Zał. nr 5 do ZW 78/2023

Attachment no. 4. to the Program of Studies

FACULTY OF CHEMISTRY

SUBJECT CARD

Name of subject in Polish Uczenie maszynowe w chemii i biologii
Name of subject in English Machine Learning for Chemistry and Biology
Main field of study (if applicable): Bioscences
Specialization (if applicable):
Profile: academic / practical*
Level and form of studies: 2nd level, full-time
Kind of subject: obligatory
Subject code W03BSS-SM2013W, W03BSS-SM2013L
Group of courses NO

	Lecture	Classes	Laboratory	Project	Seminar
Number of hours of organized classes in University (ZZU)	30		30		
Number of hours of total student workload (CNPS)	50		50		
Form of crediting	Examination		crediting with grade*		
For group of courses mark (X) final course					
Number of ECTS points	2		2		
including number of ECTS points for practical classes (P)			2		
including number of ECTS points corresponding to classes that require direct participation of lecturers and other academics (BU)	1.3		1.4		

*delete as not necessary

PREREQUISITES RELATING TO KNOWLEDGE, SKILLS AND OTHER COMPETENCES

1. Fundamentals of Physical Chemistry
2. Understanding the structure of bioorganic molecules
3. Fundamentals of mathematical analysis and linear algebra
4. Pre-intermediate experience with python scripting

\

SUBJECT OBJECTIVES

C1 To familiarize the students with the fundamentals of machine learning and deep learning methods.
C2 To familiarize the students with possible applications of machine learning models in chemistry and biology.
C3 Acquiring the ability to identify and apply the most appropriate machine learning methods to solve a given research problem or analyze data.
C4 Learning how to evaluate the trained models and interpret their results.

SUBJECT EDUCATIONAL EFFECTS

relating to knowledge:

PEU_W01 Knows the basic strategies and algorithms of supervised and unsupervised learning.

PEU_W02 Has knowledge of common applications of machine learning methods in chemistry and biology.

PEU_W03 Is able to assess the strengths, weaknesses and limitations of individual machine learning methods in applications to various problems in the field of computational biology.

PEU_W04 Has knowledge of good practices in training machine learning models to avoid overtraining and identify potential shortcomings in the training data set.

PEU_W05 Knows various forms of representation of the structure of bioorganic molecules, including commonly used geometry formats (xyz, pdb, zmat, smiles, smarts, sdf) as well as representations dedicated to machine learning.

PEU_W06 Knows the formats and representations of data that can be used to train machine learning models.

relating to skills:

PEU_U01 Is able to effectively select and prepare a representative data set in the appropriate format for a given machine learning method.

PEU_U02 Can apply supervised learning models for data classification.

PEU_U03 Can apply unsupervised learning models for data clustering.

PEU_U04 Can conceptually/schematically describe an algorithm to solve a given research problem or data analysis problem.

PEU_U05 Can implement an algorithm to solve a given research problem or data analysis problem using the Python scripting language.

PEU_U06 Can evaluate machine learning models and interpret the results they offer.

relating to social competences:

PEU_K01 Students are able to work in a group, performing various roles, including group leader

PEU_K02 Students are aware of the social role of an MSc in Bioinformatics

PEU_K03 Students are ready to critically evaluate his knowledge and the received content

PROGRAMME CONTENT

Lecture		Number of hours
Lec 1	Introduction to machine learning. Explanation of the term machine learning and its relation to the so-called artificial intelligence. To familiarize students with the general classification of supervised and unsupervised learning methods. An overview of the most popular applications of machine learning in science, engineering, and life sciences.	2
Lec 2	Machine learning datasets. Data sources and representative data formats that can be used for machine learning. Sources of data errors. Good practices in data selection.	2

