

KARTA KURSU

Nazwa	Inżynieria i Analiza Danych	
Nazwa w j. ang.	Data Science	
Koordynator	dr Roman Czapla	Zespół dydaktyczny
		dr Roman Czapla mgr Katarzyna Marczak mgr Patryk Mazurek
Punktacja ECTS*	4	

Opis kursu (cele kształcenia)

Celem kursu jest zapoznanie studentów z metodami i narzędziami przetwarzania, analizy oraz wizualizacji danych z wykorzystaniem języka Python i jego bibliotek. Kurs łączy teoretyczne podstawy analizy danych, takie jak statystyka i algebra liniowa, z praktycznymi technikami inżynierii danych, przygotowywania i czyszczenia zbiorów, eksploracji oraz prognozowania.

Studenci poznają narzędzia do pracy z danymi numerycznymi i tabelarycznymi, techniki wizualizacji oraz wybrane metody uczenia maszynowego. Zajęcia przygotowują do samodzielnego rozwiązywania rzeczywistych problemów analitycznych, integrowania danych z różnych źródeł oraz prezentowania wyników w formie raportów i wizualizacji. Kurs stanowi fundament do dalszej nauki w zakresie zaawansowanej analizy danych i data science.

Warunki wstępne

Wiedza	Student posiada podstawową wiedzę z zakresu programowania strukturalnego i obiektowego. Rozumie podstawowe pojęcia związane z algorytmiką, strukturami danych oraz podstawami matematyki wyższej, obejmującymi analizę matematyczną, statystykę opisową i algebrę liniową na poziomie wstępnym.
Umiejętności	Student potrafi pisać proste programy w języku Python, wykorzystując zmienne, struktury danych, instrukcje sterujące i funkcje. Umie korzystać z podstawowych narzędzi środowiska programistycznego (np. Jupyter Notebook) oraz potrafi analizować proste zbiory danych przy użyciu bibliotek Pandas i NumPy. Potrafi korzystać z dokumentacji technicznej i materiałów dydaktycznych w języku angielskim.
Kursy	<u>Wymagane zaliczenie kursu:</u> Języki skryptowe

Efekty kształcenia

	Efekt kształcenia dla kursu	Odniesienie do efektów kierunkowych
	Po zakończeniu kursu student:	
	W01: zna podstawowe metody inżynierii danych oraz narzędzia do przetwarzania i analizy danych w języku Python.	K_W06 K_W07
	W02: zna biblioteki NumPy i Pandas oraz rozumie ich zastosowania w pracy z danymi numerycznymi i tabelarycznymi.	K_W06 K_W07
	W03: posiada wiedzę z zakresu statystyki opisowej, rozkładów prawdopodobieństwa i podstawowych testów	K_W03 K_W06

Wiedza	statystycznych oraz zna elementy algebry liniowej istotne dla analizy danych.	K_W07
	W04: rozumie metody czyszczenia i przygotowywania danych, w tym skalowanie, normalizację i kodowanie zmiennych kategoriycznych.	K_W03 K_W06 K_W07
	W05: zna narzędzia i techniki wizualizacji danych (Matplotlib, Seaborn) oraz podstawy analizy szeregów czasowych.	K_W02 K_W05
	W06: zna wybrane metody uczenia maszynowego, w tym regresję, klasyfikację, redukcję wymiarowości i analizę skupień, oraz rozumie podstawowe metody oceny jakości modeli.	K_W02 K_W03 K_W06 K_W07
Umiejętności	Efekt kształcenia dla kursu	Odniesienie do efektów kierunkowych
	Po zakończeniu kursu student:	
	U01: potrafi pracować w środowisku Jupyter Notebook, korzystając z bibliotek Python do przetwarzania i analizy danych.	K_U04 K_U05
	U02: umie tworzyć, modyfikować i analizować struktury danych w NumPy i Pandas, a także wczytywać i zapisywać dane w różnych formatach oraz łączyć się z bazami danych SQL/NoSQL.	K_U05 K_U06 K_U11
	U03: potrafi stosować metody statystyki opisowej i testów statystycznych do analizy zbiorów danych oraz wykorzystywać elementy algebry liniowej w praktycznych obliczeniach.	K_U06 K_U09
	U04: umie czyścić, przygotowywać i transformować dane do dalszej analizy, stosując odpowiednie techniki skalowania, standaryzacji i kodowania.	K_U05 K_U06
	U05: potrafi tworzyć różne formy wizualizacji danych, dobierać odpowiednie typy wykresów do rodzaju danych oraz prezentować wyniki w sposób czytelny i estetyczny.	K_U05 K_U06
	U06: potrafi analizować szeregi czasowe oraz stosować proste metody prognozowania.	K_U02 K_U05 K_U06
Kompetencje społeczne		K_U02 K_U05 K_U06 K_U13
	Efekt kształcenia dla kursu	Odniesienie do efektów kierunkowych

