STUDIJŲ DALYKO (MODULIO) APRAŠAS

Dalyko (modulio) pavadinimas	Kodas
Įvadas į duomenų tyrybą ir mašininį mokymąsi	

Dėstytojas (-ai)	Padalinys (-iai)
Koordinuojantis: prof. habil. dr. Gintautas Dzemyda **Kitas (-i):** dr. Jolita Bernatavičienė	Matematikos ir informatikos fakultetas Duomenų mokslo ir skaitmeninių technologijų institutas

Studijų pakopa	Dalyko (modulio) tipas
Pirmoji	Pasirenkamas

Įgyvendinimo forma	Vykdymo laikotarpis	Vykdymo kalba (-os)
Auditorinė	5 arba 6 semestras	Lietuvių / Anglų

Reikalavimai studijuojančiajam	
Išankstiniai reikalavimai: Algoritmai ir duomenų struktūros, Procedūrinis programavimas, Statistiniai duomenų analizės metodai, Matematika informacinėms sistemoms, Optimizavimo metodai.	**Gretutiniai reikalavimai (jei yra):**

Dalyko (modulio) apimtis kreditais	Visas studento darbo krūvis	Kontaktinio darbo valandos	Savarankiško darbo valandos
5	133	48	85

Dalyko (modulio) tikslas: studijų programos ugdomos kompetencijos		
Dalyko tikslas – siekiama, kad studentai įgytų žinių apie mašininį mokymąsi, duomenų tyrybą ir vizualizavimą, susipažintų su daugiamačių duomenų vizualizavimo galimybėmis, daugiamačių duomenų struktūra, ugdytų praktinius gebėjimus taikyti mašininio mokymosi, duomenų tyrybos ir vizualizavimo metodus, ugdytų gebėjimą vizualiai tirti ir interpretuoti duomenis, suprastų ir analizuotų sudėtingus duomenų tyrybos uždavinius.		
Dalyko (modulio) studijų siekiniai	**Studijų metodai**	**Vertinimo metodai**
Gebės parinkti ir pritaikyti įvairius duomenų vizualizavimo sprendimus, planuoti mašininio mokymosi, duomenų tyrybos uždaviniams atlikti reikiamas veiklas.	Pavyzdžių nagrinėjimas, literatūros skaitymas ir analizė, savarankiškas darbas, laboratorinių darbų atlikimas, probleminis dėstymas.	Egzaminas, laboratorinių darbų vertinimas, darbas auditorijoje (diskusija).
Gebės atlikti literatūros, susijusios su daugiamačių duomenų struktūrų ir jų efektyviu apdorojimu paiešką ir analizę, įsisavinti žinias ir jas pritaikyti, sprendžiant vizualizavimo uždavinius.		
Gebės kritiškai įvertinti naudojamų mašininio mokymosi, duomenų tyrybos ir vizualizavimo algoritmų tinkamumą problemoms našiųjų skaičiavimų aplinkoje spręsti.		
Gebės tirti duomenų dimensiškumo įtaką klasifikavimo, klasterizavimo rezultatams interpretuoti gaunamus rezultatus.		
Gebės panaudoti esamus mašininio mokymosi, duomenų tyrybos sprendimus, skirtus apdoroti didžiuosius duomenis, atlikti jų statistinę analizę, taikyti daugiamačių duomenų vizualizavimo algoritmus.		

