

Eksploracja danych

semestr zimowy 2026/2027

Dr Anna Muranova
UWM w Olsztynie

Wstęp

Co to jest „Eksploracja danych”?

Eksploracja danych (ang. *Data Mining*) – to proces używany do wydobywania cennych informacji z dużych zbiorów danych. Eksploracja danych jest jednym z etapów procesu odkrywania wiedzy z baz danych (ang. *Knowledge Discovery in Databases, KDD*), który składa się z następujących kroków:

1. Zrozumienie dziedziny problemu i celu analizy.
2. Przygotowanie danych
3. Eksploracja danych – stosowanie dedykowanych metod w celu wydobycia z danych wzorców, reguł, zależności.
4. Interpretacja znalezionych regularności – identyfikacja najbardziej interesujących (istotnych) odkrytych wzorców.
5. Prezentacja odkrytej wiedzy – przedstawienie odkrytej wiedzy użytkownikowi za pomocą technik wizualizacji i reprezentacji danych.
6. Wykorzystanie odkrytej wiedzy.

Przygotowanie danych:

1. Czyszczenie danych (ang. *data cleaning*) – usuwanie błędów wynikających z pomyłek operatora (ortograficzne, literówki), różnych form danych oznaczających te same informacje, niezgodności wartości pola i jego opisu, brakujących wartości, błędów wynikających z ograniczeń systemu (brak pól na niektóre ważne dane), wielokrotnego wprowadzenia tej samej danej.
2. Integracja danych – łączenie danych pochodzących z różnych źródeł (baz danych), posiadających różną strukturę oraz różne modele danych.
3. Wybieranie danych – z bazy danych pobierane są dane do przeprowadzenia analiz.

Knowledge Discovery Nuggets

Dziedzina eksploracji danych rozpoczęła się od seminarium prowadzonego przez Grigorija Piatetsky-Shapiro w 1989 roku. Wcześniej Grigorij Piatetsky-Shapiro, pracując w GTE Labs, zainteresował się pytaniem: czy możliwe jest automatyczne znalezienie określonych reguł w celu przyspieszenia niektórych zapytań do dużych baz danych. Jednocześnie zaproponowano dwa terminy – *data mining* (eksploracja danych) oraz *Knowledge Discovery in Databases, KDD* (odkrywanie wiedzy z danych).

W 1993 roku ukazał się pierwszy biuletyn „Knowledge Discovery Nuggets”, a w 1994 roku powstała jedna z pierwszych stron internetowych eksploracji danych.

<https://www.kdnuggets.com/gps.html>

Techniki i metody eksploracji danych

Techniki i metody służące eksploracji danych wywodzą się głównie z dziedziny badań nad sztuczną inteligencją. Główne przykłady stosowanych rozwiązań należą do następujących obszarów:

- ▶ wizualizacje na wykresach,
- ▶ metody statystyczne,
- ▶ sieci neuronowe,
- ▶ metody uczenia maszynowego,
- ▶ metody ewolucyjne,
- ▶ logika rozmyta,
- ▶ zbiory przybliżone.

Eksploracja danych a statystyka 1

Eksploracja danych	Statystyka
Eksploracja danych to proces wydobywania przydatnych informacji, wzorców i trendów z ogromnych zbiorów danych i wykorzystywania ich do podejmowania decyzji w oparciu o dane.	Statystyka odnosi się do analizy i prezentacji danych liczbowych i stanowi główną część wszystkich algorytmów eksploracji danych.
Dane wykorzystywane w eksploracji danych są numeryczne lub nienumeryczne.	Dane wykorzystywane w statystykach mają wyłącznie charakter liczbowy.

Eksploracja danych a statystyka 2

Eksploracja danych	Statystyka
Rodzaje eksploracji danych to grupowanie, klasyfikacja, asocjacja, sieć neuronowa, analiza oparta na sekwencjach, wizualizacja itp.	Rodzaje statystyki to statystyka opisowa i statystyka wnioskowania.
Nadaje się do dużych zbiorów danych.	Nadaje się do mniejszych zestawów danych.
Eksploracja danych jest procesem indukcyjnym. Oznacza to generowanie nowej teorii na podstawie danych.	Statystyka jest procesem dedukcyjnym. Nie pozwala sobie na jakiegokolwiek przewidywania.

Eksploracja danych a statystyka 3

Eksploracja danych	Statystyka
Czyszczenie danych jest częścią eksploracji danych.	W statystyce czyste dane służą do wdrożenia metody statystycznej.
Walidacja modelu wymaga mniej interakcji z użytkownikiem, dzięki czemu można go łatwo zautomatyzować.	Walidacja modelu wymaga interakcji użytkownika, dlatego jest to złożone i automatyczne.
Zastosowania do eksploracji danych obejmują analizę danych finansowych, przemysł detaliczny, przemysł telekomunikacyjny, analizę danych biologicznych, niektóre zastosowania naukowe itp.	Zastosowanie statystyki obejmuje biostatystykę, kontrolę jakości, demografię, badania operacyjne itp.

Wstępny plan kursu

- ▶ elementy algebry liniowej – powtórzenie
- ▶ zaawansowany rachunek prawdopodobieństwa
- ▶ zaawansowana statystyka
- ▶ modelowanie regresji
- ▶ redukcja wymiaru
- ▶ analiza skupień
- ▶ metoda asocjacyjna
- ▶ sieci neuronowe (?)