Warsaw University of Technology

COURSE OFFERED IN THE DOCTORAL SCHOOL

Code of the course	4606-EW-0000000-0012	Name of the course	Polish	**Eksploracja danych i uczenie maszynowe w badaniach naukowych**	
			English	**Data exploration and machine learning in scientific research**	
Type of the course	~~przedmioty ogólne~~ / warsztat badacza / ~~przedmioty specjalnościowe~~				
Course coordinator	Dr inż. Hubert Anysz, MBA				
Implementing unit	WIL	Scientific discipline / disciplines*			
Level of education	Kształcenie doktorantów	Semester	zimowy/letni (dowolny)		
Language of the course	~~polski~~/angielski				
Type of assessment:	~~zaliczenie~~/ zaliczenie na ocenę/~~egzamin~~	Number of hours in a semester	45	ECTS credits	3
Minimum number of participants	12	Maximum number of participants	20	Available for students (BSc, MSc)	~~Yes~~/No

Type of classes		Lecture	Auditory classes	Project classes	Laboratory	Seminar
Number of hours	in a week	1	1		1	
	in a semester	15	15		15	

* does not apply to the Researcher's Workshop

1. Prerequisites

Knowledge of mathematics at the engineering level

Basic knowledge of statistics

Ability to use Microsoft Excel

Installing the Statistica package (available at PW for students and employees) from the 4th class and Solver for Excel from the second class.

Necessary computer room to conduct classes.

2. Course objectives

The aim of the course is to teach practical skills of finding the relationships between the studied phenomenon and many independent variables accompanying this phenomenon, describing them, without the need to acquire coding skills in the Python environment, R, etc. In particular, PhD students will acquire the skills to:

- find the rules governing the studied phenomenon

- analysis of the relevance of these rules

- determining the relative significance of the influence of each of the independent variables on the studied phenomenon

- finding relationships between independent variables describing a given phenomenon and the studied phenomenon itself

by:

- building predictive models,
- building classification models,
- data mining
- analysing the results of previous studies.

As a result - which is also the aim of the subject - doctoral students will acquire the skills of drawing conclusions regarding the studied phenomena on the basis of the assessed quality of the above-mentioned models. Therefore, they will also be able to assess the quality (goodness) of fitting models to the modeled phenomena, both in their works and in the works of others, presented in the scientific literature.

The stress will be put on the analysis of the results of tests, examinations with a limited number of samples (several dozen, several hundred), on the basis of which it is difficult to conclude, assess the quality of the models, rather than on the data covering several thousand samples.

3. Course content (separate for each type of classes)

Lecture

Lectures:

1.1. Introduction, course organizational issues and requirements, basic of statistics, Introduction to data exploration and machine learning
2. Classification and prediction errors as a measure of models' accuracy
3. Data pre-processing
4. Association analysis, rule finding
5. Introduction to artificial neural networks
6. Artificial neural networks (ANN) in classification problems
7. Artificial neural networks (ANN) in regression problems
8. Optimizing ANN models
9. Decision trees in classification problems
10. Decision trees in regression problems
11. Multivariate regression
12. Support vector machine (SVM); random forest (RF), generalized additive models
13. Reasoning based on models
14. Examples of machine learning applications
15. Introduction to deep learning, recurrent networks, conclusions, final remarks

Laboratory

Exercises in the computer room with the use of Microsoft Excel i Statistica software.

1. Introductory pre-test
2. Calculations of errors in Excel
3. Standardizing of datasets, Pearson's correlation matrix in Excel
4. Rule finding in Statistica software
5. Creating artificial neural networks in Excel
6. Creating ANN classification model in Statistica software
7. Creating ANN regression model in Statistica software
8. Optimizing ANN models in Statistica software
9. Creating decision tree (DT) classification model in Statistica software
10. Creating decision tree (DT) regression model in Statistica software
11. Creating multivariate regression models in Statistica and Excel
12. Creating SVM and RF in Statistica software
13. Discussing results achieved at meetings 4 to 12
14. Discussing results presented in already published scientific articles
15. Conclusions, final remarks, final test

