

Program „Metropolitalny Fundusz Wspierania Nauki”

Projekt w ramach Obszaru 3 - NOWATORSKIE METODY KSZTAŁCENIA

Zadanie nr 2: Innowacje w Nauczaniu – AI i Data Science

Podzadanie: 2.1. Opracowanie modułu: *Data Science*

(Paweł Błaszczuk, Agnieszka Kulawik)

Spis treści

1. Opis modułu *Data Science*
2. Lista przykładowych tematów projektów zaliczeniowych z modułu *Data Science*.
3. Wzór raportu z realizacji projektu.

1. Opis modułu

OPRACOWALI

PAWEŁ BŁASZCZYK, AGNIESZKA KULAWIK

nazwa kierunku	matematyka
poziom kształcenia	Drugi
profil kształcenia	ogólnoakademicki
forma prowadzenia studiów	stacjonarne

MODUŁ KSZTAŁCENIA: *Data Science*

1. Liczba punktów ECTS: 6

2. Zakładane efekty kształcenia modułu		
kod efektu kształcenia modułu	opis efektu kształcenia	kod efektu kształcenia kierunku
DATA_1	Ma pogłębioną wiedzę dotyczącą metod analizy danych i modelowania statystycznego, w tym regresji, klasyfikacji i klasteryzacji, oraz zna ich zastosowania w różnych dziedzinach nauki i gospodarki.	K_W04
DATA_2	Zna narzędzia informatyczne i środowiska programistyczne wykorzystywane w Data Science (np. Python, R, Statistica) oraz rozumie proces przetwarzania i eksploracji danych w kontekście przygotowania danych do analizy.	K_W05
DATA_3	Potrafi samodzielnie przeprowadzić pełny proces analizy danych: od ich pozyskania i wstępnego przetwarzania, przez analizę opisową, testowanie hipotez i modelowanie, aż po interpretację wyników i ich wizualizację.	K_U05
DATA_4	Potrafi wykorzystać odpowiednie metody statystyczne i algorytmy uczenia maszynowego do rozwiązywania złożonych problemów analitycznych oraz dokonać krytycznej oceny jakości i trafności uzyskanych modeli.	K_U07
DATA_5	Ma świadomość odpowiedzialności za rzetelność analiz i interpretacji danych, potrafi jasno komunikować wnioski, uwzględniając aspekty etyczne i kontekst społeczny analizy danych.	K_K04

3. Opis modułu	
Opis	Wykład (30 h = 15 x 2 h)

Wykład 1. Wprowadzenie do analizy danych i Data Science

- Czym jest Data Science? Obszary zastosowań, przykłady.
- Podstawowe pojęcia: dane, zmienne, obserwacje, modele
- Proces analizy danych: od zbierania do wnioskowania
- Rola matematyka w pracy z danymi i projektach Data Science
- Narzędzia i środowiska pracy (np. Python, R, Statistica)

Wykład 2. Przygotowanie i eksploracja danych

- Formaty i źródła danych
- Przygotowanie danych do analizy. Wstępne przetwarzanie danych.
- Podstawowe transformacje danych (filtrowanie, grupowanie, sortowanie)

Wykład 3-4. Opisowa analiza danych

- Statystyki opisowe (średnia, mediana, odchylenie standardowe)
- Rozkłady danych: histogramy, boxploty, wykresy słupkowe
- Zależności między zmiennymi: korelacja, tabele kontyngencji

Wykład 5. Wizualizacja danych

- Podstawy wizualizacji.
- Tworzenie wykresów
- Interpretacja wykresów

Wykład 6-9. Testowanie hipotez statystycznych

- Testy parametryczne (próby zależne i niezależne)
- Testy nieparametryczne (próby zależne i niezależne)
- Test niezależności chi kwadrat
- Testy normalności

Wykład 10-14. Modelowania danych

- Idea modelu – co to jest model i dlaczego go stosujemy?
- Modele regresyjne (np. regresja liniowa, logistyczna)
- Klasyfikacja (np. drzewa decyzyjne, kNN, SVM)
- Klasteryzacja (np. K-means, hierarchiczna)
- Walidacja modeli (np. cross-validation, metryki jakości)
- Redukcja wymiarowości (np. PCA, t-SNE)

Wykład 15. Praktyczne zastosowania

- Studium przypadków z różnych dziedzin (np. finanse, zdrowie, przemysł)

Laboratorium (30h = 15 x 2h)

Lab 1. Wprowadzenie do analizy danych i Data Science

- Narzędzia i środowiska pracy (np. Python, R, Statistica)

Lab 2. Przygotowanie i eksploracja danych

- Formaty i źródła danych
- Przygotowanie danych do analizy. Wstępne przetwarzanie danych.
- Podstawowe transformacje danych (filtrowanie, grupowanie, sortowanie)
- Studium przypadków. Praca nad projektem.

Lab 3-4. Opisowa analiza danych

- Statystyki opisowe (średnia, mediana, odchylenie standardowe)
- Rozkłady danych: histogramy, boxploty, wykresy słupkowe
- Zależności między zmiennymi: korelacja, tabele kontyngencji
- Studium przypadków. Praca nad projektem.

