FACULTY OF PURE AND APPLIED MATHEMATICS

SUBJECT CARD

Name of subject in Polish: Pozyskiwanie wiedzy
Name of subject in English: Data mining
Main field of study (if applicable): Applied Mathematics
Specialization (if applicable): Data engineering
Profile: academic / practical*
Level and form of studies: 2nd level / full-time /
Kind of subject: optional
Subject code
Group of courses YES

	Lecture	Classes	Laboratory	Project	Seminar
Number of hours of organized classes in University (ZZU)	30		30		
Number of hours of total student workload (CNPS)	90		60		
Form of crediting	crediting with grade				
For group of courses mark (X) final course	X				
Number of ECTS points	3		2		
including number of ECTS points for practical classes (P)	2		2		
including number of ECTS points corresponding to classes that require direct participation of lecturers and other academics (BU)	1,5		1,5		

*delete as not necessary

PREREQUISITES RELATING TO KNOWLEDGE, SKILLS AND OTHER COMPETENCES

1. Introduction to probability theory
2. Introduction to mathematical statistics
3. Introduction to programming

\

SUBJECT OBJECTIVES

C1 Knowledge of basic data mining tasks
C2 Knowledge of classical and modern approaches used for classification, dimension reduction and cluster analysis
C3 Knowledge of procedures used to evaluate the performance of classification or cluster analysis algorithms
C4 Use of acquired knowledge in solving practical problems from different areas of science, technology and economics

SUBJECT EDUCATIONAL EFFECTS

relating to knowledge:
PEU_W01 has knowledge related to different data mining tasks

PEU_W02 knows basic methods/algorithms used for classification, dimension reduction, cluster analysis, association rules discovery, and knows properties of these methods
PEU_W03 knows procedures used in evaluating quality of classification or clustering results
relating to skills:
PEU_U01 can choose appropriate methods to solve a given data exploration task
PEU_U02 knows how to use both supervised and unsupervised learning algorithms
PEU_U03 knows how to evaluate the performance of data mining procedures
relating to social competences:
PEU_K01 can, without assistance, search for necessary information in the literature and acquire knowledge independently
PEU_K02 understands the need for systematic and independent work on mastery of course material

PROGRAMME CONTENT

Lecture		Number of hours
Lec 1	Introduction to the basic concepts of data mining. The types of data exploration tasks.	2
Lec 2	Data preparation for data mining analysis: handling missing values, identification of outliers and necessary transformations.	4
Lec 3	Dimension reduction methods: Principal Components Analysis (PCA), Multidimensional Scaling (MDS).	4
Lec 4	Methods used for data classification: k-nearest neighbors (K-NN), classification tree, naive Bayes classifier, discriminant analysis, logistic regression.	6
Lec 5	Cluster analysis. Partitioning and hierarchical methods (k-means, PAM, AGNES, DIANA).	4
Lec 6	Evaluation of the quality of classification and clustering results.	2
Lec 7	Support Vector Machines (SVM).	2
Lec 8	Ensemble methods in classification: bagging, boosting, random forest.	2
Lec 9	Introduction to the association rules mining.	2
Lec 10	Final test.	2
	Total hours	30

Laboratory		Number of hours
Lab 1	Introduction to R statistical environment.	2
Lab 2	Data structures and elements of programming in R.	2
Lab 3	Exploratory analysis of multivariate data.	2
Lab 4	Data preparation (handling missing values, identification of outliers and necessary data transformations).	2
Lab 5	Dimension reduction methods (PCA, MDS).	3
Lab 6	K-nearest neighbors (K-nn) algorithm and classification trees.	2

Lab 7	Discriminant analysis and logistic regression.	3
Lab 8	Cluster analysis – partitioning algorithms (k-means, PAM).	2
Lab 9	Cluster analysis – hierarchical algorithms (AGNES, DIANA, MONA).	2
Lab 10	Evaluation of classification and cluster analysis results.	3
Lab 11	Support Vector Machines (SVM).	2
Lab 12	Classifier ensembles: bagging, boosting, random forest	3
Lab 13	Association rules discovery.	2
	Total hours	30

TEACHING TOOLS USED

N1. Lecture – traditional method
N2. Computer lab classes
N3. Consultations
N4. Student's self work – preparation for the classes

EVALUATION OF SUBJECT LEARNING OUTCOMES ACHIEVEMENT

Evaluation (F – forming during semester), P – concluding (at semester end)	Learning outcomes code	Way of evaluating learning outcomes achievement
F1	PEK_U01, PEK_U02, PEK_U03, PEK_K01, PEK_K02,	Oral presentations, written reports, individual projects.
F2	PEK_W01, PEK_W02, PEK_W03, PEK_K01, PEK_K02,	Test
P = 60%F1 + 40%F2		

PRIMARY AND SECONDARY LITERATURE

PRIMARY LITERATURE:

[1] P.-N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, Addison-Wesley, 2006.

[2] G.James, D.Witten, T.Hastie, R.Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, 2017.

[3] T.Hastie, R.Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2017.

[4] D.T. Larose, Discovering Knowledge in Data: An Introduction to Data Mining, Wiley, 2005.

[5] D.T. Larose, Data Mining Methods and Models, Wiley, 2006.

SECONDARY LITERATURE:

[1] Ch. M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics). Springer, 2006.

[2] W.N. Venables, B.D. Ripley, Modern Applied Statistics With S, Springer, 2001.

[3] R.A. Johnson, D.W. Wichern, Applied multivariate statistical analysis, Pearson Prentice Hall, 2002.

SUBJECT SUPERVISOR (NAME AND SURNAME, E-MAIL ADDRESS)

dr inż. Adam Zagdański (Adam.Zagdanski@pwr.edu.pl)