# KARTA OPISU PRZEDMIOTU - SYLABUS

Nazwa przedmiotu
Eksploracja danych [S1SI1E>EDAN]

## Przedmiot

Kierunek studiów
Sztuczna inteligencja/Artificial Intelligence

Rok/Semestr
2/4

Studia w zakresie (specjalność)
–

Profil studiów
ogólnoakademicki

Poziom studiów
pierwszego stopnia

Język oferowanego przedmiotu
angielski

Forma studiów
stacjonarne

Wymagalność
obligatoryjny

## Liczba godzin

Wykład
30

Laboratorium
30

Inne
0

Ćwiczenia
0

Projekty/seminaria
0

## Liczba punktów ECTS

5,00

## Koordynatorzy

dr hab. inż. Mikołaj Morzy prof. PP
mikolaj.morzy@put.poznan.pl

## Wykładowcy

## Wymagania wstępne

Student rozpoczynający ten przedmiot powinien posiadać podstawową wiedzę z zakresu systemów baz danych, statystyki, probabilistyki, oraz optymalizacji kombinatorycznej. Do realizacji zajęć laboratoryjnych konieczna jest podstawowa znajomość języków programowania Java oraz Python. Student powinien posiadać umiejętność rozwiązywania podstawowych problemów z zakresu przetwarzania i analizy danych oraz umiejętność pozyskiwania informacji ze wskazanych źródeł. Powinien również rozumieć konieczność poszerzania swoich kompetencji / mieć gotowość do podjęcia współpracy w ramach zespołu. W zakresie kompetencji społecznych student musi prezentować takie postawy jak uczciwość, odpowiedzialność, wytrwałość, ciekawość poznawcza, kreatywność, kultura osobista, szacunek dla innych ludzi.

## Cel przedmiotu

1. Przekazanie studentom podstawowej wiedzy z eksploracji danych, w zakresie: - typów danych, miar podobieństwa i odległości danych - metod odkrywania asocjacji, - odkrywania wzorców sekwencji, - grupowania danych. 2. Rozwijanie u studentów umiejętności rozwiązywania problemów eksploracji danych i odkrywania wiedzy z dużych repozytoriów danych. 3. Kształtowanie u studentów umiejętności pracy zespołowej oraz integracji wiedzy z różnych obszarów informatyki. 4. Rozwijanie u studentów umiejętności formułowania i testowania hipotez związanych z problemami inżynierskimi i prostymi problemami badawczymi w zakresie analizy i eksploracji danych.

## Przedmiotowe efekty uczenia się

Wiedza:
ma zaawansowaną i pogłębioną wiedzę z zakresu systemów informatycznych bazujących na uczeniu maszynowym, podstaw teoretycznych ich budowania oraz metod, narzędzi i środowisk programistycznych wykorzystywanych do ich implementacji (K2st_W1)
ma uporządkowaną i podbudowaną teoretycznie wiedzę ogólną związaną z kluczowymi zagadnieniami z zakresu statystyki i informatyki (K2st_W2)
ma zaawansowaną wiedzę szczegółową dotyczącą eksploracji danych, uczenia maszynowego, statystyki i przetwarzania danych (K2st_W3)
ma wiedzę o trendach rozwojowych i najistotniejszych nowych osiągnięciach uczenia maszynowego i eksploracji danych (K2st_W4)
zna zaawansowane metody, techniki i narzędzia stosowane przy rozwiązywaniu złożonych zadań inżynierskich i prowadzeniu prac badawczych w obszarze eksploracji danych (K2st_W6)

Umiejętności:
potrafi pozyskiwać informacje z literatury, baz danych oraz innych źródeł (w języku polskim i angielskim), integrować je, dokonywać ich interpretacji i krytycznej oceny, wyciągać wnioski oraz formułować i wyczerpująco uzasadniać opinie (K2st_U1)
potrafi planować i przeprowadzać eksperymenty, oraz interpretować uzyskane wyniki i wyciągać wnioski oraz formułować i weryfikować hipotezy związane ze złożonymi problemami osobowymi i technicznymi (K2st_U3)
potrafi wykorzystać do formułowania i rozwiązywania zadań inżynierskich i prostych problemów badawczych metody analityczne, symulacyjne oraz eksperymentalne (K2st_U4)
potrafi - przy formułowaniu i rozwiązywaniu zadań inżynierskich - integrować wiedzę z różnych obszarów informatyki i statystyki oraz zastosować podejście systemowe, uwzględniające także aspekty pozatechniczne (K2st_U5)
potrafi ocenić przydatność i możliwość wykorzystania nowych bibliotek do uczenia maszynowego (K2st_U6)
potrafi dokonać krytycznej analizy istniejących procesów uczenia maszynowego oraz zaproponować ich ulepszenia (K2st_U8)
potrafi - stosując m.in. metody uczenia maszynowego - rozwiązywać złożone zadania informatyczne, w tym zadania nietypowe oraz zadania zawierające komponent badawczy (K2st_U10)

