UČNI NAČRT PREDMETA / COURSE SYLLABUS (leto / year 2016/17)			
Predmet:	Odkrivanje znanj iz podatkov		
Course title:	Data mining		

Študijski program in stopnja **Study programme and level**	**Študijska smer** **Study field**	**Letnik** **Academic year**	**Semester** **Semester**
Interdisciplinarni magistrski študijski program Računalništvo in matematika	ni smeri	1 ali 2	drugi
Interdisciplinary Master's study programme Computer Science and Mathematics	none	1 or 2	second

Vrsta predmeta / Course type	izbirni / elective
Univerzitetna koda predmeta / University course code:	63525

Predavanja **Lectures**	**Seminar** **Seminar**	**Vaje** **Tutorial**	**Klinične vaje** **work**	**Druge oblike študija**	**Samost. delo** **Individ. work**	**ECTS**
45	**20**	**10**			**105**	**6**

Nosilec predmeta / Lecturer:	prof. dr. Blaž Zupan

Jeziki / Languages:	**Predavanja / Lectures:**	slovenski / Slovene, angleški / English
	Vaje / Tutorial:	slovenski / Slovene, angleški / English

Pogoji za vključitev v delo oz. za opravljanje študijskih obveznosti:	**Prerequisites:**
Vpis v letnik študija. Opravljen predmet Uvod v odkrivanje znanj iz podatkov.	Enrolment in the programme. Completed course Introduction to data mining.
Vsebina:	**Content (Syllabus outline):**

Predmet bo v teoriji in na praktičnih primerih predstavil sledeče vsebine: 1. Predstavitev področja in klasifikacija tehnik za odkrivanje znanj iz podatkov, pregled značilnih aplikacij 2. Tehnološke platforme in razvojne metodologije (skriptna okolja, okolja za analizo podatkov z vizualnim programiranjem) 3. Predobdelava podatkov: iskanje osamelcev, zmanjševanje dimenzij (metoda glavnih komponent), izbor in konstrukcija značilk, permutacijski pristopi, diskretizacija 4. Uvrščanje v skupine, s poudarkom na tehnikah, ki lahko obravnavajo velike množice podatkov in podatkov z velikim naborom značilk, metode podpornih vektorjev, iskanje in vizualizacija interakcij 5. Tehnike razvrščanja v skupine (metode hierarhičnega združevanja, metode voditeljev), s poudarkom na tehnikah, ki lahko obravnavajo velike množice podatkov, določanje števila skupin (metoda silhuete) 6. Ocenjevanje uspešnosti napovednih modelov, kalibracijske in diskriminantne metode, ROC analiza, permutacijski pristopi 7. Vizualizacija podatkov in modelov, tehnike gradnje, analize in vizualizacije mrež 8. Tehnike odkrivanj znanj iz zbirk besedil in spletnih strani 9. Integrativni pristopi (uporaba predznanja, integracija povezav, pridobljenih iz različnih naborov podatkov) 10. Tipične napake pri snovanju pristopov ali uporabi tehnik odkrivanja znanj iz podatkov in kako se jim izognemo Na predavanjih bodo študenti spoznavali ključne tehnologije in orodja, s katerimi bodo tekom semestra na vajah in v okviru projektov	The course will cover theoretical and practical aspects of the following data mining approaches: 1. Introduction to data mining, taxonomy of data mining approaches and tasks 2. Data mining programming environments (scripting, visual programming) 3. Data preprocessing (dimensionality reduction, feature construction, identification of outliers) 4. Classification, including support vector machines and feature interaction discovery 5. Clustering, with emphasis on techniques that can consider very large data sets, and techniques for to determine an appropriate number of clusters 6. Evaluation, including permutation-based and cross-validation approaches, statistical scoring of models 7. Data and model visualization techniques, visualization of networks 8. Text mining, text-based kernels for support vector machines 9. Integrative aspects, including ensemble methods and mining with inclusion of prior knowledge 10. Typical mistakes in data mining and how to avoid them The course will be composed of lectures in core data mining techniques and tools, which will then be employed on practical problems during lab work. We will focus on open source solutions and modern scripting languages (e.g., Python). Students will use scripting to access various data

oz. seminarskih nalog reševali praktične probleme. Poudarek bo na uporabi odprtokodnih, prosto dostopnih orodij, ki za analizo podatkov uporabljajo moderne skriptne jezike (npr. Python). V skriptnih okoljih bodo študenti z uporabo že obstoječih komponent razvijali lastne metode, uporabo teh preverjali na različnih podatkih, ter poročali o ocenah njihove uporabnosti in napovedne točnosti. Vaje se bodo izvajale v računalniški učilnici opremljeni z ustrezno strojno in programsko opremo.	mining techniques which will they, in a programming framework, combine into their own data mining procedures.

Temeljni literatura in viri / Readings:

1. Tan P-N, Steinbach M, Kumar V (2006) Introduction to data mining. Pearson & Addison Wesley, New York.Education, Boston.

2. I. H. Witten, E. Frank (2005) Data mining: practical machine learning tools and techniques, Elsevier, Amsterdam.

3. Dokumentacija okolja za odkrivanje znanj iz podatkov Orange, prosto dostopna na spletnih straneh www.ailab.si/orange/doc.