Temos	Kontaktinio darbo valandos							Savarankiškų studijų laikas ir užduotys	
	Paskaitos	Konsultacijos	Seminarai	Pratybos	Laboratoriniai darbai	Praktika	**Visas kontaktinis darbas**	**Savarankiškas darbas**	**Užduotys**
1. Duomenų tyrybos bei mašininio mokymosi galimybių apžvalga	**4**						**4**	**6**	Mokslinės literatūros skaitymas ir analizė, laboratorinių darbų atlikimas, ataskaitų rengimas, pasiruošimas laboratorinių darbų gynimui.
2. Pagrindinės duomenų statistinės charakteristikos	**2**				2		**4**	**6**	
3. Pirminio duomenų apdorojimo metodai	**2**				4		**6**	**6**	
4. Duomenų dimensijos mažinimo metodai ir vizualizavimas (tiesioginio vizualizavimo metodai; projekcijos metodai: pagrindinės komponentės, daugiamatės skalės, savireguliuojantys neuroniniai tinklai SOM, lokaliai tiesinis vaizdavimas)	**4**				4		**8**	**19**	
5. Klasterizavimas (dalinimo, hierarchiniai ir tankiu grindžiami metodai; klasterių skaičiaus įvertinimas; klasterizavimo kokybės vertinimas)	**5**				6		**11**	**14**	
6. Klasifikavimas (klasifikavimo metodai: sprendimų medžiai, bajesinis klasifikatorius, tiesinė diskriminantinė analizė, k-artimiausių kaimynų, atraminių vektorių mašina; klasifikatoriaus kokybės vertinimas; klasifikatorių palyginimas; klasifikatorių ansambliai)	**5**				8		**13**	**14**	
7. Įvadas į dirbtinius neuroninius tinklus (dirbtinis neuronas, vienasluoksnis ir daugiasluoksnis perceptronas)	**2**						**2**	**20**	
Iš viso	**24**				**24**		**48**	**85**	

Vertinimo strategija	Svoris proc.	Atsiskaitymo laikas	Vertinimo kriterijai
Laboratoriniai darbai	50	Semestro metu (pagal semestro pradžioje pateiktą atsiskaitymų tvarkaraštį)	Studentai privalo atlikti 4 praktines užduotis, paruošti rašto darbus (ataskaitas) ir jas individualiai apginti atsakant į užduodamus klausimas. • 9–10 balai – darbas tenkina visus būtinus reikalavimus: rašto darbo struktūra aiški ir logiška, yra visos reikiamos dalys (darbo tikslas, uždaviniai, eksperimentų vykdymo aprašas, rezultatai, išvados), darbas tinkamai suformatuotas, rezultatų analizė išsami, išvados pagrįstos, studentas atsako į užduotus klausimus; • 7–8 balai – yra visos reikiamos rašto darbo dalys (darbo tikslas, uždaviniai, eksperimentų vykdymo aprašas, rezultatai, išvados), tačiau yra trūkumų: ne visos darbo dalys tinkamai suformatuotos, yra loginių klaidų, rezultatų analizė nepakankamai išsami, ne visos išvados yra pagrįstos, studentas atsako ne į visus užduotus klausimus. • 5–6 balai – yra ne visos būtinos rašto darbo dalys, analizė yra paviršutiniška ir fragmentiška, ne visos išvados pagrįstos, yra esminių loginių klaidų, ne į visus esminius klausimus studentas geba atsakyti. • 1–4 balai – rašto darbas blogai struktūrizuotas, turi labai daug esminių klaidų, studentas neatsako į klausimus. • 0 balų – rašto darbas nėra pateiktas arba nėra apgintas. Yra laikoma, kad studentas yra atsiskaitęs už darbą, kai rašto darbas įkeltas į sistemą ir apgintas. Pavėlavus 2 savaites atsiskaityti, įvertinimas mažinamas dviem balais, 3 savaites – trimis balais. Vėliau darbas nebus vertinamas. Praktines užduotis būtina atsiskaityti ne vėliau kaip paskutinę praktinių užsiėmimų paskaitą.
Aktyvumas per paskaitas ir laboratorinius darbus (diskusija)	10	Kiekviena semestro savaitė	Aktyvumas kiekvieno užsiėmimo metu vertinamas iki 1 balo. Suminis vertinimas apskaičiuojamas surinktą balų skaičių dalinant iš užsiėmimų skaičiaus.
Egzaminas	40	Egzaminų sesijos metu	Egzaminą leidžiama laikyti studentams, atsiskaičiusiems visus laboratorinius darbus ir surinkusiems ne mažiau kaip 2 balus semestro metu (įskaitant aktyvumą per užsiėmimus). Egzamino metu studentai atsako į 3 pateiktus klausimus.

