Program kursu *Data mining dla niezaawansowanych*

Informacje wstępne

Liczba godzin	30
Język wykładowy	Polski
Punkty ECTS dla kształcenia na odległość	3
Wymagania wstępne	– Znajomość matematyki na poziomie szkoły średniej. – Podstawowa umiejętność posługiwania się komputerem z systemem Windows. – Samodyscyplina, umiejętność gospodarowania własnym czasem. – Umiejętność czytania ze zrozumieniem tekstów napisanych językiem formalnym. – Otwartość na nowe wyzwania. – Świadomość konieczności samodoskonalenia się. – Odwaga w zadawaniu pytań oraz wypowiadaniu się na forum publicznym.
Skrócony opis	Duże zbiory danych pojawiają się aktualnie w każdym obszarze, w którym gromadzone są informacje. Oznacza to konieczność wykształcenia powszechniej sprawności w pracy z takimi danymi. Celem przedmiotu jest zapoznanie uczestników zajęć z najważniejszymi algorytmami data-miningu oraz wykształcenie umiejętności analizy danych z wykorzystaniem programu PS IMAGO PRO (IBM SPSS Statistics).
Całkowity nakład pracy	– Zajęcia prowadzone synchronicznie – 0 h, – zapoznanie się z materiałami dostępnymi na platformie Moodle – 30 h, – praca własna: studiowanie literatury, wykonywanie ćwiczeń, rozwiązywanie zadań, konsultacje z prowadzącymi zajęcia – 50 h, – rozwiązywanie testów i zadań zaliczeniowych 10 h. Razem: 90 h (3 pkt. ECTS)

Efekty uczenia się osiągane poprzez uczestnictwo w kursie

Wiedza	W1. Rozumie potrzebę pozyskiwania wiedzy z danych. Zna podstawowe problemy eksploracji danych. W2. Zna wybrane algorytmy eksploracji danych i wie, które z nich stosują się do określonego typu zagadnień z tego zakresu. W3. Ma wiedzę na temat dostępnego oprogramowania służącego do eksploracji danych.
Umiejętności	U1. Potrafi znaleźć potrzebne dane w zbiorach danych ogólnie dostępnych, umie pobrać dane i poddać je analizie. U2. Umie zaproponować odpowiednie algorytmy eksploracji danych do konkretnego zagadnienia, w tym klasyfikacji, grupowania, szacowania i budowania

	reguł, oraz wyselekcjonować z ich użyciem najlepszy model. U3. Umie posługiwać się w stopniu podstawowym programem do eksploracji danych PS IMAGO PRO (IBM SPSS Statistics). U4. Potrafi przygotować raport z wynikami swoich analiz.
Kompetencje społeczne	K1. Potrafi sformułować problem eksploracji danych w sposób zrozumiały zarówno dla osób, z którymi współpracuje w tym obszarze, jak i ekspertów analityków. K2. Potrafi poddać krytycznej ocenie pozyskane dane. K3. Potrafi czerpać wiedzę z danych i na tej podstawie formułować propozycje rozwiązania sytuacji problemowych.

Metody i techniki kształcenia stosowane przy realizacji kursu

Dydaktyczne	– Metody odnoszące się do autentycznych lub fikcyjnych sytuacji (zbiory danych rzeczywistych i przykłady ich analiz), – metody wymiany i dyskusji (fora dyskusyjne), – metody ewaluacyjne (testy i zadania podlegające ocenie).
Dydaktyczne eksponujące	Pokaz (filmy instruktażowe demonstrujące działanie programu).
Dydaktyczne podające	Wykład informacyjny (nagrania krótkich wykładów dotyczących teorii).
Dydaktyczne poszukujące	– Ćwiczeniowa (testy i zadania zlecone przez prowadzących), – studium przypadku.

Program przedmiotu

Szczegółowy program przedmiotu	W czasie kursu uczestnik zapozna się z programem PS IMAGO PRO oraz jego możliwościami. Realizowane będą następujące zagadnienia: 1. **Otwarte źródła informacji**: big data i data mining, obowiązek ustawowy eksploracji danych, zadania eksploracji danych, jawne źródła danych (2 godziny) - celem zajęć jest przedstawienie założeń i możliwości eksploracji danych oraz uświadomienie słuchaczy co do dostępności danych m.in. w Internecie. 2. **Podstawy pracy w PS IMAGO PRO**: tabele danych, poziom pomiaru zmiennych, specyfikacja właściwości zmiennych w PS IMAGO PRO (2 godziny) - celem zajęć jest przedstawienie podstawowych możliwości okna danych programu PS IMAGO PRO oraz zwrócenie uwagi na konieczność sprawdzenia jakości danych i specyfikacji zmiennych. 3. **Statystyka opisowa**: miary tendencji centralnej, miary rozproszenia, wykres skrzynka z wąsami, histogramy, standaryzacja i normalizacja zmiennych, zasada trzech sigma, rozkład normalny (2 godziny) - celem zajęć jest

