COURSE UNIT (MODULE) DESCRIPTION

Course unit (module) title	Code
Practical Data Analysis with R and Python	

Lecturer(s)	Department(s) where the course unit (module) is delivered
Coordinator: dr. Andrius Buteikis **Other(s):**	Faculty of Mathematics and Informatics Institute of Applied Mathematics Department of Statistical Analysis

Study cycle	Type of the course unit (module)
1st (BA)	Compulsory

Mode of delivery	Period when the course unit (module) is delivered	Language(s) of instruction
Face-to-face	4th semester	Lithuanian / English

Requirements for students	
Prerequisites: Statistics and probability theory	**Additional requirements (if any):** Microeconomics

Course (module) volume in credits	Total student's workload	Contact hours	Self-study hours
5	140	64	76

Purpose of the course unit (module): programme competences to be developed		
Ability to model phenomena using mathematical, statistical, and machine learning tools. Proficiency in using specialized statistical and machine learning software. Ability to independently analyze empirical data, select and apply appropriate machine learning methods. Ability to analyze and interpret the results of the conducted analysis.		
Learning outcomes of the course unit (module)	**Teaching and learning methods**	**Assessment methods**
To become familiar with the theoretical and practical aspects of supervised and unsupervised machine learning models.	Lectures, exercises using computer software, independent problem solving and studying of theoretical material, work in a computer class/laboratory.	Statistical program mastery testing, computer-based midterm and final exams, research project.
Be able to apply machine learning methods for sampling, model validation and forecasting.		
Be able to analyze textual information.		
Be able to perform model adequacy analysis.		

Content: breakdown of the topics	Contact hours							Self-study work: time and assignments	
	Lectures	Tutorials	Seminars	Exercises	Laboratory work	Internship/work placement	**Total contact hours**	**Self-study hours**	**Literature**
1. Statistical data types and models. R and Python programming languages, modern statistical modelling tools (RStudio, JupyterLab, VSCode, Docker).	2				2		**4**	**6**	[1]
2. Linear regression with Python: model functional form specification, parameter estimation, coefficient interpretation, standardized coefficients, confidence intervals, instrumental variables. Model adequacy analysis – statistical tests for autocorrelation and heteroskedasticity, checking for multicollinearity, model specification tests, coefficient of determination and information criterions.	4				2		**6**	**8**	[1] and [2]
3. Omitted variables, overfitting, outlier detection methods.	2				1		**3**	**2**	[1] and [2]
4. Classification, clustering and discrete response models.	6				2		**8**	**10**	[1] and [2]
5. Sampling-based methods for model validation: Monte Carlo method, cross-validation, bootstrap.	3				2		**5**	**6**	[1] and [2]
6. Dimension reduction methods and regularization.	3				2		**5**	**6**	[1] and [2]
7. Ensemble methods: gradient boosting, random forest, bootstrap aggregation (bagging), voting, stacking.	4				2		**6**	**6**	[1] and [2]
8. Deep learning: neural networks, cost function, weight (parameter) estimation, hyperparameters, activation functions, backpropagation algorithm.	6				3		**9**	**14**	[1] and [2]
9. Natural language processing: preparing, analysing and modelling text data.	8				4		**12**	**12**	[3]
10. Neural networks: MLP (multilayer perceptrons) and KAN (Kolmogorov-Arnold networks).	2				1		**3**	**3**	[1] and [4]
11. Introduction to generative (video, text and audio) artificial intelligence models.	2				1		**3**	**3**	[1]
Total	**42**				**22**		**64**	**76**	

Assessment strategy	Weight, %	Deadline	Assessment criteria
General evaluation scheme: To get a passing grade, at least 45 points are necessary (out of a maximum of 100 points). Taking the course on an external basis is not allowed.			
Intermediate exam	35	In the middle of the of semester	Examination using computer software. A total of 10 tasks from the topics covered during the lectures up to that point. Students analyse specific data and apply appropriate methods, statistical tests and write down the results.
Project presentation (defense)	30	The end of the semester	The student must complete the individually assigned project work by the specified time. If the written project is not submitted by the specified date, no points will be awarded. The assessment takes into account the suitability of the chosen methods, the consistency of the analysis performed, the suitability of the written program code and the formulation of the conclusions.
Final exam	35	Exam session	Examination using computer software with a total of 10 tasks. Students analyse specific data and apply appropriate methods, statistical tests and write down the results.

Author	Year of publication	Title	Issue of a periodical or volume of a publication	Online access or VU library
Compulsory reading				
1. Buteikis A.		Learning material		VLE
2. James, G., Witten, D., Hastie, T. and Tibshirani, R.	2021	An Introduction to Statistical Learning, 2nd Edition		Available Online: [with R] and [with Python]
3. Jurafsky D., Martin J. H.	2025	Speech and Language Processing, 3rd Ed.		Available Online: https://web.stanford.edu/~jurafsky/slp3/
Optional reading				
4. Liu Z., Wang Y., Vaidya S., Ruehle F., Halverson J., Soljačić M., Hou T. Y., Tegmark M.	2024	KAN: Kolmogorov-Arnold Networks		https://arxiv.org/abs/2404.19756 and https://github.com/KindXiaoming/pykan
5. Abu-Mostafa Y. S., Magdon-Ismail M., Lin H.T.	2012	Learning From Data: A short Course		https://amlbook.com/