**Nazwa przedmiotu:**

Metody eksploracji danych w odkrywaniu wiedzy

**Koordynator przedmiotu:**

Rajmund Kożuszek

**Status przedmiotu:**

Fakultatywny dowolnego wyboru

**Poziom kształcenia:**

Studia II stopnia

**Program:**

Informatyka

**Grupa przedmiotów:**

Przedmioty techniczne - zaawansowane

**Kod przedmiotu:**

MED

**Semestr nominalny:**

2 / rok ak. 2021/2022

**Liczba punktów ECTS:**

4

**Liczba godzin pracy studenta związanych z osiągnięciem efektów uczenia się:**

1. liczba godzin kontaktowych – 60 godz., w tym
obecność na wykładach 30 godz.,
obecność na zajęciach projektowych 30 godz.,
2. praca własna studenta – 60 godz., w tym
wykonanie projektu 46 godz.
przygotowanie dokumentacji projektowej 5 godz.
przygotowanie do sprawdzianów 9 godz.
Łączny nakład pracy studenta wynosi 120 godz., co odpowiada 4 pkt. ECTS.

**Liczba punktów ECTS na zajęciach wymagających bezpośredniego udziału nauczycieli akademickich:**

2,0 pkt. ECTS, co odpowiada 60 godz. kontaktowym

**Język prowadzenia zajęć:**

polski

**Liczba punktów ECTS, którą student uzyskuje w ramach zajęć o charakterze praktycznym:**

2,7 pkt. ECTS, co odpowiada 30 godz. zajęć projektowych, 46 godz. wykonywania projektu i 5 godz. przygotowania dokumentacji projektowej

**Formy zajęć i ich wymiar w semestrze:**

|  |  |
| --- | --- |
| Wykład: | 30h |
| Ćwiczenia: | 0h |
| Laboratorium: | 0h |
| Projekt: | 30h |
| Lekcje komputerowe: | 0h |

**Wymagania wstępne:**

Wymagana jest umiejętność programowania w języku C, C++, C# lub Python.

**Limit liczby studentów:**

30

**Cel przedmiotu:**

Celem wykładu jest zapoznanie studentów z podstawami teoretycznymi oraz wydajnymi metodami i algorytmami eksploracji danych. Celem projektu jest nabycie przez studentów umiejętności ich wydajnego implementowania, usprawniania, planowania i prowadzenia badań o charakterze naukowym, właściwego wysnuwania wniosków i dokumentowania prowadzonych prac.

**Treści kształcenia:**

Wykład:
1. Eksploracja danych. Eksploracja danych a proces odkrywania wiedzy. Zadania eksploracji danych. Eksploracja danych jako dziedzina multidyscyplinarna. Aktualne wyzwania w eksploracji danych.
2. Częste wzorce i reguły asocjacyjne. Skalowalne metody odkrywania częstych wzorców i reguł asocjacji w transakcyjnych i relacyjnych bazach danych, z uwzględnieniem hierarchii pojęć i negacji. Określanie ograniczeń w języku eksploracji danych. Wykorzystywanie nałożonych ograniczeń w celu zwiększenia wydajności procesu odkrywania wzorców. Odkrywanie wzorców wysokiej użyteczności.
3. Miary oceny reguł asocjacyjnych. Właściwości miar oceny reguł asocjacyjnych, takich jak współczynnik podniesienia, współczynnik pewności, współczynnik zależności, iloraz szans i współczynnik wzrostu.
4. Zwięzłe reprezentacje częstych wzorców. Generatory, zbiory zamknięte, uogólnione zbiory wolne od dysjunkcji i minimalne dysjunkcyjne jako podstawowe elementy bezstratnych reprezentacji częstych wzorów. Metody odkrywania zwięzłych reprezentacji częstych wzorców. Zastosowanie reprezentacji do wyprowadzania wszystkich częstych wzorców.
5. Zwięzłe reprezentacje reguł asocjacyjnych. Generatory i zbiory zamknięte jako elementy składowe bezstratnych reprezentacji reguł asocjacyjnych, takich jak reguły reprezentatywne, minimalne nieredundantne reguły i szablony reguł. Mechanizmy wyprowadzania reguł asocjacyjnych z tych reprezentacji.
6. Wzorce i reguły sekwencyjne. Odkrywanie częstych wzorców i reguł sekwencyjnych. Reprezentacje częstych wzorców i reguł sekwencyjnych. Odkrywanie wzorców i reguł sekwencyjnych wysokiej użyteczności
7. Klasyfikacja. Wykorzystanie wzorów kontrastowych w klasyfikacji.
8. Predykcja. Wykorzystanie impulsowych sieci neuronowych do prognozowania.
9. Miary podobieństwa i odległości obiektów. Skuteczne metody odkrywania obiektów najbardziej podobnych (lub najbliższych) ze względu na miary, takie jak odległość Minkowskiego i miary podobieństwa: Jaccard, Tanimoto i kosinus.
10. Grupowanie. Metody gęstościowego grupowania obiektów i odkrywania anomalii, takie jak DBSCAN i NBC, oraz ich wydajne modyfikacje, takie jak TI-DBSCAN i TI-NBC, wykorzystujące własność nierówności trójkąta wprost lub przy użyciu drzewa VP-tree.
11. Zależności funkcyjne i przybliżone. Metody wykrywania zależności funkcyjnych i przybliżonych pomiędzy zbiorami atrybutów w bazach danych.
12. Zbiory przybliżone. Metody odkrywania reduktów zbioru atrybutów warunkowych i reguł decyzyjnych.
13. Wnioskowanie w przypadku niekompletnej informacji. Uprawnione podejścia do wnioskowania na podstawie danych z brakującymi wartościami oraz na podstawie częściowej wiedzy.
Projekt:
Projekt obejmuje:
- dogłębne zapoznanie się z algorytmami eksploracji danych, które prowadzący projekt zlecił do realizacji,
- wymóg konsultacji interpretacji tematu i zakresu projektu,
- konsultowanie z prowadzącym planowanego rozwiązania zadania projektowego,
- wydajną implementację zadanych algorytmów eksploracji danych,
- przeprowadzenie eksperymentów umożliwiających ocenę jakości i wydajności zrealizowanych algorytmów,
- przeprowadzenie tej oceny i wysnucie wniosków na podstawie przeprowadzonych eksperymentów,
- przygotowanie dokumentacji projektu,
- prezentację uzyskanych wyników.

**Metody oceny:**

(-)

**Egzamin:**

nie

**Literatura:**

1. Morzy T., Eksploracja danych, Metody i algorytmy, Wydawnictwo Naukowe PWN, 2013
2. Han J., Kamber M., Pei, J., Data Mining: Concepts and Techniques, The Morgan Kaufmann Series in Data Management Systems, 3rd edition, Morgan Kaufmann, 2011
3. Kryszkiewicz M., Concise Representations of Frequent Patterns and Association Rules, Prace Naukowe, Elektronika, Oficyna Wydawnicza Politechniki Warszawskiej, z. 142 (2002)
oraz dostępne na stronach Internetowych publikacje z dziedziny eksploracji danych, o których wykładowca będzie informował w materiałach wykładowych lub na bieżąco w trakcie zajęć.

**Witryna www przedmiotu:**

https://usosweb.usos.pw.edu.pl/kontroler.php?\_action=katalog2/przedmioty/pokazPrzedmiot&prz\_kod=103C-INxxx-MSP-MED

**Uwagi:**

(-)

## Charakterystyki przedmiotowe