	przypomnienie podstawowych statystyk i wykresów używanych w analizie statystycznej zmiennych jakościowych i ilościowych. 4. **Eksploracyjna analiza danych:** opis danych jedno- i wielowymiarowych, współczynniki korelacji Pearsona i Spearmana, wizualizacja danych (2 godziny) - celem zajęć jest omówienie najważniejszych typów wizualizacji danych oraz przedstawienie możliwości programu PS IMAGO PRO w zakresie tworzenia wizualizacji danych. 5. **Klasyfikacja metodą *k* najbliższych sąsiadów**: przygotowanie zmiennych do analizy, zasada działania algorytmu, algorytm k najbliższych sąsiadów w PS IMAGO PRO (2 godziny) - celem zajęć jest omówienie algorytmu *k* najbliższych sąsiadów oraz jego implementacji w programie PS IMAGO PRO. 6. **Ocena jakości modelu**: miary jakości klasyfikacji i szacowania (2 godziny) - celem zajęć jest omówienie procesu oceny modelu, w tym podziału na zbiory uczący i testowy, oraz najważniejszych miar jakości klasyfikacji (trafność, czułość, swoistość) i szacowania (średni błąd bezwzględny i pierwiastek z błędu średniokwadratowego). 7. **Drzewa klasyfikacyjno-regresyjne CRT**: budowa i interpretacja drzew klasyfikacyjno-regresyjnych, zapobieganie przeuczeniu, poprawa jakości drzew, budowa drzew z wykorzystaniem programu PS IMAGO PRO, predykcja z użyciem zbudowanych modeli (2 godziny) - celem zajęć jest omówienie modelu drzew klasyfikacyjno-regresyjnych oraz wykształcenie praktycznej umiejętności tworzenia takich modeli w programie PS IMAGO PRO i stosowania ich do klasyfikacji i szacowania. 8. **Sieci neuronowe**: budowa i uczenie modelu perceptronu wielowarstwowego (2 godziny) - celem zajęć jest zapoznanie słuchaczy z tematyką sieci neuronowych na przykładzie perceptronu wielowarstwowego, omówienie budowy tego typu sieci oraz tworzenie gotowych modeli w PS IMAGO PRO. 9. **Prosta regresja liniowa**: model prostej regresji liniowej, interpretacja jego parametrów oraz ocena jakości (2 godziny) - celem zajęć jest wprowadzenie modeli liniowych na przykładzie prostej regresji liniowej. 10. **Wieloraka regresja liniowa**: model regresji wielorakiej, ocena jakości modelu oraz metody wyboru zmiennych (2 godziny) - celem zajęć jest rozszerzenie modelu prostej regresji liniowej na przypadek z wieloma predyktorami oraz omówienie dostępnych w programie PS IMAGO PRO metod wyboru zmiennych do modelu. 11. **Grupowanie metodą *k*-średnich**: cel i zastosowania grupowania, algorytm k średnich, jego własności, wady i zalety (2 godziny) - celem zajęć jest zapoznanie słuchaczy z algorytmem *k* średnich, możliwością wykorzystania programu PS IMAGO PRO do wyznaczenia grup z wykorzystaniem tego algorytmu oraz wykształcenie umiejętności interpretacji uzyskanych klastrów. 12. **Hierarchiczna analiza skupień**: miary

	podobieństwa/niepodobieństwa dla zmiennych ilościowych i binarnych, metody określania odległości grup, tworzenie i interpretacja dendrogramów (2 godziny) - celem zajęć jest wprowadzenie metody grupowania hierarchicznego oraz wykształcenie umiejętności odczytywania dendrogramów. 13. **Dwustopniowa analiza skupień**: opis algorytmu dwustopniowej analizy skupień oraz interpretacja otrzymanych grup (2 godziny) - celem zajęć jest wykształcenie umiejętności znajdowania grup z wykorzystaniem implementacji algorytmu dwustopniowej analizy skupień w programie PS IMAGO PRO, a także interpretacji i ewaluacji otrzymanych klastrów. 14. **Analiza koszykowa**: typy i miary jakości reguł asocjacyjnych, algorytm A priori (2 godziny) - celem zajęć jest wprowadzenie do zagadnienia budowy reguł asocjacyjnych na przykładach z zakresu analizy koszykowej, demonstracja działania algorytmu A priori oraz ocena jakości reguł uzyskanych z wykorzystaniem tego algorytmu. 15. **Studium przypadku** (2 godziny) - celem zajęć jest utrwalenie zdobytej przez słuchaczy wiedzy poprzez wykonanie w programie PS IMAGO PRO zadania wymagającego wykorzystania poznanych wcześniej algorytmów. Każdy temat będzie zawierał krótki wstęp teoretyczny, materiał instruktażowy, jak rozwiązać dany problem za pomocą PS IMAGO PRO, przykładowe zbiory danych, ćwiczenia do wykonania i zadania do rozwiązania.
Literatura tradycyjna	1. Bedyńska S., Książek M. (red.): *Statystyczny drogowskaz 1. Praktyczne wprowadzenie* do *wnioskowania statystycznego*. Warszawa, Wydawnictwo Akademickie Sedno, 2012. 2. Górniak J., Wachnicki J.: *Pierwsze kroki w analizie danych*. Kraków, SPSS Polska, 2004. 3. Daniel T. Larose: *Odkrywanie wiedzy z danych*, Wydawnictwo Naukowe PWN, Warszawa, 2013. 4. Daniel T. Larose: *Metody i modele eksploracji danych*, Wydawnictwo Naukowe PWN, Warszawa, 2012. 5. Tadeusz Morzy: *Eksploracja danych. Metody i algorytmy*. Wydawnictwo Naukowe PWN, Warszawa, 2013. 6. Bartosz Saramak: *Wykorzystanie otwartych źródeł informacji w działalności wywiadowczej: historia, praktyka, perspektywy*. WDiNP UW, Warszawa, 2015 (dostępne on-line: https://wnpism.uw.edu.pl/wp-content/uploads/2019/08/Wykorzystanie-otwartych-zrodel.pdf, dostęp z dn. 22.10.2020). 7. Sobczyk M.: *Statystyka, Wydanie piąte uzupełnione*. Warszawa, Wydawnictwo Naukowe PWN, 2020, str. 12-16. 8. Stephane Tuffery: *Data Mining and Statistics for Decision Making*. Wiley, 2011. 9. Xindong Wu, Vipin Kumar: *The Top Ten Algorithms in Data Mining*. Chapman & Holl/CRC, 2009.