Lec 3	Supervised learning - artificial neural networks I. A brief history of artificial neural networks and similarities to biological networks. Research directions and applications of neural networks. Linear networks.	2
Lec 4	Supervised learning - artificial neural networks II. Training a neural network using the gradient descent method and back propagation. Rosenblatt perceptron. Multilayer and deep networks. Detailed application examples.	2
Lec 5	Supervised learning - other methods. Support vector machines, kernel ridge regression, decision trees, random forest.	2
Lec 6	Unsupervised learning. Description of the basic methods of unsupervised learning. Classification and grouping. Train the model to recognize features that characterize the data set.	2
Lec 7	Structural biology I. Introduction/review of selected issues in structural biology concerning the structure and dynamics of proteins and nucleic acids. Predicting the secondary structure of peptides from sequences.	2
Lec 8	Structural biology II. Predicting the structure of biomolecules - AlphaFold and nucleic acids.	2
Lec 9	Machine learning models in molecular simulations I. Introduction/review of the elements of computational chemistry. Potential and free energy surfaces. Classification of various methods in computational chemistry including machine learning potentials.	2
Lec 10	Machine learning models in molecular simulations II. Representation of the geometry/structure of molecules in machine learning. Training of models to reproduce the shape of the potential energy surface and selection of the data set. Advantages and disadvantages of neural networks and kernel ridge regression.	2
Lec 11	Machine learning models in molecular simulations III. Learning molecular properties. Non-bonding interactions, oxidation states and electron configurations.	2
Lec 12	Drug design. Interaction of the drug with the active site. Methods of estimating the free energy of active substance binding in the active site.	2
Lec 13	Prediction of synthetic pathways to organic molecules. Reaxys database. SMARTS and SMILES structure formats. Approaches to prediction of organic synthesis pathways using retrosynthesis.	2
Lec 14	Image analysis and medical applications. Examples and methods of analyzing diagnostic images using machine learning	2
Lec 15	Revision of the most important topics presented during the lectures. Preparation for the exam, discussion and questions.	2
	Total hours	30

Laboratory		**Number of hours**
Lab 1	Organization of work in the computer laboratory and computing center. Discussion of the principles of occupational health and safety. Account distribution and basic information about available operating systems. Reminder of elements and selected commands of the LINUX operating system. Basic information about the operating system. Using Anaconda and Jupyter Notebooks.	2
Lab 2	Introduction to the basics of statistics using the Pandas module. Tasks: histograms, block plots, exploration of pseudo-random number generation, meshing histograms with Pandas; binomial, Poisson and normal distributions. Introduction to the SciKit-learn library in python.	4

Lab 3	Data visualization and dimensionality reduction - introduction and exercises. Tasks: use of block charts to visualize many variables simultaneously. Correlation analysis between data based on heat maps.	4
Lab 4	Data classification - introduction and exercises. Tasks: classification of white and red wines on the basis of physical and chemical properties. Assessment of the accuracy of the trained model.	4
Lab 5	Regression methods - introduction and exercises. Tasks: regularization.	4
Lab 6	Structural biology: Grouping of biomolecular structures using the DBSCAN algorithm. Sequence based peptide secondary structure prediction.	4
Lab 7	Training models for molecular simulations based on DFT calculations - models based on kernel ridge regression (AQML) and neural networks (ANI).	4
Lab 8	Work on individual projects. Presentation of reports on the implementation of individual projects.	4
	Total hours	30

TEACHING TOOLS USED
N1. Presentation. N2. Problem solving in a small-group setting. N3. Implementation of solutions to problems and realization of tasks in a computer laboratory.

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	PEU_U01-PEU_U06, PEU_K01-03	Grading mid-term reports (max 50 points)
P1	PEU_U01-PEU_U06	Grading the final report and project (max 50 points)
P2	PEU_W01-PEU_W06	Exam grade (max 100 points)
P (lab classes) 2.0 if (F1+P1) < 50 points 3.0 if (F1+P1) = 50 - 59 points 3.5 if (F1+P1) = 60 - 69 points 4.0 if (F1+P1) = 70 - 79 points 4.5 if (F1+P1) = 80 - 89 points 5.0 if (F1+P1) = 90 - 97 points 5.5 if (F1+P1) = 98 - 100 points P (lecture) 2.0 if (P2) < 50 points 3.0 if (P2) = 50 - 59 points 3.5 if (P2) = 60 - 69 points 4.0 if (P2) = 70 - 79 points 4.5 if (P2) = 80 - 89 points		

5.0 if (P2) = 90 - 97 points
5.5 if (P2) = 98 - 100 points

PRIMARY AND SECONDARY LITERATURE

PRIMARY LITERATURE:

[1] A. Géron, Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, O'Reilly Media, Sebastopol, CA, 2020.
[2] B. Ramsunda, P. Eastman, P. Walters, V. Pande, Deep Learning for the Life Sciences, O'Reilly Media, Sebastopol, CA, 2019.

SECONDARY LITERATURE:

[1] Lafuente D. et al., A Gentle Introduction to Machine Learning for Chemists: An Undergraduate Workshop Using Python Notebooks for Visualization, Data Processing, Analysis, and Modeling, J. Chem. Educ. 2021, 98, 2892–2898
[2] Keith J.A. et al., Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems, Chem. Rev. 2021, 121, 9816–9872.
[3] Artrith N. et al., Best practices in machine learning for chemistry, Nat. Chem. 2021, 13, 505-508.

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

dr inż. Rafał Szabla, rafal.szabla@pwr.edu.pl