



3. KARTA PRZEDMIOTU

Kod przedmiotu	Z-IDN-U-504
Nazwa przedmiotu	Wstępna eksploracja i przygotowanie danych do analiz
Nazwa przedmiotu w języku angielskim	Preliminary Exploration and Preparation of Data for Analyses
Obowiązuje od roku akademickiego	2019/2020

USYTUOWANIE MODUŁU W SYSTEMIE STUDIÓW

Kierunek studiów	INŻYNIERIA DANYCH
Poziom kształcenia	I stopień
Profil studiów	Praktyczny
Forma i tryb prowadzenia studiów	Studia stacjonarne
Zakres	Wszystkie specjalności
Jednostka prowadząca przedmiot	Katedra Informatyki i Matematyki Stosowanej
Koordinator przedmiotu	Dr hab. Marzena Nowakowska
Zatwierdził	Dr hab. inż. Artur Bartosik, prof. PŚk

OGÓLNA CHARAKTERYSTYKA PRZEDMIOTU

Przynależność do grupy/bloku przedmiotów	Przedmiot kierunkowy
Status przedmiotu	Obowiązkowy
Język prowadzenia zajęć	Polski
Usytuowanie modułu w planie studiów - semestr	Semestr V
Wymagania wstępne	Algebra liniowa, Elementy rachunku prawdopodobieństwa i statystyki, Bazy danych, Logika, Zaawansowane zastosowania arkuszy kalkulacyjnych, Algorytmy i struktury danych, Akwizycja danych pomiarowych, Zarządzanie bazami danych - SQL
Egzamin (TAK/NIE)	TAK
Liczba punktów ECTS	4

Forma prowadzenia zajęć	wykład	ćwiczenia	laboratorium	projekt	inne
Liczba godzin w semestrze	9		18		

EFEKTY UCZENIA SIĘ

Kategoria	Symbol efektu	Efekty kształcenia	Odniesienie do efektów kierunkowych
Wiedza	W_01	Student wie, jak poszukiwać i pozyskiwać dane potrzebne do analiz.	ID1_W05 ID1_W12
	W_02	Student ma świadomość znaczenia jakości danych. Ma podstawową wiedzę w zakresie diagnozowania jakości danych oraz narzędzi wykorzystywanych do czyszczenia danych.	ID1_W05 ID1_W09 ID1_W12
	W_03	Student ma wiedzę dotyczącą przygotowania danych do analiz poprzez ich ewidencję (inventaryzację) i wstępną eksplorację.	ID1_W05 ID1_W13
	W_04	Student wie, jak i jakie narzędzia programistyczne stosować do eksploracji i przygotowania danych do analiz.	ID1_W12 ID1_W13
Umiejętności	U_01	Student potrafi pozyskać dane do analiz oraz scharakteryzować ich strukturę.	ID1_U04
	U_02	Student potrafi wykonać podstawową diagnozę jakości danych i wykonać podstawowe operacje czyszczenia tych danych.	ID1_U06 ID1_U16
	U_03	Student potrafi wykonać wstępną eksplorację danych ilościowych i jakościowych, wykorzystując właściwe oprogramowanie.	ID1_U06 ID1_U14
	U_04	Student potrafi zidentyfikować podstawowe problemy w danych oraz zaproponować rozwiązanie tych problemów, również z wykorzystaniem właściwego oprogramowania.	ID1_U14 ID1_U16
	U_05	Student umie opracować wyniki swojej pracy w formie dokumentacji lub raportu.	ID1_U02
Kompetencje społeczne	K_01	Student ma świadomość odpowiedzialności za pracę własną oraz ponoszenia odpowiedzialności za wspólnie realizowane zadania projektowe.	ID1_K04
	K_02	Student umie komunikować się w zespole interdyscyplinarnym w zakresie wykraczającym poza zagadnienia czysto techniczne.	ID1_K05