Wykorzystywane e-materiały	– Biecek P.: *Odkrywać! Ujawniać! Objaśniać! Zbiór esejów o sztuce prezentowania danych*. http://www.biecek.pl/Eseje/ (dostęp z dn. 23.10.2020). – *Polska Statystyka Publiczna* https://bip.stat.gov.pl/dzialalnosc-statystyki-publicznej/polska-statystyka-publiczna/ (dostęp z dn. 21.10.2020). – Strona internetowa *Discovering Statistics* z tutorialami, filmami i zbiorami danych https://www.discoveringstatistics.com/ (dostęp z dn. 24.08.2020). – Strona internetowa IBM Knowledge Center z dokumentacją produktu IBM SPSS Statistics https://www.ibm.com/support/knowledgecenter/pl/SSLVMB_26.0.0/statistics_kc_ddita/spss/product_landing.html (dostęp z dn. 25.08.2020). – Predictive Solutions: *Samouczek*. http://samouczek.predictivesolutions.pl/ (dostęp z dn. 12.09.2020). – Predictive Solutions: *Blog*. https://predictivesolutions.pl/2,blog (dostęp z dn. 29.10.2020). – StatSoft: *Internetowy Podręcznik Statystyki*. https://www.statsoft.pl/textbook/stathome.html (dostęp z dn. 25.10.2020). – Strona internetowa *Towards Data Science*: https://towardsdatascience.com/ (dostęp z dn. 30.10.2020). – Portal *Sztuczna Inteligencja* https://www.sztucznainteligencja.org.pl/ (dostęp z dn. 29.10.2020). – Narzędzia doboru kolorów na wykresach: - http://paletton.com (dostęp z dn. 23.10.2020), - https://colorbrewer2.org/ (dostęp z dn. 23.10.2020). – Przemysław Biecek: Przewodnik po pakiecie R 4.0 http://pbiecek.github.io/Przewodnik/ (dostęp z dn. 30.10.2020). – Strona domowa projektu R https://www.r-project.org/ (dostęp z dn. 27.10.2020).

Weryfikacja efektów uczenia się i ewaluacja

Kryteria oceniania	Ocenie podlegają aktywności zamieszczone w kolejnych modułach, w tym: testy zawierające pytania teoretyczne oraz pytania odnoszące się do ćwiczeń praktycznych, krótkie zadania do wykonania w programie oraz końcowe zadanie zaliczeniowe w formie studium przypadku.
Sposób przeprowadzania oceny	Ocena jest przeprowadzana zdalnie. Kursant otrzymuje punkty za wszystkie wymienione aktywności, przy czym przy każdej z nich podana jest informacja o maksymalnej możliwej do zdobycie liczbie punktów. Wszystkie aktywności powinny być realizowane zgodnie z harmonogramem kursu. Prowadzący zajęcia może obniżyć ocenę za aktywności wykonane po terminie.

	Do zaliczenia kursu wymagane jest uzyskanie co najmniej połowy punktów możliwych do zdobycia.
Ewaluacja zajęć	– Fora zamieszczone przy każdym module i służące do zadawania pytań pozwolą na bieżącą ewaluację zajęć oraz reagowanie w sytuacjach wymagających korekty materiałów. – Po każdym module zostanie zamieszczona krótka ankieta dotycząca poziomu trudności zajęć oraz stopnia ich zrozumienia. – Kurs będzie kończył się ankieta pozwalająca na wszechstronną ocenę zajęć: poziomu merytorycznego, jakości materiałów, stopnia zadowolenia słuchacza z kursu.