Kompetencje społeczne:
rozumie, że w uczeniu maszynowym wiedza, umiejętności i narzędzia bardzo szybko stają się przestarzałe (K2st_K1)
rozumie znaczenie wykorzystywania najnowszej wiedzy z zakresu uczenia maszynowego w rozwiązywaniu problemów badawczych i praktycznych (K2st_K2)

## Metody weryfikacji efektów uczenia się i kryteria oceny

Efekty uczenia się przedstawione wyżej weryfikowane są w następujący sposób:

Ocena formująca:
a) w zakresie wykładów:
- na podstawie ocen realizowanych ćwiczeń/zadań przy tablicy
b) w zakresie laboratoriów:
- na podstawie oceny bieżącego postępu realizacji zadań,

Ocena podsumowująca:
a) w zakresie wykładów weryfikowanie założonych efektów kształcenia realizowane jest przez:

- ocenę wiedzy i umiejętności wykazanych na otwartym egzaminie pisemnym o charakterze problemowym (student może korzystać z dowolnych materiałów dydaktycznych), Egzamin składa się z 6-8 zadań problemowych, za które można uzyskać 10 pkt. Łącznie można uzyskać od 60-80 pkt. Zaliczenie na ocenę 3.0 wymaga uzyskania 50% maksymalnej liczby punktów.
- dodatkowe punkty za obecność na wykładach
- dodatkowe 20% punktów uzyskanych na zaliczenie laboratorium
- omówienie wyników egzaminu,

b) w zakresie laboratoriów weryfikowanie założonych efektów kształcenia realizowane jest przez:
- ocenę stopnia przyswojenia wiedzy prezentowanej w trakcie laboratorium poprzez krótki quiz zawierający pytania dotyczące zagadnień poruszanych w trakcie danego tygodnia zajęć
- realizację indywidualnych zadań samodzielnych o charakterze projektowym lub problemowym po każdych zajęciach,
- realizację większego zadania o charakterze projektowym lub problemowym.
Uzyskiwanie punktów dodatkowych za aktywność podczas zajęć, a szczególnie za:
- poprawne rozwiązywanie zagadek tematycznie związanych ze statystyką, uczeniem maszynowym i eksploracją danych,
- udział w międzynarodowych konkursach programistycznych, ze szczególnym naciskiem na pracę zespołową.

## Treści programowe

Program wykładu obejmuje następujące zagadnienia:
Wprowadzenie do eksploracji danych: metody i zastosowania. Odkrywanie asocjacji: sformułowanie problemu i definicja reguł asocjacyjnych. Tablica obserwacji. Odkrywanie asocjacji binarnych: reguła asocjacyjna, miary oceny reguł. Algorytm odkrywania binarnych reguł asocjacyjnych Apriori. Algorytm odkrywania binarnych reguł asocjacyjnych FP-Growth. Domknięte i maksymalne reguły asocjacyjne. Odkrywanie wielopoziomowych reguł asocjacyjnych. Odkrywanie wielowymiarowych reguł asocjacyjnych. Binaryzacja i dyskretyzacja danych. Miary atrakcyjności reguł asocjacyjnych. Omówienie typów danych. Miary podobieństwa i miary odległości dla różnych typów danych danych, przekleństwo wymiarowości.
Typy danych sekwencyjnych. Odkrywanie wzorców sekwencji: sformułowanie problemu. Podstawowy algorytm odkrywania wzorców sekwencji. Prefiksowy algorytm odkrywania wzorców sekwencji. Odkrywanie wzorców sekwencji z ograniczeniami czasowymi - sformułowanie problemu. Algorytm odkrywania wzorców sekwencji z ograniczeniami czasowym. Proces grupowania i składowe procesu grupowania. Klasyfikacja metod grupowania. Grupowanie hierarchiczne: aglomeracyjne i podziałowe. Algorytmy grupowania hierarchicznego. Grupowanie iteracyjno-optymalizacyjne. Metody grupowania gęstościowego. Metody oparte ma modelu. Grupowanie obiektów opisanych atrybutami kategorycznymi. Wykrywanie punktów osobliwych.
Zajęcia laboratoryjne prowadzone są w formie piętnastu 2-godzinnych spotkań odbywających się w laboratorium. Program laboratorium obejmuje następujące zagadnienia:
Wstępne przygotowanie danych do procesów eksploracji danych: dyskretyzacja, normalizacja, zastępowanie wartości brakujących, wyznaczanie i eliminacja wartości odstających na przykładach środowisk RapidMiner, Orange Data Mining i KNIME. Wstępne przetwarzanie atrybutów z poziomu języka Python. Ocena ważności atrybutów, metody ważenia atrybutów, test chi-kwadrat, zasada minimalizacji długości opisu (MDL), ważenie atrybutów za pomocą entropii. Odkrywanie reguł asocjacyjnych i algorytmy Apriori oraz FP-Growth. Podstawowe algorytmy grupowania, praktyczne ograniczenia algorytmów k-średnich i k-medoidów, algorytmy grupowania bazujące na gęstości, rodzina algorytmów EM grupownia. Niskopoziomowy interfejs programistyczny do uczenia maszynowego w języku Python Sci-Kit. Metody ekstrakcji cech: rodzina algorytmów PCA, SVD i NNMF.

