3. critically evaluate strengths and limitations of statistical and machine learning methods in health data science projects;   4. critically assess the application of different techniques and choose an appropriate algorithm to answer a specific health data science question.

4. Indicative Syllabus

Session Content
The module is expected to cover the following topics: • Regression with the generalised linear model (GLM)   • Support vector machines   • K-nearest neighbours   • Decision trees and random forests   • Regularised/penalised methods for feature selection (e.g. LASSO)   • Ensemble methods (boosting and superlearner)   • Principles of machine learning algorithms   • Cross-validation and prediction

5. Teaching and Learning

Notional Learning Hours		
Type of Learning Time	**Number of Hours**	**Expressed as Percentage (%)**
Contact time	40	27
Directed self-study	50	33
Self-directed learning	30	20
Assessment, review and revision	30	20
Total	150	100

Teaching and Learning Strategy

Flipped classroom

- Prior to the introduction of each topic, students will be expected to do some background reading first in order to engage with a formative quiz at the beginning of, and group discussion during, each teaching session.

Interactive lectorials

- The teaching sessions will be run as lectorials, an interactive format that alternates between lecture-based delivery of material and hands-on practical work (optionally in small groups)
- Students will be expected to bring laptop computers to class to engage with the material in the appropriate software environment

Adaptive release

- Practical worksheets will only become available to students after completing the quiz based on the background reading to ensure that students are engaging critically with the material.

Some sessions mini-presentations (small groups):
- different data sets for different groups
- mini-presentations, peer assessment/discussion

Indicative Breakdown of Contact Time:

Type of delivery	Total (hours)	
Lecture	0	
Seminar	0	
Tutorial	0	
Computer Practical	0	
Laboratory Practical	0	
Fieldwork	0	
Project Supervision	0	
Lectorial (interactive)	36	
Group mini-projects/presentations	4	
Total	**40**	

6. Assessment

Assessment Strategy
The module will employ a number of different formative and summative assessment strategies. To ensure students critically engage with the pre-reading material, formative quizzes will be used during interactive contact sessions. To allow students to practice relevant skills, the module will include a mini-project, similar to the one used in the summative assessment, part-way through. Students will undertake the mini-project in groups and present their findings to other student groups. Staff and other students will provide feedback on the presentations. The summative assessment will be a data analysis project, assessed by oral presentation in front of 2 lecturers and the submission of a reproducible data analysis report (Rmarkdown or Jupyter notebook) (individual work) on which the presentation is based.

Summative assessment			
Assessment Type *(delete as appropriate)*	**Assessment Length (i.e. Word Count, Length of presentation in minutes)**	**Weighting** (%)	**Intended Module Learning Outcomes Tested**
Individual Presentation	10-minute presentation and 5-minute Q&A session. + Accompanying 6-page data analysis report.	100	1- 4

Resitting assessment
Resits will accord with the LSHTM's [Resits Policy in Chapter 8a, PGT Regulations, of the Academic Manual](#)

For individual students resitting a group assessment there will be an approved alternative assessment as detailed below.

Assessment being replaced	Approved Alternative Assessment Type	Approved Alternative Assessment Length (i.e. Word Count, Length of presentation in minutes)
NA – no assessed group work		

7. Resources

Indicative reading list
James, G., D. Witten, T. Hastie, and R. Tibshirani. An Introduction to Statistical Learning: With Applications in R. Springer Texts in Statistics. Springer New York, 2014. http://faculty.marshall.usc.edu/gareth-james/.

Kuhn, M., and Johnson, K. Applied Predictive Modeling. SpringerLink : Bücher. Springer New York, 2013.

Kenett, R. and Redman, T. The Real Work of Data Science. Wiley, 2019.

Other resources
Module information, including timetables, lecture notes, practical instructions and key literature for each session will be made available via the Virtual Learning Environment (Moodle).

8. Teaching for Disabilities and Learning Differences

- Interactive sessions will be recorded using Panopto in line with the LSHTM's policy on Lecture Recording.
- The module manual will be made available in advance of the start of the module and will be produced in accessible format.
- All material will be made available through Moodle.