KARTA OPISU MODUŁU KSZTAŁCENIA

Nazwa modułu/przedmiotu	Kod
Eksploracja danych	**1010512321010510542**

Kierunek studiów	Profil kształcenia (ogólnoakademicki, praktyczny)	Rok / Semestr
Informatyka	**ogólnoakademicki**	**1 / 2**

Ścieżka obieralności/specjalność	Przedmiot oferowany w języku:	Kurs (obligatoryjny/obieralny)
Technologie przetwarzania danych	**polski**	**obligatoryjny**

Stopień studiów:	**Forma studiów** (stacjonarna/niestacjonarna)
II stopień	**niestacjonarna**

Godziny	Liczba punktów
Wykłady: **16** Ćwiczenia: **8** Laboratoria: **6** Projekty/seminaria: **6**	**5**

Status przedmiotu w programie studiów (podstawowy, kierunkowy, inny)	(ogólnouczelniany, z innego kierunku)
kierunkowy	**z danego kierunku**

Obszar(y) kształcenia i dziedzina(y) nauki i sztuki	**Podział ECTS (liczba i %)**
nauki techniczne	**5 100%**
nauki techniczne	**5 100%**

Odpowiedzialny za przedmiot / wykładowca:

prof. dr hab. inż. Tadeusz Morzy
email: Tadeusz.Morzy@put.poznan.pl
tel. +48 61 665 2906
Informatyki
Piotrowo 2, 60-965 Poznań

Odpowiedzialny za przedmiot / wykładowca:

dr hab. inż. Mikołaj Morzy, prof. nadzw.
email: Mikolaj.Morzy@put.poznan.pl
tel. +48 61 665 2961
Informatyki
Piotrowo 2, 60-965 Poznań

Wymagania wstępne w zakresie wiedzy, umiejętności, kompetencji społecznych:

1	**Wiedza:**	Efekty kształcenia ze studiów I: K1st_W1-8, K1st_U2-14, weryfikowane w procesie rekrutacji na studia 2 stopnia ? efekty te prezentowane są w serwisie internetowym wydziału www.fc.put.poznan.pl. Student rozpoczynający ten przedmiot powinien posiadać podstawową wiedzę z zakresu systemów baz danych, statystyki, probabilistyki, oraz optymalizacji kombinatorycznej.
2	**Umiejętności:**	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.
3	**Kompetencje społeczne**	W zakresie kompetencji społecznych student musi prezentować takie postawy jak uczciwość, odpowiedzialność, wytrwałość, ciekawość poznawcza, kreatywność, kultura osobista, szacunek dla innych ludzi.

Cel przedmiotu:

Cel przedmiotu:

1. Przekazanie studentom podstawowej wiedzy z eksploracji danych, w zakresie:

- metod odkrywania asocjacji,

- odkrywania wzorców sekwencji,

- klasyfikacji danych,

- 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.

Efekty kształcenia i odniesienie do kierunkowych efektów kształcenia

Wiedza:

1. zna zaawansowane metody, techniki i narzędzia stosowane przy rozwiązywaniu złożonych zadań inżynierskich i prowadzeniu prac badawczych w obszarze eksploracji danych - [K2st_W6]

Umiejętności:

1. 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]

2. potrafi wykorzystać do formułowania i rozwiązywania zadań inżynierskich i prostych problemów badawczych metody analityczne, symulacyjne oraz eksperymentalne - [K2st_U4]

3. 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]

4. potrafi ocenić przydatność i możliwość wykorzystania nowych bibliotek do uczenia maszynowego - [K2st_U6]

5. potrafi - stosując m.in. metody uczenia maszynowego - rozwiązywać złożone zadania informatyczne, w tym zadania nietypowe oraz zadania zawierające komponent badawczy - [K2st_U10]

Kompetencje społeczne:

1. rozumie, że w uczeniu maszynowym wiedza, umiejętności i narzędzia bardzo szybko stają się przestarzałe - [K2st_K1]

2. rozumie znaczenie wykorzystywania najnowszej wiedzy z zakresu uczenia maszynowego w rozwiązywaniu problemów badawczych i praktycznych - [K2st_K2]

Sposoby sprawdzenia efektów kształcenia

Efekty kształcenia 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 / ćwiczeń:

- 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.
- omówienie wyników egzaminu,

b) w zakresie laboratoriów / ćwiczeń 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 (realizacja zadań samodzielnych ma charakter opcjonalny),
- ocenę notek blogowych publikowanych na wspólnym blogu poświęconym przedmiotowi, notatki dotyczą artykułów naukowych prezentujących rozszerzenie i uszczegółowienie zagadnień poruszanych w trakcie zajęć laboratoryjnych, lub wybranych problemów i narzędzi eksploracji danych.

