COURSE UNIT (MODULE) DESCRIPTION

Course unit (module) title	Code
Data mining	

Lecturer(s)	Department(s) where the course unit (module) is delivered
Coordinator: dr. Jurgita Markevičiūtė **Other(s):**	Department of Statistical Analysis Institute of Applied Mathematics Faculty of Mathematics and Informatics Naugardukas st., 24

Study cycle	Type of the course unit (module)
Second	Compulsory

Mode of delivery	Period when the course unit (module) is delivered	Language(s) of instruction
Face-to- face	First (autumn) semester	English/Lithuanian

Requirements for students	
Prerequisites:	**Additional requirements (if any):** Parametric and nonparametric statistics or econometrics Multivariate statistics

Course (module) volume in credits	Total student's workload	Contact hours	Self-study hours
5	150	42	108

Purpose of the course unit (module): programme competences to be developed

The main goal of the course is a) to learn to find and analyse empirical data from various sources; b) to define the process of developing a model in a way that one can understand c) to create and evaluate appropriate models and use best methods in a situation, evaluate adequacy of the results. Enhance critical and analytic thinking.

Competences

- creatively solve nonstandard theoretical and empirical problems (B1)
- critically analyze and correctly apply the results presented in the scientific literature (B2)
- concisely and clearly present the results of the analysis (B3.4)
- to combine knowledge of statistics, economics, mathematics and other sciences for solving practical problems (B4.2)
- to know the static and dynamic models and methods of analysis: a) in the time and frequency domain; b) in a continuous and a discrete time (D5.1)
- to understand the underlying probabilistic laws and statistical principles used for the stochastic models (D5.2)
- an understanding of empirical adequacy testing principles of economic and statistical models and can apply them in practice (D8.1)
- knows various model adequacy tests for the identification of potential problems (D8.2)
- to create and supervise statistical and / or economic models (D9.1)
- to create and supervise machine learning algorithms (D9.2)
- prepare raw empirical data for the econometric analysis and professionally operate the econometric software (D10)

Learning outcomes of the course unit (module)	Teaching and learning methods	Assessment methods
• To prepare training set data for further analysis.	Lectures, labs, case studies.	Individual project, exam.
• To quantify measure of model prediction performance.	Lectures, labs, case studies.	Individual project, exam.

• To build and analyse linear regression, partial least squares and penalized models. • To build and use nonlinear regression models using neural networks, multivariate adaptive regression spline, and support vector machine and k-nearest neighbours. • To build and understand regression trees and rule-based models.		
• To build and understand linear classification models. • To build and understand nonlinear classification models. • To build and understand classification trees.	Lectures, labs, case studies.	Individual project, exam.
• To make suitable conclusions, interpretation of the model. • To understand and explain the pitfalls related to the modelling techniques and big data.	Lectures, labs, case studies.	Individual project, exam.

Content: breakdown of the topics	**Contact hours**							**Self-study work: time and assignments**	
	Lectures	Tutorials	Seminars	Exercises	Laboratory work	Internship/work placement	**Contact hours**	**Self-study hours**	**Assignments**
1. Introduction. Prediction Versus Interpretation. Key Ingredients of Predictive Models. Terminology	2				4		**6**	**18**	[1] ch. 1.4 analyse the examples
2. General strategies. Data pre-processing. Over fitting and model tuning.	4				8		**12**	**30**	[1] ch. 3 and 4, exercises, p. 58-59 and 89-92
3. Regression models. Linear and nonlinear regression. Regression trees.	4				8		**12**	**30**	[1] ch. 6, 7 and 8, exercises, p. 137-139, 168-171, 218-220.
4. Classification models. Linear and nonlinear classification. Classification trees.	4				8		**12**	**30**	[1] ch. 12, 13 and 14 exercises, p. 326-328, 366-367, 411-413.
Total	**14**				**28**		**42**	**108**	

Assessment strategy	**Weight,%**	**Deadline**	**Assessment criteria**
Individual project	60%	During semester	Students are evaluated according projects originality, compliance to subject methods and goal, project quality and presentation of the project and/or its parts on time.
Exam	40%	June	A test 30 questions. Each question is worth of one point.

Author	**Year of publication**	**Title**	**Issue of a periodical or volume of a publication**	**Publishing place and house or web link**
Compulsary reading				
[1] Max Kuhn and Kjell Johnson	2013	Applied Predictive Modeling		Springer Science+Business Media New York
Optional reading				
Max Kuhn and Kjell	2015	Documentation of R package AppliedPredictiveModeling		https://cran.r-project.org/web/packages/Appl

Johnson				iedPredictiveModeling/AppliedPredictiveModeling.pdf
Max Kuhn	2016	Documentation of R package carret		http://topepo.github.io/caret/index.html