Lab 5. Wizualizacja danych

- Podstawy wizualizacji.
- Tworzenie wykresów
- Interpretacja wykresów
- Studium przypadków. Praca nad projektem.

Lab 6-9. Testowanie hipotez statystycznych

- Testy parametryczne (próby zależne i niezależne)
- Testy nieparametryczne (próby zależne i niezależne)
- Test niezależności chi kwadrat
- Testy normalności
- Studium przypadków. Praca nad projektem.

Lab 10-14. Modelowania danych

- Idea modelu – co to jest model i dlaczego go stosujemy?
- Modele regresyjne (np. regresja liniowa, logistyczna)
- Klasyfikacja (np. drzewa decyzyjne, kNN, SVM)
- Klasteryzacja (np. K-means, hierarchiczna)
- Walidacja modeli (np. cross-validation, metryki jakości)

	• Redukcja wymiarowości (np. PCA, t-SNE) • Studium przypadków. Praca nad projektem. Lab 15. Praca końcowa nad projektem analizy danych
--	---

4. Sposoby weryfikacji efektów kształcenia modułu

nazwa (typ) sposobu weryfikacji	opis szczegółowy	efekt(-y) kształcenia modułu
projekt	Przygotowanie projektu indywidualnego lub zespołowego, który pozwala zastosować pełny proces analizy danych – od eksploracji, przez modelowanie, aż po wnioskowanie i prezentację wyników. Efektem końcowym jest raport, który dokumentuje przebieg analizy i kluczowe wnioski.	DATA_1, DATA_2, DATA_3, DATA_4, DATA_5
egzamin	Weryfikacja znajomości podstawowych pojęć, metod, narzędzi i technik analizy danych oraz umiejętność ich właściwego zastosowania. Oceniana jest zarówno wiedza teoretyczna jak i rozumienie procesów analitycznych.	DATA_1, DATA_2

5. Formy prowadzenia zajęć (studia prowadzone w formie stacjonarnej)

forma prowadzenia zajęć		praca własna studenta		punkty ECTS
opis (z uwzględnieniem metod dydaktycznych)	liczba godzin	opis	liczba godzin	Liczba punktów

wykład	wykład, z wykorzystaniem pomocy audiowizualnych, prezentujący pojęcia i fakty z zakresu treści programowych wymienionych w opisie modułu i ilustrujący je licznymi przykładami	30	Przygotowanie do zajęć. Konsultowanie programu i organizacji zajęć. Przygotowanie do weryfikacji efektów uczenia się. Konsultowanie wyników weryfikacji efektów uczenia się.	60	3
laboratorium	Laboratoria służą praktycznemu zastosowaniu zagadnień omawianych na wykładach. Zajęcia odbywają się w pracowni komputerowej i polegają na samodzielnym rozwiązywaniu zadań z wykorzystaniem narzędzi i języków programowania stosowanych w analizie danych. W ramach zajęć stosowane są metody dydaktyczne takie jak ćwiczenia praktyczne, rozwiązywanie problemów oraz konsultacje z prowadzącym.	30	Przygotowanie do zajęć. Konsultowanie programu i organizacji zajęć. Przygotowanie do weryfikacji efektów uczenia się. Konsultowanie wyników weryfikacji efektów uczenia się.	60	3
		60	suma godzin:	120	6

6. Wymagania wstępne: Brak

7. Zalecana literatura	
1.	Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani "Introduction to Statistical Learning with application in R", Springer 2021
2.	Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani "Introduction to Statistical Learning with application in Python", Springer 2023
3.	Saltz, Stanton, "An Introduction to Data Science", Thousand Oaks: SAGE Publications 2017
4.	Peter Bruce, Andrew Bruce, Peter Gedeck, "Practical Statistics for Data Scientists", O'Reilly 2020
5.	Christopher M. Bishop, "Pattern Recognition and Machine Learning", Springer

6.	R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification 2ed., Wiley, New York 2000.
7.	Hastie, Tibshirani, Friedman, "The Elements of Statistical Learning", Springer 2009
8.	V. N. Vapnik, The Nature of Statistical Learning Theory 2ed., Springer, 2000.
9.	A. Stanis, Przystępny kurs statystyki z zastosowaniem STATISTICA PL na przykładach z medycyny, StatSoft, Kraków 2006 (3 tomy).
10.	M. Sobczyk "Statystyka", PWN Warszawa 2007.
11.	Internetowy podręcznik statystyki, StatSoft, http://www.statsoft.pl/textbook/

2. Przykładowe tematy projektów zaliczeniowych

OPRACOWALI

PAWEŁ BŁASZCZYK, AGNIESZKA KULAWIK

Przykładowe tematy projektów zaliczeniowych

1. Wpływ czynników zdrowotnych i stylu życia na ryzyko chorób serca.
2. Analiza danych edukacyjnych: czynniki wpływające na wyniki uczniów.
3. Analiza opinii klientów.
4. Wykrywanie nieprawidłowości w danych finansowych.
5. Przewidywanie cen mieszkań na podstawie ich cech.
6. Analiza jakości powietrza w polskich miastach – eksploracja i wizualizacja danych.
7. Analiza danych demograficznych i edukacyjnych w Polsce.
8. Analiza danych o wypadkach drogowych w Polsce.

