GDAŃSK UNIVERSITY OF TECHNOLOGY

Subject card

Subject name and code	, PG_00062084						
Field of study	Mathematics						
Date of commencement of studies	October 2023		Academic year of realisation of subject	2023/2024			
Education level	second-cycle studies		Subject group	Optional subject group Subject group related to scientific research in the field of study			
Mode of study	Full-time studies		Mode of delivery	at the university			
Year of study	1		Language of instruction	Polish			
Semester of study	1		ECTS credits	5.0			
Learning profile	general academic profile		Assessment form	assessment			
Conducting unit	Katedra Fizyki Teoretycznej i Informatyki Kwant. -> Faculty Of Applied Physics And Mathematics -> Wydziały Politechniki Gdańskiej						
Name and surname of lecturer (lecturers)	Subject supervisor		dr inż. Patryk Jasik				
	Teachers		dr inż. Patryk Jasik				
Lesson types and methods of instruction	Lesson type	Lecture	Tutorial	Laboratory	Project	Seminar	SUM
	Number of study hours	30.0	15.0	15.0	0.0	0.0	60
	E-learning hours included: 0.0						
Learning activity and number of study hours	Learning activity	Participation in didactic classes included in study plan		Participation in consultation hours	Self-study		SUM
	Number of study hours	60		5.0	60.0		125
Subject objectives	The main aim of the course is to introduce students to the tools and methods used to process and analyze large volumes of data (Big Data).						

Learning outcomes

Course outcome	Subject outcome	Method of verification
[K7_U13] Understands the mathematical foundations of the analysis of algorithms and computational processes, can construct algorithms with good numerical properties, used to solve typical and unusual mathematical problems.	The student understands the functioning of machine learning algorithms from a mathematical perspective.	[SU1] Assessment of task fulfilment [SU4] Assessment of ability to use methods and tools
[K7_W11] Knows the mathematical foundations of information theory, the theory of algorithms and cryptography and their practical applications, i.a. in programming and computer science.	The student is familiar with selected artificial intelligence methods and can apply them in practice. The student is also proficient in the Python language.	[SW3] Assessment of knowledge contained in written work and projects
[K7_W10] Knows the numerical methods used to find approximate solutions to mathematical problems (e.g. differential equations) posed by applied fields (e.g. industrial technologies, management, etc.).	The student is familiar with selected numerical methods that underlie machine learning algorithms, such as regression, classification, and clustering.	[SW3] Assessment of knowledge contained in written work and projects
[K7_W08] Knows advanced computation techniques, supporting the work of a mathematician and understand their limitations.	The student is knowledgeable about advanced computational tools and techniques for processing large volumes of data.	[SW2] Assessment of knowledge contained in presentation
[K7_K03] Can work as a team; understands the necessity of systematic work on all projects that are long-term in nature, understands and appreciates the importance of intellectual honesty in one’s own activities and the activities of other people; behaves ethically.	The student can carry out data science projects within a team and understands and appreciates the ethical aspects of teamwork.	[SK5] Assessment of ability to solve problems that arise in practice

Subject contents	1. Big Data a) What are large volumes of data - definitions. b) Scale. c) Advantages of using big data methods. d) Problems and challenges. 2. Data mining methods. 3. Data a) Data sources, data types, data quality. b) ETL (Extract, Transform, Load) process, data verification and validation, data cleaning, data consistency, data profiling, data standardization, data formatting. 4. Python Language a) Basic data types and operations on them. The print() function, the input() function, conditional statements, different types of loops, exceptions, lists, tuples, dictionaries, functions. b) Data analysis from a selected dataset, loading observations for selected variables, checking basic statistics for individual variables, plotting histograms, identifying variables with potentially erroneous data (observations) or missing data, data cleaning, calculating normalized correlations between variables, conducting linear regression for selected variables, including charts. c) scikit-learn package and linear regression model, R-squared coefficient, MSE, MAE, splitting the data into training and testing sets, predicting values using the created model. d) scikit-learn package and preprocessing, polynomial model, generating new features, reducing model variables - Bayesian Information Criterion (BIC), polynomial model operation in practice. e) scikit-learn package, k-nearest neighbors method, decision trees, and random forests, classification problem, feature selection - predictors and target variables, model parameters, model quality evaluation - confusion matrix, sensitivity, specificity, precision, accuracy, ROC curve, LIFT curve, cross-validation: k-fold, n-fold, and Monte-Carlo (bootstrap). f) scikit-learn package and k-means algorithm as a case of unsupervised learning, cluster analysis - clustering, model parameters, Fowlkes-Mallows index - concordance between two dataset divisions into clusters, Principal Component Analysis (PCA). g) Hyperparameter optimization of models. h) Elements of explainable artificial intelligence. i) Analyzing and modeling time series.
Prerequisites and co-requisites	Basic programming skills.
Assessment methods and criteria	

Subject passing criteria	Passing threshold	Percentage of the final grade
From Data to Insights with Google Cloud	50.0%	55.0%
Programming test	50.0%	15.0%
Class attendance	50.0%	15.0%
Presentation	50.0%	15.0%

<table>
<tr><td rowspan="3">Recommended reading</td><td>Basic literature</td><td>[1] Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer

[2] Nathan Marz, James Warren, Big Data. Najlepsze praktyki budowania skalowalnych systemów obsługi danych w czasie rzeczywistym, Helion

[3] Stanisław Osowski, Metody i narzędzia eksploracji danych, BTC</td></tr>
<tr><td>Supplementary literature</td><td>[1] Alan Agresti, An Introduction to Categorical Data Analysis, Wiley - Interscience

[2] Bradley Efron, Trevor Hastie, "Computer Age Statistical Inference. Algorithms, Evidence, and Data Science"</td></tr>
<tr><td>eResources addresses</td><td>Adresy na platformie eNauczanie:
Big Data 2023 - Moodle ID: 24037
https://enauczanie.pg.edu.pl/moodle/course/view.php?id=24037</td></tr>
<tr><td>Example issues/ example questions/ tasks being completed</td><td colspan="2">1. Completion of the "From Data to Insights with Google Cloud" course.
2. Programming exam: Prepare a selected dataset for analysis; perform exploratory data analysis on the chosen dataset; create a regression or classification model.
3. Presentation:
• Random Forest Algorithm.
• Neural Networks.</td></tr>
<tr><td>Work placement</td><td colspan="2">Not applicable</td></tr>
</table>

Document generated electronically. Does not require a seal or signature.