Autorius	Leidimo metai	Pavadinimas	Periodinio leidinio Nr. ar leidinio tomas	Leidimo vieta ir leidykla ar internetinė nuoroda
Privaloma literatūra				
M. J. Zaki, W. Meira Jr.	2020	Data mining and analysis. fundamental concepts and algorithms, Second Edition		Cambridge University Press https://dataminingbook.info/book_html/
Mohri, Mehryar; Rostamizadeh, Afshin; Talwalkar, Ameet	2018	Foundation of Machine Learning		London, The MIT Press https://virtualibiblioteka.vu.lt/permalink/f/1pp6fcs/VUB01000955876 https://d1rkab7tlqy5f1.cloudfront.net/EWI/Over%20de%20faculteit/Afdelingen/Intelligent%20Systems/Pattern%20Recognition%20Laboratory/PR/Reading%20Group/Foundations_of_Machine_Learning.pdf
G. Dzemyda, O. Kurasova, J. Žilinskas	2013	Multidimensional Data Visualization: Methods and Applications	Springer Optimization and Its Applications, Vol. 75	Springer https://www.researchgate.net/publication/266046175_Multidimensional_data_visualization_Methods_and_applications
G. Dzemyda, O. Kurasova, J. Žilinskas	2008	Daugiamačių duomenų vizualizavimo metodai		Matematikos ir informatikos institutas
Papildoma literatūra				
J. Leskovec, A. Rajaraman, J. Ullman	2015	Mining of Massive Datasets		http://www.mmds.org/
J. Han, M. Kamber, J. Pei	2012	Data Mining. Concepts and Techniques	Third Edition	Elsevier http://hanj.cs.illinois.edu/bk2/toc.pdf

COURSE UNIT (MODULE) DESCRIPTION

Course unit (module) title	Code
Introduction to data mining and machine learning	

Lecturer(s)	Department(s) where the course unit (module) is delivered
Coordinator: Prof. Habil. Dr. Gintautas Dzemyda **Other(s):** Dr. Jolita Bernatavičienė	Faculty of Mathematics and Informatics Institute of Data Science and Digital Technologies

Study cycle	Type of the course unit (module)
First	Optional

Mode of delivery	Period when the course unit (module) is delivered	Language(s) of instruction
face-to-face	5th or 6th semester	Lithuanian / English

Requirements for students	
Prerequisites: Algorithms and Data Structures, Procedural Programming, Statistical Data Analysis Methods, Math for Information Systems Engineering, Optimization Methods	**Additional requirements (if any):**

Course (module) volume in credits	Total student's workload	Contact hours	Self-study hours
5	133	48	85

Purpose of the course unit (module): programme competences to be developed		
The aim of the course unit is to introduce to a process during which the useful knowledge is extracted from data, conclusions and generalizations are made. Introduce to and develop abilities to use the data mining and machine learning techniques and methods that allow the better understanding of the structure of analysed data – clusters, outliers, similarities (dissimilarities) of objects in the bigger context, make reasoning on data visually and interpret the data from different standpoints.		
Learning outcomes of the course unit (module)	**Teaching and learning methods**	**Assessment methods**
Ability to select and apply various data visualization solutions, plan data collection, pre-processing activities to solve data mining tasks.	Case studies, analysis of scientific literature, individual work, laboratory tasks, problem-oriented teaching	Exam, evaluation results of laboratory tasks, discussion
Ability to conduct a proper literature search and analysis with the view to depict the appropriate multidimensional data structures, to gather new knowledge and apply it for data visualization.		
Ability to evaluate data mining, machine learning methods suitability and applicability for high-performance computing tasks.		
Ability to investigate the data dimensionality influence to data classification, clustering, interpret results.		
Ability to apply existing data mining, machine learning tools for multidimensional data analysis, big data processing, to perform statistical analysis, to solve data mining and data visualization tasks.		

