



COURSE UNIT (MODULE) DESCRIPTION

Course unit (module) title	Code
Data mining	

Lecturer(s)	Department(s) where the course unit (module) is delivered
Coordinator: Erinija Pranckevičienė, PhD [EP] Other(s):	Department of Human and Medical Genetics, Institute of Biomedical Sciences, Faculty of Medicine, Vilnius University Santariskiu str. 2, LT-08661

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

Mode of delivery	Period when the course unit (module) is delivered	Language(s) of instruction
Face-to-face, self-study Lectures, seminars and practice	First semester	English

Requirements for students	
Prerequisites: English B2 level Linear algebra basics Programming basics Python coding basics UNIX/Linux basics Statistics	Additional requirements (if any):

Course (module) volume in credits	Total student's workload	Contact hours	Self-study hours
5	130	60	70

Purpose of the course unit (module): programme competences to be developed		
The unit aims at providing the basic concepts of data mining, teaching the students how to explore relevant concepts and sources for further development through theoretical lectures, exercises and case studies. The unit is also oriented towards applying the data analysis concepts on real life biomedical datasets.		
Learning outcomes of the course unit (module)	Teaching and learning methods	Assessment methods
Be able to work autonomously and as a part of a multidisciplinary team; act honestly and according to ethical obligations	Exercises, course projects	Completion of practical assignments, oral and written presentation of course project
Comprehend advanced data processing and programming techniques; be able to apply technologies used in data mining and basic data handling	Lectures, exercises, seminars with case studies	Completion of practical assignments, oral and written presentation of course project
Be able perform practical calculations using modern high-performance open computing platforms	Lectures, exercises, seminars with case studies	Completion of practical assignments, oral and written presentation of course project
Be able to analyse, manage and model data from the	Lectures, exercises, seminars	Completion of practical

field of system biology	with case studies	assignments, oral and written presentation of course project
Be able select an appropriate modelling strategy for a given biological domain and problem	Lectures, exercises, seminars with case studies	Completion of practical assignments, oral and written presentation of course project
Be able to gather and analyse information on subjects related to system biology with a critical approach, and to carry out a technological watch	Lectures, exercises, seminars with case studies	Completion of practical assignments, oral and written presentation of course project

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
Use of data mining in biomedical sciences. Technologies used in data mining, basic data handling: cleaning, integration, reduction. Outlier detection. Classification model selection and evaluation, bias-variance trade-off, over fitting, model evaluation by cross-validation. Public datasets. [EP]	4			4			8	6	Data handling exercises: data format transformations, cleaning, dealing with missing values and outliers.
Theory of statistical learning, linear modelling and parameter estimation, Lasso, regularization. [EP]	2			4			6	5	Exercises attempting to reproduce case studies from several selected papers
Supervised learning. Bayes rule. Fundamental state-of-the-art parametric linear and quadratic discriminant and non-parametric K nearest neighbours (KNN) and classification and decision tree (CART) classification methods. Introduction to kernel methods and Support Vector Machines. Multiple classifier systems. Mining frequent patterns. Classification model selection and evaluation, bias-variance trade-off, over fitting, model evaluation by cross-validation. [EP]	8			8			16	6	Classifier training, testing and validation exercises, case study from selected papers.
Unsupervised learning: K-means, hierarchical clustering, PCA, multidimensional scaling, data dimensionality reduction, manifold learning UMAP and tSNE. [EP]	4		2				6	6	Unsupervised learning exercises, case study from selected papers
Deep learning introduction: artificial neural networks (ANN), multi-task learning, deep learning frameworks [EP]	4			4			8	6	Textbook study, computational exercises
Deep learning models : convolutional neural networks, recurrent neural networks and transformers, representational learning, autoencoders and generative neural networks [EP].	6			6			12	6	Analysis of real-world datasets, case studies review from textbook examples.
Course project. [EP]			4				4	35	A data mining project, presentation and discussion of results
Total	28		6	26			60	70	

Assessment strategy	Weight, %	Deadline	Assessment criteria
8 completed assignments	30	Every week	Each assignment fully completed and uploaded by a due date 10 points.
Data mining project	30	6 th week of the course	<p>Written report consists of 2-3 pages; duration of presentation is 15 minutes;</p> <p>Project requirements outline:</p> <ol style="list-style-type: none"> 1) Same data set and analysis task will be given to each student 2) Student selects 4 methods introduced in the course and applies them to analyse the provided data set 3) Student compares results from different methods and selects the best method 4) Student submits analysis code to github <p>Project evaluation criteria:</p> <ol style="list-style-type: none"> 1) Mandatory description the dataset 2) Rationale for the chosen 4 methods 3) Description of cross-validation strategy 4) Reporting train and test errors/confusion table 5) Description of evidence supporting the best method for the given data analysis task. <p>Maximum grade 10 points distributed as follows: 2 points for quality of presentation 2 points for quality of visual material 4.5 points for ability to explain the problem and defend the chosen solution providing rationale and results 1.5 points for ability to answer questions</p>
Exam	40		10 open and multiple choice mixed questions

Author	Year of publication	Title	Issue of a periodical or volume of a publication	Publishing place and house or web link
Compulsary reading				
A.Geron	2019	Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition	ISBN: 9781492032649	O'Reilly Media, Inc.
I.Goodfellow et al.	2016	Deep learning	ISBN: 9780262035613	MIT Press
F.Chollet	2021	Deep learning with Python	ISBN: 9781617296864	Manning Publications
Optional reading				
		Code examples		https://courses.dataschool.io/ ; https://learn-scikit.oneoffcoder.com/index.html
Duda et al.	November 9, 2000	Pattern Classification	ISBN: 978-0471056690	Wiley-Interscience; 2 edition
G. James et al.	July 5, 2023	An Introduction to Statistical Learning with Applications in Python https://www.statlearning.com	EISBN: 978-3-031-38747-0	Springer Nature