Cilji in kompetence:	**Objectives and competences:**
Cilj predmeta je študente seznaniti z osnovnimi in naprednimi metodami odkrivanja znanj iz podatkov, s poudarkom na njihovi praktični uporabi. Pri predmetu se bodo naučili uporabljati moderna skriptna orodja za analizo podatkov. Spoznali bodo, kako je z njimi moč implementirati nove metode za odkrivanje znanj, oziroma kako je moč obstoječe tehnike prilagoditi za obravnavo konkretnih podatkov.	Students will learn a number of core techniques for data mining. The course will include an introduction to data mining as well as a detailed study of several selected methods. It will also focus on practical use of these methods on real-life problems. The course will use a scripting data mining environment, where students will learn how to use the existing data mining libraries and design and implement in code their own data mining solutions.

Predvideni študijski rezultati:	**Intended learning outcomes:**
Znanje in razumevanje: Poznavanje metod in orodij odkrivanja znanj iz podatkov, uporaba teh v skriptnih okoljih, poznavanje načinov gradnje sistemov za analizo podatkov iz obstoječih komponent za vizualizacijo, statistiko in strojno učenje.	Knowledge and understanding: Knowledge of methods and tools for data mining, their utility within modern data mining suites, engineering skills to construct (develop in code) data mining process from existing data analysis components.

Uporaba:

Uporaba tehnik odkrivanja znanj iz podatkov na praktičnih primerih s področja družboslovja, tehnike in biomedicine.

Refleksija:

Razumevanje primernosti teoretičnih metod za reševanje praktičnih primerov ter njihovih omejitev, sposobnost analitičnega razmišljanja, sposobnost analize in reševanja praktičnih problemov z razvojem inteligentnih sistemov.

Prenosljive spretnosti - niso vezane le na en predmet:

Kombiniranje znanj, pridobljenih pri predmetih Strojnega učenja in Umetna inteligenca. Spretnosti iskanja in uporabe domače in tuje literature, uporaba primerne (predvsem odprtokodne) programske opreme, identifikacija in reševanje kompleksnih problemov.

Application:

Application of data mining methods and tools on real-life data.

Reflection:

Which are appropriate practical applications of theoretical methods of data analysis? What are their limitations? How can intelligent data analysis systems be used in practice?

Transferable skills:

Students will be able to combine the knowledge from other courses that cover machine learning and artificial intelligence. The course will require students to acquire skills in literature search and search for existing algorithmic solutions and code snippets, and engineering skills for solving real-life complex problems.

Metode poučevanja in učenja:

Predavanja s podporo avdio-vizualne opreme, sprotni razvoj programskih rešitev, laboratorijske vaje v računalniški učilnici z ustrezno programsko opremo. Delo posamezno in v skupinah. Velik poudarek na praktičnem delu (npr. razvoj skript za pregledovanje in analizo podatkov) in

Learning and teaching methods:

Combined lecturing with simultaneous use of the blackboard and computer projection (coding, visualization of models, results). Lab work in computer-equipped lecture rooms. Individual and work in team. Emphasis on practical problem solving.

reševanju praktičnih problemov.	

Načini ocenjevanja:	Delež (v %) / Weight (in %)	**Assessment:**
Način (pisni izpit, ustno izpraševanje, naloge, projekt):		Type (examination, oral, coursework, project):
Sprotno preverjanje (domače naloge,		Continuing (homework, midterm exams,
kolokviji in projektno delo)		project work)
Končno preverjanje (pisni in ustni izpit)	50%	Final (written and oral exam)
Ocene: 6-10 pozitivno, 1-5 negativno		Grading: 6-10 pass, 1-5 fail (according to
(v skladu s Statutom UL)	50%	the rules of University of Ljubljana)

Reference nosilca / Lecturer's references:

BELLAZZI, Riccardo, ZUPAN, Blaž. Predictive data mining in clinical medicine : current issues and guidelines. International journal of medical informatics, ISSN 1386-5056. [Print ed.], 2008, vol. 77, no. 2, str. 81-97, ilustr. [COBISS.SI-ID 6280788]

MRAMOR, Minca, LEBAN, Gregor, DEMŠAR, Janez, ZUPAN, Blaž. Visualization-based cancer microarray data calssification analysis. Bioinformatics, ISSN 1367-4803. [Print ed.], 2007, vol. 23, no. 16, str. 2147-2154, ilustr. [COBISS.SI-ID 6087252]

VAN DRIESSCHE, Nancy, DEMŠAR, Janez, BOOTH, Egzi O., HILL, Paul, JUVAN, Peter, ZUPAN, Blaž, KUSPA, Adam, SHAULSKY, Gad. Epistasis analysis with global transcriptional phenotypes. Nature genetics, ISSN 1061-4036, May 2005, vol. 37, no. 5, str. 471-477, ilustr. [COBISS.SI-ID 4712532]

ZUPAN, Blaž, DEMŠAR, Janez, BRATKO, Ivan, JUVAN, Peter, HALTER, John A., KUSPA, Adam, SHAULSKY, Gad. GenePath : a system for automated construction of genetic networks from mutant data. Bioinformatics, ISSN 1367-4803. [Print ed.], 2003, vol. 19, no. 3, str. 383-389. [COBISS.SI-ID 3415124]

ZUPAN, Blaž, BOHANEC, Marko, DEMŠAR, Janez, BRATKO, Ivan. Learning by discovery concept hierarchies. Artificial intelligence, ISSN 0004-3702. [Print ed.], 1999, vol. 109, str. 211-242. [COBISS.SI-ID 14228007]