	Po zakończeniu kursu student: K01: rozumie znaczenie jakości danych i odpowiedniego doboru metod analitycznych w procesie podejmowania decyzji. K02: potrafi pracować indywidualnie i zespołowo nad zadaniami związanymi z analizą i inżynierią danych, prezentując wyniki w sposób zrozumiały dla różnych odbiorców. K03: dostrzega konieczność dalszego doskonalenia swoich umiejętności w zakresie zaawansowanej analizy danych, inżynierii danych i uczenia maszynowego.	K_K01 K_K03 K_K02
--	---	--

Studia stacjonarne

Organizacja											
Forma zajęć	Wykład (W)	Ćwiczenia w grupach									
		A		K		L		S		P	E
Liczba godzin	20					30					

Studia niestacjonarne

Organizacja											
Forma zajęć	Wykład (W)	Ćwiczenia w grupach									
		A		K		L		S		P	E
Liczba godzin	10					30					

Opis metod prowadzenia zajęć

Zajęcia mają formę wykładów oraz ćwiczeń laboratoryjnych. Podczas wykładów prowadzący wprowadza nowe zagadnienia teoretyczne oraz prezentuje przykłady problemów analitycznych wraz z możliwymi metodami ich rozwiązania.

Ćwiczenia koncentrują się na praktycznym zastosowaniu omawianych metod. Studenci samodzielnie lub w małych grupach przygotowują skrypty w języku Python, rozwiązujące zadane problemy. Następnie odbywa się wspólna analiza i dyskusja nad przygotowanymi rozwiązaniami, z naciskiem na poprawność, czytelność oraz efektywność kodu.

Formy sprawdzania efektów kształcenia

	E – learning	Gry dydaktyczne	Ćwiczenia w szkole	Zajęcia terenowe	Praca laboratoryjna	Projekt indywidualny	Projekt grupowy	Udział w dyskusji	Referat	Praca pisemna (esej)	Egzamin ustny	Egzamin pisemny	Inne
W01					x			x					
W02					x			x					
W03					x			x					

W04					X			X					
W05					X			X					
W06					X			X					
U01					X	X		X					X
U02					X	X		X					X
U03					X	X		X					X
U04					X	X		X					X
U05					X	X		X					X
U06					X	X		X					X
U07					X	X		X					X
K01								X					
K02								X					
K03								X					

Kryteria oceny	Podstawą zaliczenia kursu jest uzyskanie pozytywnej oceny z projektów zaliczeniowych związanych z analizą danych oraz z kolokwium końcowego obejmującego materiał teoretyczny i praktyczny. Osiągnięcie efektów kształcenia podanych powyżej uprawnia studentów do uzyskania oceny nie wyższej niż dostateczna. Ocenę dobrą lub bardzo dobrą może uzyskać student, który spełnia wymagania zaliczenia podstawowego, a ponadto uzyskuje wysoką liczbę punktów z projektów i kolokwium, potwierdzając umiejętność samodzielnego stosowania metod inżynierii i analizy danych. Na wyższą ocenę wpływa również: • poprawność merytoryczna i techniczna przygotowanych projektów, • umiejętność trafnej interpretacji wyników analizy, • stosowanie dobrych praktyk programistycznych (czytelność i organizacja kodu, komentarze, obsługa błędów), • umiejętność integrowania różnych narzędzi i metod w ramach jednego rozwiązania. Obecność na wykładach jest warunkiem koniecznym zaliczenia tej części kursu.
----------------	---

