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):** Vita Tomkutė, PhD, [VT] Juozas Gordevičius, PhD [JG]	Department of Human and Medical Genetics, Institute of Biomedical Sciences, Faculty of Medicine, Vilnius University Santariskiu str. 2, LT-08661 Department of Botany and genetics, Institute of Biosciences, Life Sciences Center, Vilnius University Saulėtekio al. 7, LT-10257

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 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 epxlore relevant concepts and sources for further development through theoretical lectures, excercises 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. [VT]	6			2			**8**	**6**	Data handling exercises:
Theory of statistical learning, linear modelling and parameter estimation, Lasso, regularization. [VT]	2		2	2			**6**	**5**	Exercises involving traditional statistical analysis of biomedical data, case study from 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		4	4			**16**	**6**	Classifier training, testing and validation exercises, case study from selected papers
Unsupervised learning: K-means, hierarchical clustering, PCA, multidimensional scaling. Assembly methods. [VT]	4			2			**6**	**6**	Unsupervised learning exercises, case study from selected papers
Deep learning introduction: artificial neural networks (ANN), Hebb rule (reinforcement learning). [EP]	4		4				**8**	**6**	Textbook study, computational exercises
Significant data mining applications in life sciences and biomedicine. [EP 50%, VT 50%]	6		6				**12**	**6**	Analysis of real-world biomedical datasets as part of data mining project, case studies review from selected papers
Course project. [VT]			4				**4**	**35**	A data mining project, presentation and discussion of results

Total	**30**		**20**	**10**			**60**	**70**	

Assessment strategy	Weight,%	Deadline	Assessment criteria
Oral discussion during a seminar lecture presenting a case study from selected papers	15	The final week	Students in groups of two analyse a provided article and presents the article overview (20 minutes for the article presentation – 10 minutes for each student followed by approximately 40 minutes discussion and questions). Presentation must include: 1) The main problem of the article 2) Dataset presentation 3) Methods in depth (highlight the methods presented in the course) 4) Results and their interpretation Maximum grade 10 consists of points distributed as follows: 2 points for the quality of presentation 2 points for the quality of visual material 2.5 points for explanation of the methods 2.5 points for explanation of results and interpretation questions 1 point for answers to questions
Data mining project	50	The final two lectures	Written report consists of 2-3 pages; duration of presentation is 15 minutes; Project requirements outline: 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 Project evaluation criteria: 1) Figure describing 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. 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
Exam	35		One open question (25%) and 10 multiple choice questions (75%)

Author	Year of publication	Title	Issue of a periodical or volume of a publication	Publishing place and house or web link
Compulsary reading				
J. Han	July 6, 2011	Data mining concepts and techniques (Third edition)	ISBN 978-9380931913	Morgan Kaufmann
T. Hastie et al.	April 12, 2011	The Elements of Statistical Learning	ISBN 978-0387848570	Springer; 2nd ed. 2009. Corr. 7th printing 2013 edition
G. James et al.	2017	An Introduction to Statistical Learning	ISBN 978-1461471370	http://www-bcf.usc.edu/~gareth/ISL/

		with Applications in R		
Selected papers for case studies				
Optional reading				
Manning et al.	July 7, 2008	Introduction to Information Retrieval	ISBN 978-0521865715	Cambridge University Press; 1 edition
Duda et al.	November 9, 2000	Pattern Classification	ISBN 978-0471056690	Wiley-Interscience; 2 edition
Goodfellow et al.	2016	Deep learning	ISBN 9780262035613	MIT Press