TREŚCI PROGRAMOWE

Forma zajęć	Treści programowe
wykład	1. Źródła danych i metody pozyskiwania danych do analiz (dane oportunistyczne, eksperymentalne). Diagnozowanie struktur danych. Klasyfikacja danych wg skal pomiarowych.
	2. Elementy rachunku prawdopodobieństwa i statystyki we wstępnym diagnozowaniu danych (repetitorium). Wstępna eksploracja danych kwantytatywnych; miary statystyczne, histogramy, miary współzależności cech ilościowych.
	3. Sposoby kodowania (przekształcania) danych jakościowych, kodowania stratne i niestratne. Wstępna eksploracja danych jakościowych; rozkłady, ocena zmiennych jakościowych, tablica wielodzielcza, badanie niezależności zmiennych jakościowych, miary asocjacyjne cech kategoriycznych i korelacje cech porządkowych.
	4. Spójność wewnętrzna i integralność referencyjna w bazach danych. Metody i narzędzia do weryfikacji i naprawy danych numerycznych (w tym danych typu data/godzina). Metody i narzędzia do weryfikacji i naprawy danych tekstowych.
	5. Jakość danych. Cechy definiujące jakość danych; poprawność, adekwatność, kompletność, spójność, jednolitość. Wzorce wyrażeń regularnych. Podstawowe typy zanieczyszczeń w danych; błędy syntaktyczne, semantyczne i rejestracji. Znaczenie słowników w procesie diagnozy jakości danych.

	6. Czyszczenie danych. Przyczyny złej jakości danych. Etapy czyszczenia danych. Zarządzanie wartościami odstającymi. Operacje na danych tekstowych w procesie czyszczenia danych. Pozyskiwanie i tworzenie zasobów zewnętrznych do weryfikacji danych. Możliwości wykorzystania zasobów wewnętrznych w weryfikacji danych. Raport z czyszczenia danych.
	7. Zarządzanie danymi brakującymi. Podstawowe mechanizmy generowania danych brakujących; MCAR, MAR, MNAR. Obsługa danych brakujących. Rodzaje imputacji.
	8. Pozostałe problemy z danymi; różnorodność mian, duże zbiory danych, liczne atrybuty, liczne rekordy, nierównomierny rozkład atrybutów. Metody rozwiązywania innych problemów z danymi. Przekształcenia normalizacyjne, metryki, kategoryzacja danych ilościowych, agregacja kategorii cechy jakościowej, agregacja cech jakościowych. Metody próbkowania i ich stosowanie.
laboratorium	1. Pozyskanie danych do analiz. Charakterystyka źródła danych i sposobu ich pozyskania, opis struktury pozyskanego zbioru danych i charakterystyka jego zawartości, Klasyfikacja danych wg formatów, strukturalności i skal pomiarowych. <u>Opracowanie raportu</u> . Sprawozdanie z wykonania prac z tematyki nr 1. Wnioski i zalecenia.
	2. Obsługa środowiska programu SAS Enterprise Guide - przygotowanie do inwentaryzacji i wstępnej eksploracji danych. Eksploracja danych ilościowych. Miary statystyczne i ilustracja graficzna. <u>Opracowanie raportu</u> . Sprawozdanie z wykonania prac z tematyki nr 2. Wnioski i zalecenia.
	3. Eksploracja danych jakościowych. Korelacje między cechami jakościowymi. Ilustracja graficzna. <u>Opracowanie raportu</u> . Sprawozdanie z wykonania prac z tematyki nr 3. Wnioski i zalecenia.
	4. Ocena i poprawa jakości danych z wykorzystaniem zasobów wewnętrznych repozytorium. <u>Opracowanie raportu</u> . Sprawozdanie z wykonania prac z tematyki nr 4. Wnioski i zalecenia.
	5. Akwizycja danych zewnętrznych na potrzeby weryfikacji jakości danych. Ocena i poprawa jakości danych z wykorzystaniem pozyskanych zasobów zewnętrznych. Statystyka poprawek. <u>Opracowanie raportu</u> . Sprawozdanie z wykonania prac z tematyki nr 5. Wnioski i zalecenia.
	6. Obsługa danych brakujących. Metody prostej imputacji. Porównanie i ocena wybranych metod imputacji. Obliczenia miar oceny dla wybranego przykładu. <u>Opracowanie raportu</u> . Sprawozdanie z wykonania prac z tematyki nr 6. Wnioski i zalecenia.
	7. Samodzielna praca studentów (zespoły dwuosobowe). Opracowanie raportu omawiającego przykłady nieprawidłowych krotek w bazie danych
	8. Samodzielna praca studentów (zespoły dwuosobowe). Opracowanie raportu omawiającego trzy przykłady skutków złej jakości danych w wybranym obszarze aktywności zawodowej człowieka.

