GDAŃSK UNIVERSITY OF TECHNOLOGY

Subject card

Subject name and code	, PG_00033237		
Field of study	Mathematics		
Date of commencement of studies	October 2022	Academic year of realisation of subject	2022/2023
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	4.0
Learning profile	general academic profile	Assessment form	assessment
Conducting unit	Department of Theoretical Physics and Quantum Information -> Faculty of Applied Physics and Mathematics		
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

Address on the e-learning platform: https://enauczanie.pg.edu.pl/moodle/course/view.php?id=11093

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	35.0	100

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_W10	The student knows the numerical methods used to process, analyze and model big data coming from various sources.	[SW3] Assessment of knowledge contained in written work and projects
K7_W07	The student knows the connections between data science and theoretical as well as applied mathematics.	[SW3] Assessment of knowledge contained in written work and projects
K7_U13	Student understands the mathematical fundamentals of algorithms and computational processes, can apply algorithms for processing, analyzing, and modeling of big data.	[SU1] Assessment of task fulfilment [SU4] Assessment of ability to use methods and tools

Subject contents	1. Big Data a) large volumes of data - definitions b) scale c) assets of using the big data methods d) problems and challenges 2. Data mining methodologies a) SEMMA b) CRISP-DM 3. Data a) data sources, the type of data, data quality b) ETL process (Extract, Transform, Load) • verification and validation of data • data cleaning • data consistency • data profiling • data standardization • formatting data c) loading data into databases and data warehouses 4. Data Mining (SAS Enterprise Miner, SAS Visual Analytics, SAS Visual Statistics, R, Python) a) Tasks • description • evaluating • prediction • categorizing • clustering • exploring rules b) Methods • data aggregation • correlation analysis • decision trees and random forests • regression models • neural networks • optimization models • time series forecasting models • analysis of textual data 5. Apache Hadoop a) the main functionalities of the platform • Hadoop Common • HDFS – Hadoop Distributed File System • Hadoop YARN • Hadoop MapReduce b) the ecosystem of the Hadoop platform on the example of Apache Spark

Prerequisites and co-requisites	Knowledge of the SAS software. Basic programming skills.

Assessment methods and criteria

Subject passing criteria	Passing threshold	Percentage of the final grade
Presentation	50.0%	20.0%
Class attendance	80.0%	20.0%
Test	50.0%	20.0%
Project	50.0%	40.0%

Recommended reading

Basic literature	[1] Trevor Hastie, Robert Tibshirani, Jerome Friedman, „The Elements of Statistical Learning: Data Mining, Inference, and Prediction”, Springer 2008. [2] Kristina Chodorow, „Mongodb: The Definitive Guide”, O'Reilly Media 2013
Supplementary literature	[1] Alan Agresti, “An Introduction to Categorical Data Analysis”, Wiley - Interscience 2007.
eResources addresses	Adresy na platformie eNauczanie:

Example issues/ example questions/ tasks being completed

1. Prepare the selected data set for analysis.
2. Perform the exploratory analysis of the chosen data set.
3. Describe the random forest algorithm.
4. Neural networks (presentation).

Work placement	Not applicable