# 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: **16** Projekty/seminaria: **-** | **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%** |

**Odpowiedzialny za przedmiot / wykładowca:**

prof. dr hab. inż. Tadeusz Morzy
email: tadeusz.morzy@put.poznan.pl
tel. 61 6652906
Instytut Informatyki
ul. Piotrowo 2, 60-965 Poznań

**Odpowiedzialny za przedmiot / wykładowca:**

dr hab. inż. Mikołaj Morzy
email: mikolaj.morzy@put.poznan.pl
tel. 61 6653447
Instytut Informatyki
ul. Piotrowo 2, 60-965 Poznań

## Wymagania wstępne w zakresie wiedzy, umiejętności, kompetencji społecznych:

| | | |
|---|---|---|
| 1 | **Wiedza:** | Efekty kształcenia ze studiów I stopnia zdefiniowane w Uchwale Senatu PP, a szczególnie efekty K_W1-2, K_W4, K_W6-15, weryfikowane w procesie rekrutacji na studia 2 stopnia ? efekty te prezentowane są w serwisie internetowym wydziału www.fc.put.poznan.pl<br>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, PL/SQL oraz Python. |
| 2 | **Umiejętności:** | Efekty kształcenia ze studiów I stopnia zdefiniowane w Uchwale Senatu PP, a szczególnie efekty K_U1-2, K_U4, K_U7-8, K_U14-20, K_U22-23, K_U26,weryfikowane w procesie rekrutacji na studia 2 stopnia ? efekty te prezentowane są w serwisie internetowym wydziału www.fc.put.poznan.pl<br>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** | Efekty kształcenia ze studiów I stopnia zdefiniowane w Uchwale Senatu PP, a szczególnie efekty K_K1-9, weryfikowane w procesie rekrutacji na studia 2 stopnia ? efekty te prezentowane są w serwisie internetowym wydziału www.fc.put.poznan.pl<br>Ponadto 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:
- 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. ma szczegółową wiedzę w zakresie wybranych działów matematyki, ze szczególnym naciskiem na statystykę i teorię prawdopodobieństwa. - [K_W3]

2. ma uporządkowaną, podbudowaną teoretycznie wiedzę ogólną w zakresie podstawowych metod i algorytmów eksploracji danych - [K_W4]

3. ma podbudowaną teoretycznie szczegółową wiedzę związaną z wybranymi zagadnieniami z zakresu eksploracji danych, takimi jak: odkrywanie asocjacji, odkrywanie wzorców sekwencji, klasyfikacja danych, grupowanie danych. - [K_W5]

4. ma wiedzę o trendach rozwojowych i najistotniejszych nowych osiągnięciach w dziedzinie eksploracji danych. - [K_W6]

5. ma podstawową wiedzę o cyklu życia i etapach systemu odkrywania wiedzy w dużych repozytoriach danych - [K_W7]

6. zna podstawowe metody, techniki i narzędzia stosowane przy rozwiązywaniu złożonych zadań inżynierskich z dziedziny eksploracji danych - [K_W8]

**Umiejętności:**

1. potrafi pozyskiwać informacje z literatury, baz danych oraz innych źródeł (w języku ojczystym i angielskim), integrować je, dokonywać ich interpretacji i krytycznej oceny, wyciągać wnioski oraz formułować i wyczerpująco uzasadniać opinie - [K_U1]

2. potrafi określić kierunki dalszego uczenia się i zrealizować proces samokształcenia w zakresie problematyki systemów eksploracji danych - [K_U5]

3. potrafi planować i przeprowadzać eksperymenty obliczeniowe, w tym pomiary, walidacje i ewaluacje opracowanych modeli wiedzy, umie poprawnie interpretować uzyskane wyniki i wyciągać wnioski - [K_U8]

4. potrafi wykorzystać do formułowania i rozwiązywania zadań inżynierskich i prostych problemów badawczych metody analityczne, symulacyjne oraz eksperymentalne - [K_U9]

5. potrafi ? przy formułowaniu i rozwiązywaniu zadań inżynierskich ? integrować wiedzę z różnych obszarów informatyki (a w razie potrzeby także wiedzę z innych dyscyplin naukowych) oraz zastosować podejście systemowe, uwzględniające także aspekty pozatechniczne - [K_U10]

6. potrafi formułować i testować hipotezy związane z problemami inżynierskimi i prostymi problemami badawczymi w zakresie analizy i eksploracji danych - [K_U12]

7. potrafi ocenić przydatność i możliwość wykorzystania nowych osiągnięć (metod i narzędzi) oraz nowych produktów informatycznych z dziedziny eksploracji danych - [K_U13]