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. Klasyfikacja typów wiedzy: wiedza pozytywna i negatywna. Asocjacje negatywne: negatywne reguły asocjacyjne i negatywnie skorelowane. Miary atrakcyjności reguł asocjacyjnych. Typy danych sekwencyjnych. Odkrywanie wzorców sekwencji: sformułowanie problemu. Podstawowy algorytm odkrywania wzorców sekwencji. Prefiksowy algorytm odkrywania wzorców sekwencji. Odkrywanie domkniętych wzorców sekwencji. Odkrywanie wzorców sekwencji z ograniczeniami czasowymi ? sformułowanie problemu. Algorytm odkrywania wzorców sekwencji z ograniczeniami czasowym. Odkrywanie uogólnionych wzorców sekwencji. Problemy odkrywania innych wzorców sekwencji. Wprowadzenie do klasyfikacji danych. Metody klasyfikacji danych. Klasyfikacja danych poprzez indukcję drzew decyzyjnych. Algorytmy indukcji drzew decyzyjnych z wykorzystaniem miar entropii i indeksu Gini. Zjawisko przeuczenia klasyfikatora. Metody przycinania drzew decyzyjnych. Klasyfikatory regułowe: definicje podstawowych pojęć. Wywodzenie klasyfikatorów regułowych z drzew decyzyjnych. Algorytm sekwencyjnego pokrycia i ogólny algorytm ekstrakcji reguł klasyfikacyjnych. Klasyfikacja asocjacyjna: definicja problemu. Algorytmy klasyfikacji asocjacyjnej. Klasyfikatory bayesowskie. Sieci bayesowskie. Klasyfikator najbliższego sąsiedztwa. Kombinacja klasyfikatorów. Ocena jakości klasyfikatorów: miary oceny, przestrzeń i krzywa ROC. Składowe procesu grupowania. Definicje miar niepodobieństwa obiektów. 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.

Zajęcia laboratoryjne prowadzone są w formie piętnastu 2-godzinnych ćwiczeń, 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 Weka, RapidMiner, Oracle Data Mining. Wstępne przetwarzanie atrybutów z poziomu języka PL/SQL. 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. Algorytmy znajdowania zbiorów częstych i asocjacji w bazie danych Oracle. Wprowadzenie do problemów klasyfikacji, podział zbioru danych na zbiór uczący i testujący. Klasyfikatory regułowe, proste klasyfikatory drzewiaste, metody indukcji drzew decyzyjnych, miary oceny jakości podziału zbioru: indeks Giniego, entropia, Information Gain. Naiwny klasyfikator Bayesa, optymalny klasyfikator Bayesa, sieci bayerowskie. Metody oceny i testowania klasyfikatorów, wielokryterialna ocena nauczonych modeli. Miary Lift, ROC, Precision-Recall w ocenie jakości modeli. Uczenie klasyfikatorów przy pomocy macierzy kosztów. Rodzina algorytmów SVM. Zaawansowane metody klasyfikacji: metody agregacji wielu modeli poprzez głosowanie, rodzina metod ensemble, klasyfikatory wielowarstwowe. Podstawowe algorytmy analizy skupień, praktyczne ograniczenia algorytmów k-średnich i k-medoidów, algorytmy analizy skupień bazujące na gęstości, rodzina algorytmów EM analizy skupień. Niskopoziomowe interfejsy programistyczne do eksploracji danych: Java Data Mining API, Orange Data Mining Python API, Sci-Kit API, wykorzystanie narzędzi Weka i RapidMiner do pisania własnych programów wykorzystujących algorytmy eksploracji danych. Wprowadzenie do systemu R, podstawy języka, operatory i typy danych, wektoryzacja operatorów algebraicznych. Podstawowe pakiety R do eksploracji danych. Metody ekstrakcji cech: rodzina algorytmów PCA, SVD i NNMF.

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.

Literatura 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

Czynność	Czas (godz.)
1. udział w zajęciach laboratoryjnych / projektowych	12
2. udział w wykładach/ćwiczeniach audytoryjnych	24
3. udział w konsultacjach związanych z realizacją ćwiczeń laboratoryjnych	2
4. omówienie wyników egzaminu	2
5. przygotowanie do ćwiczeń laboratoryjnych	24
6. wypełnienie quizów, przygotowanie zadań samodzielnych do poszczególnych zajęć laboratoryjnych	16
7. napisanie i testowanie programów w ramach konkursów algorytmicznych	16
8. przygotowanie do egzaminu i obecność na egzaminie	16
9. zapoznanie się ze wskazaną literaturą / materiałami dydaktycznymi, 100 stron	10

Obciążenie pracą studenta

forma aktywności	godzin	ECTS
Łączny nakład pracy	122	5
Zajęcia wymagające bezpośredniego kontaktu z nauczycielem	38	2
Zajęcia o charakterze praktycznym	68	3