GDAŃSK UNIVERSITY OF TECHNOLOGY

Subject card

Subject name and code	, PG_00062084		
Field of study	Big Data		
Date of commencement of studies	October 2025	Academic year of realisation of subject	2025/2026
Education level	second-cycle studies	Subject group	Specialty 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. Marcin Nowakowski dr inż. Patryk Jasik	

Lesson types

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

eNauczanie source addresses:
Moodle ID: 2417 Big Data 2025
https://enauczanie.pg.edu.pl/2025/course/view.php?id=2417

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 aim of the course is to familiarize students with tools and methods for processing, analyzing and modeling large volumes of data (Big Data).

Learning outcomes

Course outcome	Subject outcome	Method of verification
[K7_U09] constructs mathematical models used in specific advanced applications of mathematics, can use stochastic processes as a tool for modeling phenomena and analyzing their evolution,constructs mathematical models used in specific advanced applications of mathematics, uses stochastic processes as a tool for modeling phenomena and analyzing their evolution, recognizes mathematical structures in physical theories	The student creates models using machine learning and artificial intelligence methods to solve specific problems.	[SU4] Ocena umiejętności korzystania z metod i narzędzi
[K7_W06] analyzes 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 knows the mathematical foundations of information theory, algorithm theory and cryptography and is able to apply them practically using the Python language and its packages as well as selected data engineering tools.	[SW3] Ocena wiedzy zawartej w opracowaniu tekstowym i projektowym
[K7_W03] demonstrates knowledge advanced computation techniques, supporting the work of a mathematician and understand their limitations.	The student demonstrates knowledge of the Python language and its packages, as well as selected data engineering tools, and is able to use advanced computational techniques that support the work of a mathematician, understanding their limitations.	[SW3] Ocena wiedzy zawartej w opracowaniu tekstowym i projektowym

Subject contents

Course content – lecture
Big Data: what large data volumes aredefinitions; scale; benefits of applying big data methodologies; problems and challenges.
Data mining methods.
Working with data: data sources, data types, data quality.
ETL processes (extract, transform, load): data verification and validation; data cleaning; data consistency; data profiling; data standardization; data formatting.
Introduction to quantum computing.
Introduction to quantum optimization methods and quantum machine learning.

Course content – exercises
Overview of machine learning methods for regression.
Overview of machine learning methods for classification.
Overview of machine learning methods for clustering.
Overview of methods for time series analysis and modeling.
Overview of quantum optimization methods and quantum machine learning.

Course content – laboratory

Python
Basic data types and operations on them. Conditional statements. Various types of loops. Exceptions. Lists, tuples, dictionaries.
Functions. Functions with optional arguments. Modules. Creating your own module and using it. Reading/ writing data to/from files. YAML files. Classes and objects. Introduction to multithreading.
Analysis of data from a selected dataset. Loading observations for selected variables. Checking basic statistics for individual variables. Plotting histograms. Identifying variables with potentially erroneous or missing data (observations).
Data cleaning. Computing normalized correlations between variables. Performing linear regression for selected variables, with plots.
The scikit-learn package and the linear regression model. Coefficient of determination R2, MSE, MAE. Splitting the dataset into training and test parts. Predicting values using the trained model.
The scikit-learn package and preprocessing. Polynomial models. Feature generation. Reducing model variablesthe Schwarz criterion (BICBayesian Information Criterion). Practical use of a polynomial model.
The scikit-learn package: k-nearest neighbors, decision trees, and random forests. A classification problem. Feature selectionpredictors and the target variable. Model parameters. Model evaluationconfusion matrix, sensitivity, specificity, precision, accuracy, ROC curve, LIFT curve. Cross-validation: k-fold, n-fold, and Monte Carlo (bootstrap).
The scikit-learn package and the k-means algorithm as a case of unsupervised learning. Cluster analysisclustering. Model parameters. The FowlkesMallows index (agreement between two clusterings). Principal Component AnalysisPCA.
Hyperparameter optimization.
Elements of Explainable Artificial Intelligence (XAI).
Time series analysis and modeling.
Working with the Qiskit package.

Prerequisites and co-requisites

Basic programming skills.

<table>
<tr><td>Assessment methods and criteria</td><td colspan="2"><table>
<tr><th>Subject passing criteria</th><th>Passing threshold</th><th>Percentage of the final grade</th></tr>
<tr><td>Completion of selected specialized courses on the DataCamp portal</td><td>60.0%</td><td>70.0%</td></tr>
<tr><td>Attendance at classes</td><td>80.0%</td><td>30.0%</td></tr>
</table></td></tr>
<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></td></tr>
<tr><td>Example issues/ example questions/ tasks being completed</td><td colspan="2">DataCamp - Introduction to Python (4h)
DataCamp - Data Manipulation with pandas (3h)
DataCamp - Exploratory Data Analysis in Python (4h)
DataCamp - Supervised Learning with scikit-learn (3h)</td></tr>
<tr><td>Practical activites within the subject</td><td colspan="2">Not applicable</td></tr>
</table>

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