4. Learning outcomes

	Learning outcomes description	Reference to the learning outcomes of the WUT DS	Learning outcomes verification methods*
	Knowledge		
K01	to the extent enabling the revision of the existing paradigms - global achievements, including theoretical foundations and general issues and selected specific issues - appropriate for the represented scientific	SD_W2	Auditorium exercises (evaluation of the quality of models

	discipline, including the latest scientific achievements in the field of research		from the scientific literature)
K02	main development trends of the research discipline pursued and related research methodologies	SD_W3	Auditorium exercises (evaluation of the quality of models from the scientific literature)
K03	to the extent enabling the revision of the existing paradigms - global achievements, including theoretical foundations and general issues and selected specific issues - appropriate for the represented scientific discipline, including the latest scientific achievements in the field of research	SD_W2	Auditorium exercises (evaluation of the quality of models from the scientific literature)
Skills			
S01	use knowledge from various fields for creative identification, formulation and innovative solving of complex problems or performing research tasks, in particular: • define the aim and subject of research, formulate a research hypothesis; • develop and creatively use research methods, techniques and tools; • correctly conclude on the basis of research results	SD_U1	Auditorium exercises (evaluation of the quality of models from the scientific literature) and test of the lecture
S02	perform a critical analysis and evaluation of the results of scientific research, expert activity and other creative works and their contribution to the development of knowledge, in particular assess the usefulness and possibility of using the results of theoretical work in practice	SD_U2	Auditorium exercises (evaluation of the quality of models from the scientific literature) and test of the lecture
S03	communicate on specialist topics relevant to the represented scientific discipline to a degree enabling active participation in the national and international scientific community, including international consortia of research universities	SD_U4	Auditorium exercises (evaluation of the quality of models from the scientific literature)
Social competences			
SC01	critical evaluation of the achievements of the represented scientific discipline, including one's own contribution to the development of this discipline	SD_K1	Auditorium exercises (evaluation of the quality of models from the scientific literature)

*Allowed learning outcomes verification methods: exam; oral exam; written test; oral test; project evaluation; report evaluation; presentation evaluation; active participation during classes; homework; tests

5. Assessment criteria
60 % praca na ćwiczeniach (w sali komputerowej), 40 % wynik testu z wykładów

6. Literature
Literatura podstawowa: [1] D.T. Larouse, Discovering Knowledge in Data, an Introduction to Data Mining, Wiley, 2014 [2] User's guide for Statistica (https://docs.tibco.com/pub/stat/14.0.0/doc/html/UsersGuide/suitehelp_topic_list.html) [3] A.D. Aczel, Complete Business Statistics, IRWIN, Boston MA, 1993 [4] A. Saha, Shareware Excel models for Artificial Neural Networks classification, Artificial Neural Networks predictions and Decision Trees Literatura uzupełniająca: [1] H. Anysz, Machine Learning and data mining tools applied for databases of low number of records, 2022 Advanced Engineering Research [2] H. Anysz, Ł. Brzozowski, W. Kretowicz, P. Narloch, Feature Importance of Stabilised Rammed Earth Components Affecting the Compressive Strength Calculated with Explainable Artificial Intelligence Tools, 2020 Materials [3] H. Anysz, M. Dąbrowska, The Risk Indicators of Construction Projects' Cost Overruns assessed with PCA, Decision Trees, and Pearson's Correlations, XXX Russian-Polish-Slovak Seminar Theoretical Foundation of Civil Engineering (RSP 2021) Publisher: Springer International Publishing [4] H. Anysz, J. Rosłon, A. Foremny, 7-Score Function for Assessing the Strength of Association Rules Applied for Construction Risk Quantifying, Applied Sciences, 2022 Publikacje naukowe z uznanych czasopism wykorzystujące narzędzia eksploracji danych i uczenia maszynowego.

7. PhD student's workload necessary to achieve the learning outcomes**		
No.	Description	Number of hours
1	Hours of scheduled instruction given by the academic teacher in the classroom	45
2	Hours of consultations with the academic teacher, exams, tests, etc.	5
3	Amount of time devoted to the preparation for classes, preparation of presentations, reports, projects, homework	30
4	Amount of time devoted to the preparation for exams, test, assessments	10
	Total number of hours	**90**
	ECTS credits	**3**

** 1 ECTS = 25-30 hours of the PhD students work (2 ECTS = 60 hours; 4 ECTS = 110 hours, etc.)