Uwagi	
-------	--

Treści merytoryczne (wykaz tematów)

1. Podstawy inżynierii danych i środowiska pracy • wprowadzenie do inżynierii danych, znaczenie i zastosowania, • praca w środowisku <code>Jupyter Notebook</code>, • biblioteki <code>NumPy</code> i <code>Pandas</code>: – tablice, wektory i macierze (tworzenie, indeksowanie, kształtowanie, operacje matematyczne i agregujące), – struktury danych <code>Pandas</code> (<code>DataFrame</code> i <code>Series</code>), zarządzanie indeksami, operacje filtrowania, grupowania i sortowania. 2. Podstawy statystyki i algebry liniowej w Pythonie • pojęcia statystyczne: typy danych, miary tendencji centralnej (średnia, mediana, moda), miary zmienności (wariancja, odchylenie standardowe, kwartyle), • analiza rozkładów danych: skośność, kurtoza, kowariancja, współczynnik korelacji, centralne twierdzenie graniczne, • rozkłady prawdopodobieństwa i estymacja parametrów, podstawowe testy statystyczne (t-test, test chi-kwadrat, analiza wariancji), • elementy algebry liniowej: operacje na macierzach (dodawanie, mnożenie, transpozycja, macierze odwrotne, wyznaczniki), rozwiązywanie układów równań liniowych, wartości i wektory własne, • generowanie liczb losowych i próbek danych (moduł <code>numpy.random</code>).

3. Praca z danymi – inżynieria danych
 - odczyt i zapis danych w formatach tekstowych (CSV, TXT), binarnych i JSON,
 - obsługa interfejsów sieciowych (API),
 - praca z bazami danych: łączenie z bazami SQL i NoSQL, wykonywanie zapytań w Pythonie (SQLAlchemy, pymongo), porównanie SQL i operacji Pandas,
 - czyszczenie i przygotowanie danych: identyfikacja i obsługa wartości brakujących, imputacja, usuwanie duplikatów, konwersje typów danych,
 - przekształcanie danych: skalowanie, normalizacja, standaryzacja, kodowanie zmiennych kategorycznych (one-hot encoding, label encoding),
 - zaawansowane przygotowanie danych: transformacja i rozdzielanie cech, tworzenie nowych zmiennych,
 - agregacja danych i operacje na grupach (groupby, funkcje agregujące, analiza dużych zbiorów danych).
4. Wizualizacja danych i szeregi czasowe
 - podstawy wizualizacji w Matplotlib: wykresy liniowe, słupkowe, punktowe, histogramy, wykresy pudełkowe, dostosowywanie wyglądu i podwykresy,
 - wizualizacje w Seaborn: heatmaps, wykresy rozrzutu, boxplots, pairplots,
 - zasady doboru wykresów do rodzaju danych, estetyka i czytelność prezentacji,
 - analiza szeregów czasowych: tworzenie i manipulowanie danymi czasowymi, analiza sezonowości i trendów, prognozowanie (ARIMA, SARIMA).
5. Uczenie nadzorowane
 - regresja liniowa i wielomianowa: modelowanie zależności, ocena jakości modelu (R^2 , MSE),
 - regresja logistyczna: klasyfikacja zmiennych binarnych, metryki jakości modelu (accuracy, precision, recall),
 - wybrane algorytmy klasyfikacyjne: drzewa decyzyjne, lasy losowe, naiwny klasyfikator Bayesa, k-najbliższych sąsiadów (KNN), maszyna wektorów nośnych (SVM),
 - podział danych na zbiory uczące i testowe, walidacja krzyżowa, ocena wyników modeli.
6. Uczenie nienadzorowane
 - redukcja wymiarowości danych: analiza głównych składowych (PCA) i inne metody,
 - analiza skupień: klasteryzacja K-means, DBSCAN, metody hierarchiczne i widmowe,
 - ocena jakości grupowania, wizualizacja wyników analizy skupień.