3. Wzór raportu

OPRACOWALI

PAWEŁ BŁASZCZYK, AGNIESZKA KULAWIK

Spis treści

Nazwa i cel badania

Skład zespołu badawczego

Wstęp

Urządzenia i oprogramowanie wykorzystane do badania

Opis analizowanego zbioru

Hipotezy badawcze i testy statystyczne

Analiza współzależności zjawisk

Modelowanie danych

Wizualizacja wyników

Wnioski i podsumowanie wyników analiz

Załączniki

Słowniczek

Nazwa i cel badania

Temat badania

Sformułowano następujące cele badawcze

cel nr 1. Zbadanie zależności między płcią a poziomem dochodów

cel nr 2.

cel nr 3. Ocena współzależności między

Skład zespołu badawczego

(Imiona i nazwiska studentów, numery indeksów, podział ról – opcjonalnie)

Wstęp

W tym etapie przeprowadzone zostały badania

Opisać krótko: cel analizy (przedstawienie celu projektu), zakres analizy, źródło danych (skąd pochodzi zbiór np. dane fikcyjne, open data, dane z uczelni), krótki opis zbioru (Co zawiera zbiór danych, ile jest obserwacji, ile zmiennych i jakiego typu.)

Urządzenia i oprogramowanie wykorzystane do badania

W trakcie prowadzonych prac wykorzystano następujące urządzenia:

1.
2.

Oprogramowanie wykorzystane w przeprowadzonych pracach B+R:

1.
2.

Opis analizowanego zbioru

Opis zbioru danych w tym ilość zmiennych, obserwacji, opis zmiennych (nazwa, typ, jednostka, opis,) najlepiej w formie tabeli. Należy również sprawdzić czy w zbiorze są braki w danych i wartości odstające. Powinna się tu również znaleźć tabela z opisem zmiennych.

Analiza eksploracyjna

Tu opisujemy podstawowe statystyki opisowe dla kilku wybranych zmiennych zarówno mierzalnych jak i niemierzalnych w tym m.in.

- Średnie, mediany, odchylenia standardowe, wariancja, skośność, kurtoza, kwartyle dla zmiennych ilościowych,
- Tabele częstości i wykresy słupkowe dla zmiennych jakościowych
- Wykresy: histogramy, ramka-wąsy, wykresy macierzowe, normalności
- Opis zauważalnych rozkładów, odstających wartości, zależności itp.

Hipotezy badawcze i testy statystyczne

Tu formułujemy hipotezy badawcze i przeprowadzamy testy statystyczne np.

Hipoteza 1 (np. test t-Studenta):

H0: Średni dochód mężczyzn i kobiet jest taki sam.

H1: Średni dochód mężczyzn i kobiet się różni.

Hipoteza 2 (np. test chi-kwadrat):

H0: Płeć i poziom wykształcenia są niezależne.

H1: Płeć i poziom wykształcenia są zależne.

Hipoteza 3 (np. korelacja Pearsona):

H0: Nie ma związku między wiekiem a dochodem.

H1: Istnieje związek między wiekiem a dochodem.

Analiza współzależności zjawisk

W tej części analizujemy siłę i kierunek związku między dwoma wybranymi zmiennymi przy pomocy odpowiednich współczynników korelacji (Pearsona, Spearmana, Yula).

Modelowanie danych

Tu opisujemy przygotowane modele w tym te dotyczące regresji, klasyfikacji i klasteryzacji. Opisujemy:

- podział danych: trening/test lub walidacja krzyżowa
- Parametryzacja modeli (jeśli dotyczy)

Dokonujemy redukcji wymiarowości. Opisujemy

- zastosowane podejście (selekcja, ekstrakcja)
- wybrane metody (np. PCA, K-means)
- Liczba komponentów/klastrów

Dokonujemy oceny jakości modelu:

- metryki (np. RMSE, accuracy, precision, recall, AUC...)
- macierze pomyłek (jeśli klasyfikacja)

Wizualizacja wyników

Tu przedstawiamy najważniejsze wykresy i diagramy końcowe. Należy opisać co wynika z wykresów.

Wnioski i podsumowanie wyników analiz

Tu powinno się znaleźć podsumowanie najważniejszych wyników analizy i ich interpretacja, odpowiedzi na pytania badawcze, możliwe ograniczenia badania (np. brak danych, mała próba)

W wyniku przeprowadzonych prac B+R określono następujące wnioski:

1. wniosek 1
2. wniosek 2

Załączniki

Tu umieszczamy listę załączników np. z kodami źródłowymi, dodatkowymi danymi

1. Załącznik nr 1 - Kod źródłowy w języku python
2. Załącznik nr 2 - Kod źródłowy w języku R
3. Załącznik nr 3 - Dodatkowy model danych
4. Załącznik nr 4 - Dodatkowa tabela z danymi

Słowniczek

ABC	statystyczne
DEF	Statystyczna
ZXC	Measurement