## Tematyka zajęć

Program wykładu obejmuje następujące zagadnienia:
Wprowadzenie do eksploracji danych: metody i zastosowania. Odkrywanie asocjacji: sformułowanie problemu i definicja reguł asocjacyjnych. Tablica obserwacji. Odkrywanie asocjacji binarnych: reguła asocjacyjna, miary oceny reguł. Algorytm odkrywania binarnych reguł asocjacyjnych Apriori. Algorytm odkrywania binarnych reguł asocjacyjnych FP-Growth. Domknięte i maksymalne reguły asocjacyjne. Odkrywanie wielopoziomowych reguł asocjacyjnych. Odkrywanie wielowymiarowych reguł asocjacyjnych. Binaryzacja i dyskretyzacja danych. Miary atrakcyjności reguł asocjacyjnych. Omówienie

typów danych. Miary podobieństwa i miary odległości dla różnych typów danych danych, przekleństwo wymiarowości.
Typy danych sekwencyjnych. Odkrywanie wzorców sekwencji: sformułowanie problemu. Podstawowy algorytm odkrywania wzorców sekwencji. Prefiksowy algorytm odkrywania wzorców sekwencji. Odkrywanie wzorców sekwencji z ograniczeniami czasowymi - sformułowanie problemu. Algorytm odkrywania wzorców sekwencji z ograniczeniami czasowym. Proces grupowania i składowe procesu grupowania. Klasyfikacja metod grupowania. Grupowanie hierarchiczne: aglomeracyjne i podziałowe. Algorytmy grupowania hierarchicznego. Grupowanie iteracyjno-optymalizacyjne. Metody grupowania gęstościowego. Metody oparte ma modelu. Grupowanie obiektów opisanych atrybutami kategorycznymi. Wykrywanie punktów osobliwych.
Zajęcia laboratoryjne prowadzone są w formie piętnastu 2-godzinnych spotkań odbywających się w laboratorium. Program laboratorium obejmuje następujące zagadnienia:
Wstępne przygotowanie danych do procesów eksploracji danych: dyskretyzacja, normalizacja, zastępowanie wartości brakujących, wyznaczanie i eliminacja wartości odstających na przykładach środowisk RapidMiner, Orange Data Mining i KNIME. Wstępne przetwarzanie atrybutów z poziomu języka Python. Ocena ważności atrybutów, metody ważenia atrybutów, test chi-kwadrat, zasada minimalizacji długości opisu (MDL), ważenie atrybutów za pomocą entropii. Odkrywanie reguł asocjacyjnych i algorytmy Apriori oraz FP-Growth. Podstawowe algorytmy grupowania, praktyczne ograniczenia algorytmów k-średnich i k-medoidów, algorytmy grupowania bazujące na gęstości, rodzina algorytmów EM grupownia. Niskopoziomowy interfejs programistyczny do uczenia maszynowego w języku Python Sci-Kit. Metody ekstrakcji cech: rodzina algorytmów PCA, SVD i NNMF.

## Metody dydaktyczne

Wykład: prezentacja multimedialna, ilustrowana przykładami podawanymi na tablicy oraz przy użyciu skryptów Python
Laboratorium: praca samodzielna na podstawie przykładów dostarczonych przez prowadzącego, tutoriale, quizy, zadania do realizacji samodzielnej, samodzielna praca w grupach projektowych.

## Literatura

Podstawowa:
1. Eksploracja danych: metody i algorytmy, T. Morzy, PWN, 2013.
2. Introduction to Data Mining, Tan, P-N., Steinbach, M., Kumar, V., Pearson Education, 2006.
3. Data Mining: Concepts and Techniques, Han, J., Kamber, M., Pei, J., Morgan Kaufmann, 2012.
4. Systemy uczące się, Cichosz, P., WNT, 2000.
5. Data Mining: Practical Machine Learning Tools and Techniques, Witten, I., Frank, E., Morgan Kaufmann, 2005.

Uzupełniająca:
1. Statystyczne systemy uczące się, Koronacki, J., Ćwik, J., WNT, 2005.
2. Uczenie maszynowe i sieci neuronowe, Krawiec, K., Stefanowski, J., Wydawnictwo PP, 2003.
3. Programmer’s Guide to Data Mining, Zacharski, R. http://guidetodatamining.com/
4. Machine Learning, Ng, A., https://www.coursera.org/course/ml

## Bilans nakładu pracy przeciętnego studenta

| | Godzin | ECTS |
|---|---|---|
| Łączny nakład pracy | 125 | 5,00 |
| Zajęcia wymagające bezpośredniego kontaktu z nauczycielem | 62 | 2,50 |
| Praca własna studenta (studia literaturowe, przygotowanie do zajęć laboratoryjnych/ćwiczeń, przygotowanie do kolokwiów/egzaminu, wykonanie projektu) | 63 | 2,50 |