Wykaz literatury podstawowej

Po zakończeniu kursu student:

1. D. Y. Chen, *Jak analizować dane z biblioteką Pandas. Praktyczne wprowadzenie. Wydanie II*, Helion, Gliwice 2024;
2. D. Nolan, J. Gonzalez, S. Lau, *Poznaj Data Science. Przekształcanie, eksplorowanie, wizualizacja i modelowanie danych w Pythonie*, Promise, Warszawa 2024;
3. W. McKinney, *Python w analizie danych. Przetwarzanie danych za pomocą pakietów pandas i NumPy oraz środowiska Jupyter. Wydanie III*, Helion, Gliwice 2023;
4. T. Nield, *Podstawy matematyki w data science. Algebra liniowa, rachunek prawdopodobieństwa i statystyka*, Helion, Gliwice, 2023;
5. J. VanderPlas, *Python Data Science. Niezbędne narzędzia do pracy z danymi. Wydanie II*, Helion, Gliwice 2023;
6. K. Gallatin, Ch. Albon, *Uczenie maszynowe w Pythonie. Receptury. Od przygotowania danych do deep learningu. Wydanie II*, Helion, Gliwice 2023;
7. A.h Navlani, A. Fandango, I. Idris, *Python i praca z danymi. Przetwarzanie, analiza, modelowanie i wizualizacja. Wydanie III*, Helion, Gliwice 2022.

Wykaz literatury uzupełniającej

1. M. Walker, *Czyszczenie danych w Pythonie. Receptury. Nowoczesne techniki i narzędzia Pythona do wykrywania i eliminacji zanieczyszczeń oraz wydobywania kluczowych cech z danych*, Helion, Gliwice 2021;
2. R. Johansson, *Matematyczny Python. Obliczenia naukowe i analiza danych z użyciem NumPy, SciPy i Matplotlib*, Helion, Gliwice 2021;

Bilans godzinowy zgodny z CNPS (Całkowity Nakład Pracy Studenta) **studia stacjonarne**

liczba godzin w kontakcie z prowadzącymi	Wykład	20
	Konwersatorium (ćwiczenia, laboratorium itd.)	30
	Pozostałe godziny kontaktu studenta z prowadzącym	2
liczba godzin pracy studenta bez kontaktu z prowadzącymi	Lektura w ramach przygotowania do zajęć	15
	Przygotowanie krótkiej pracy pisemnej lub referatu po zapoznaniu się z niezbędną literaturą przedmiotu	
	Przygotowanie projektu lub prezentacji na podany temat (praca w grupie)	18
	Przygotowanie do egzaminu/zaliczenia	15
Ogółem bilans czasu pracy		100
Liczba punktów ECTS w zależności od przyjętego przelicznika		4

Bilans godzinowy zgodny z CNPS (Całkowity Nakład Pracy Studenta) **studia niestacjonarne**

liczba godzin w kontakcie z prowadzącymi	Wykład	10
	Konwersatorium (ćwiczenia, laboratorium itd.)	20
	Pozostałe godziny kontaktu studenta z prowadzącym	2
liczba godzin pracy studenta bez kontaktu z prowadzącymi	Lektura w ramach przygotowania do zajęć	25
	Przygotowanie krótkiej pracy pisemnej lub referatu po zapoznaniu się z niezbędną literaturą przedmiotu	
	Przygotowanie projektu lub prezentacji na podany temat (praca w grupie)	18
	Przygotowanie do egzaminu/zaliczenia	25
Ogółem bilans czasu pracy		100
Liczba punktów ECTS w zależności od przyjętego przelicznika		4