METODY WERYFIKACJI EFEKTÓW UCZENIA SIĘ

Symbol efektu	Metody sprawdzania efektów kształcenia (zaznaczyć X)					
	Egzamin ustny	Egzamin pisemny	Kolokwium	Projekt	Sprawozdanie	Inne
W_01					X	
W_02		X			X	
W_03		X			X	
W_04					X	
U_01		X			X	
U_02					X	
U_03		X			X	
U_04					X	
U_05					X	
K_01					X	
K_02					X	

FORMA I WARUNKI ZALICZENIA

Forma zajęć	Forma zaliczenia	Warunki zaliczenia
wykład	egzamin	Uzyskanie co najmniej 50% punktów z egzaminu.
laboratorium	zaliczenie z oceną	Uzyskanie co najmniej 50% punktów ze sprawozdań opracowywanych na zajęciach, będąc członkiem zespołu dwuosobowego oraz uzyskanie co najmniej 50% punktów z każdego raportu samodzielnego wskazanego w treściach programowych laboratorium w punktach 7 i 8.

NAKŁAD PRACY STUDENTA

Bilans punktów ECTS							
Lp.	Rodzaj aktywności	Obciążenie studenta					Jednostka
		W	C	L	P	S	
1.	Udział w zajęciach zgodnie z planem studiów	9		18			h
2.	Inne (konsultacje, egzamin)	4		2			h
3.	Razem przy bezpośrednim udziale nauczyciela akademickiego	33					h
4.	Liczba punktów ECTS, którą student uzyskuje przy bezpośrednim udziale nauczyciela akademickiego	1,3					ECTS
5.	Liczba godzin samodzielnej pracy studenta	67					h
6.	Liczba punktów ECTS, którą student uzyskuje w ramach samodzielnej pracy	2,7					ECTS
7.	Nakład pracy związany z zajęciami o charakterze praktycznym	67					h
8.	Liczba punktów ECTS, którą student uzyskuje w ramach zajęć o charakterze praktycznym	2,7					ECTS
9.	Sumaryczne obciążenie pracą studenta	100					h
10.	Punkty ECTS za moduł <i>1 punkt ECTS=25 godzin obciążenia studenta</i>	4					ECTS

LITERATURA

1. Cichosz P., *Systemy uczące się*, Wydawnictwa Naukowo-Techniczne, Warszawa 2000.
2. Frątczak E. (red.), *Zaawansowane metody analiz statystycznych*, Oficyna Wydawnicza Szkoła Główna Handlowa w Warszawie, Warszawa 2013.
3. Larose D.T., *Odkrywanie wiedzy z danych*. Wydawnictwo Naukowe PWN, Warszawa 2006.
4. Larose D.T., *Metody i modele eksploracji danych*, Wydawnictwo Naukowe PWN, Warszawa 2012.
5. Hand D., Manila H., Smyth P., *Eksploracja danych*, Wydawnictwa Naukowo-Techniczne, Warszawa 2005.
6. Olhost F.J., *Big Data Analytics. Turning Big Data into Big Money*, John Wiley & Sons, Inc., Hoboken, New Jersey 2013.
7. Panek T., *Statystyczne metody wielowymiarowej analizy porównawczej*, Szkoła Główna Handlowa w Warszawie, Warszawa 2009.
8. Szeliga M., *Data Science i uczenie maszynowe*, Wydawnictwo Naukowe PWN SA, Warszawa 2017.