8. umie rozwiązywać złożone zadania informatyczne, w tym zadania nietypowe, wymagające integracji wiedzy dziedzinowej, oraz zadania zawierające komponent badawczy - [K_U23]

**Kompetencje społeczne:**

1. rozumie, że w informatyce, a szczególnie w zakresie systemów eksploracji danych, wiedza i umiejętności bardzo szybko stają się przestarzałe i potrafi systematycznie zdobywać nową wiedzę i umiejętności w tej dziedzinie - [K_K1]

2. zna przykłady i rozumie przyczyny wadliwie działających systemów eksploracji danych, które doprowadziły do poważnych strat finansowych, społecznych lub też do poważnej utraty zdrowia, a nawet życie - [K_K4]

3. potrafi odpowiednio określić priorytety służące realizacji określonego przez siebie lub innych zadania - [K_K6]

**Sposoby sprawdzenia efektów kształcenia**

Efekty kształcenia przedstawione wyżej weryfikowane są w następujący sposób:

Ocena formująca:

a) w zakresie wykładów:

- na podstawie odpowiedzi na pytania dotyczące materiału omówionego na poprzednich wykładach oraz ćwiczeń realizowanych 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. Wykrywanie punktów osobliwych.

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.

Cześć wymienionych wyżej treści programowych realizowana jest w ramach pracy własnej studenta.

Metody dydaktyczne:

1. wykład: prezentacja multimedialna, prezentacja ilustrowana przykładami podawanymi na tablicy, wspólne rozwiązywanie zadań na tablicy,
2. ćwiczenia: wspólne rozwiązywanie zadań praktycznych, dyskusja i analiza przypadków użycia.
3. ćwiczenia laboratoryjne: prezentacja multimedialna, ćwiczenia praktyczne, wykonywanie eksperymentów, dyskusja, gry obliczeniowe, konkursy programistyczne, quizy.

## Literatura podstawowa:

1. Eksploracja danych: metody i algorytmy, T. Morzy, PWN, 2013.
2. Data Mining: Concepts and Techniques, Han, J., Kamber, M., Pei, J., Morgan Kaufmann, 2012.
3. Uczenie maszynowe i sieci neuronowe, Krawiec, K., Stefanowski, J., Wydawnictwo PP, 2003.

## Literatura uzupełniająca:

1. Statystyczne systemy uczące się, Koronacki, J., Ćwik, J., WNT, 2005.
2. Programmer?s Guide to Data Mining, Zacharski, R. http://guidetodatamining.com/
3. Machine Learning, Ng, A., https://www.coursera.org/course/ml
4. Introduction to Data Mining, Tan, P-N., Steinbach, M., Kumar, V., Pearson Education, 2006.
5. Data Mining: Concepts and Techniques, Han, J., Kamber, M., Pei, J., Morgan Kaufmann, 2012.
6. Systemy uczące się, Cichosz, P., WNT, 2000.

## Bilans nakładu pracy przeciętnego studenta

| Czynność | Czas (godz.) |
|---|---|

| | |
|---|---|
| 1. udział w zajęciach laboratoryjnych / ćwiczeniach | 24 |
| 2. przygotowanie do ćwiczeń laboratoryjnych | 16 |
| 3. dokończenie (w ramach pracy własnej) ćwiczeń laboratoryjnych, wypełnienie quizów, przygotowanie zadań samodzielnych | 15 |
| 4. udział w konsultacjach związanych z realizacją procesu kształcenia, w szczególności ćwiczeń laboratoryjnych | 5 |
| 5. napisanie i testowanie programów w ramach konkursów algorytmicznych (czas poza zajęciami laboratoryjnymi) | 10 |
| 6. przygotowanie do sprawdzianów / kolokwium | 5 |
| 7. zapoznanie się ze wskazaną literaturą / materiałami dydaktycznymi (10 stron tekstu naukowego = 1 godz.), 100 stron | 10 |
| 8. omówienie wyników egzaminu | 2 |
| 9. udział w wykładach | 16 |
| 10. przygotowanie do egzaminu i obecność na egzaminie: 20 godz. + 2 godz. | 22 |

**Obciążenie pracą studenta**

| forma aktywności | godzin | ECTS |
|---|---|---|
| Łączny nakład pracy | 125 | 5 |
| Zajęcia wymagające bezpośredniego kontaktu z nauczycielem | 49 | 2 |
| Zajęcia o charakterze praktycznym | 65 | 3 |