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. Overview of data mining and machine learning opportunities	4						**4**	**6**	Work with scientific literature and its analysis, preparation of laboratory works and their reports, preparation for the defence of laboratory work.
2. Main statistical characteristics of the multidimensional data	2				2		**4**	**6**	
3. Data pre-processing	2				4		**6**	**6**	
4. Dimensionality reduction methods and visualisation (direct visualisation methods; projection methods: principal components PCA, multidimensional scaling MDS, self-organizing neural networks SOM, locally linear embedding LLE)	4				4		**8**	**19**	
5. Clustering (partitioning, hierarchical and density-based methods; estimating the number of clusters; assessing the quality of clustering)	5				6		**11**	**14**	
6. Classification (classification methods: decision trees, Bayesian classifier, linear discriminant analysis, k-nearest neighbours classifier, support vector machine; assessing the quality of a classifier; comparison of classifiers; ensemble of classifiers)	5				8		**13**	**14**	

7. Introduction to artificial neural networks (artificial neuron, single-layer and multilayer perceptron)	2						**2**	**20**	
Total	**24**				**24**		**48**	**85**	

Assessment strategy	**Weight, %**	**Deadline**	**Assessment criteria**
Laboratory works	50%	During the semester (according to the timetable given at the beginning of the semester)	The students must perform 4 practical tasks, present the reports, and defend them individually by answering the questions asked. • 9-10 points – the report satisfies all the necessary requirements: the report structure is clear and logical, there are all the necessary parts (work aim, tasks, description of experiments, results, conclusions), the report is properly formatted, the analysis of the results is comprehensive, the conclusions are reasoned, the student answers all questions asked. • 7-8 points – all the required parts of the report are presented (work aim, tasks, description of experiments, results, conclusions), but there are shortcomings: not all parts of the report are properly formatted, there are logical errors, the analysis of the results is not sufficiently detailed, not all the conclusions are justified, and the student does not answer all the questions asked. • 5-6 points – not all the necessary parts of the report are presented, the analysis is superficial and fragmentary, not all the conclusions are justified, there are fundamental logical errors, the student is not able to answer all the essential questions. • 1-4 points – the report is poorly structured, it has a large number of fundamental errors, the student does not answer the questions. • 0 points – the report is not presented, or it is not defended. A student is considered to have completed the practical task when the report is uploaded to the system and defended. If the delay is 2 weeks, the grade is reduced by 2 points; if the delay is 3 weeks, the grade is reduced by 3 points. Thereafter the practical task will not be evaluated. Practical tasks must be submitted no later than the last practical lecture.
Activeness (discussion) during lectures and laboratory works	10%	Each week of the semester	Activeness during each lecture or laboratory work is evaluated maximum up to 1 point. The total mark is derived by summing the points above and dividing by maximal possible number of lectures and laboratory works.
Exam	40%	During exam session	In the exam, students answer 3 questions. To take the exam is possible only for the student which has successfully defended all semester exercises during lectures and laboratory works and has the total sum of marks for laboratory works and activeness not less than 2.

Author	**Year of publication**	**Title**	**Issue of a periodical or volume of a publication**	**Publishing place and house or web link**
Compulsory reading				
M.J.Zaki, W.Meira Jr.	2020	Data mining and analysis. fundamental concepts and algorithms, Second Edition		Cambridge University Press https://dataminingbook.info/book_html/
Mohri, Mehryar ; Rostamizadeh, Afshin ; Talwalkar, Ameet	2018	Foundation of Machine Learning		London, The MIT Press https://virtualibiblioteka.vu.lt/permalink/f/1pp6fcs/VUB01000955876 https://d1rkab7tlqy5f1.cloudfront.net/EWI/Over%20de%20faculteit/Afdelingen/Intelligent%20Systems/Pattern%20Recognition%20Laboratory/PR/Reading%20Group/Foundations_of_Machine_Learning.pdf
G. Dzemyda, O. Kurasova, J. Žilinskas	2013	Multidimensional Data Visualization: Methods and Applications	Springer Optimization and Its Applications, Vol. 75	Springer https://www.researchgate.net/publication/266046175_Multidimensional_data_visualization_Methods_and_applications
Optional reading				
J. Leskovec, A. Rajaraman, J. Ullman	2015	Mining of Massive Datasets		http://www.mmds.org/
J. Han, M. Kamber, J. Pei	2012	Data Mining. Concepts and Techniques	Third Edition	Elsevier http://hanj.cs.illinois.edu/bk2/toc.pdf