Politechnika Krakowska im. Tadeusza Kościuszki

Karta przedmiotu

obowiązuje studentów rozpoczynających studia w roku akademickim 2022/2023

Wydział Informatyki i Telekomunikacji

Kierunek studiów: Matematyka | Profil: Ogólnoakademicki

Forma sudiów: stacjonarne | Kod kierunku: M

Stopień studiów: II

Specjalności: Modelowanie matematyczne,Matematyka w finansach i ekonomii

1 Informacje o przedmiocie

Nazwa przedmiotu	Zaawansowana eksploracja dużych zbiorów danych
Nazwa przedmiotu w języku angielskim	Advanced exploration of big datasets
Kod przedmiotu	WIiT M oIIS C11 22/23
Kategoria przedmiotu	Przedmioty kierunkowe
Liczba punktów ECTS	4.00
Semestry	4

2 Rodzaj zajęć, liczba godzin w planie studiów

Semestr	Wykład	Ćwiczenia	Laboratorium	Laboratorium komputerowe	Seminarium	Projekt
4	30	0	0	0	0	30

3 Cele przedmiotu

Cel 1 Zapoznanie studentów z ważniejszymi algorytmami i metodami stosowanymi obecnie do przechowywania, przetwarzania, analizy, modelowania i wizualizacji olbrzymich ilości danych.

Cel 2 Zapoznanie studentów z ważniejszym oprogramowaniem stosowanym do przechowywania, przetwarzania, analizy, modelowania i wizualizacji olbrzymich ilości danych. Studenci zdobędą praktyczną wiedzę, jak zainstalować RStudio oraz darmowe, gotowe i nowoczesne (2021 i 2022) biblioteki (tzw. pakiety) w języku R,

oraz zdobędą praktyczne umiejętności, jak wykorzystywać wspomniane biblioteki, zwłaszcza do przetwarzania, analizy, tworzenia modeli i wizualizacji danych, zarówno na komputerze, jak i online w Chmurze; jak poprawnie interpretować uzyskane z algorytmów wyniki oraz zdobędą praktyczne umiejętności, jak tworzyć wizualizację danych, zwłaszcza za pomocą internetowych notebooków i dashboard'ów udostępnionych przez RStudio dzięki bibliotekom rmarkdown i shiny.

Cel 3 Celem będzie nabycie umiejętności samodzielnego poszerzania swojej wiedzy i doskonalenia umiejętności w dziedzinie modelowania matematycznego, także docenienia wiedzy z tego zakresu w kształtowaniu współczesnej matematyki. Celem będzie tez nabycie umiejętności, aby w sposób staranny i terminowy realizować powierzone sobie zadania, oraz aby być gotowym do rozwiązywania problemów ze wspomnianego zakresu, zarówno w ramach pracy indywidualnej, jak i grupowej, a także nabycie umiejętności poszukiwania niezbędnej w tym zakresie wiedzy oraz umiejętności pracy w małych zespołach.

4 Wymagania wstępne w zakresie wiedzy, umiejętności i innych kompetencji

1 Znajomość podstaw obsługi komputera.

2 Podstawowa znajomość języka angielskiego.

3 Podstawowa wiedza z zakresu statystyki.

5 Efekty kształcenia

EK1 Wiedza Student będzie potrafił wytłumaczyć pojęcia oraz stosowane obecnie metody i modele służące przechowywania, przetwarzania, analizy i wizualizacji olbrzymich ilości danych. Student będzie potrafił wytłumaczyć działanie oraz zinterpretować wyniki ważniejszych algorytmów stosowanych przy przetwarzaniu, analizie i wizualizacji danych.

EK2 Umiejętności Student będzie posiadał umiejętność zastosowania do przechowywania, przetwarzania, analizy i wizualizacji olbrzymich ilości danych ważniejszych algorytmów (tzw. funkcji) zawartych w wybranych bibliotekach (pakietach) języka R i środowiska zintegrowanego RStudio.

EK3 Umiejętności Student będzie potrafił poprzez RStudio łączyć się internetowo zarówno ze SPARKiem, tj. z uniwersalnym silnikiem dla Big data, jak i z darmowymi serwerami RStudio, wspomagającymi przetwarzanie dużych zbiorów danych w Chmurze. Student będzie również posiadał umiejętność samodzielnego programowania w języku R w środowisku RStudio.

EK4 Kompetencje społeczne Student nabędzie umiejętność pracy w grupie, pracy indywidualnej,samokształcenia, umiejętność komunikacji z nauczycielem i środowiskiem pozauczelnianym w celu popularyzacji i przedstawiania uzyskanych rezultatów w zrozumiały sposób, samodzielnego poszerzania swojej wiedzy i doskonalenia umiejętności w swojej dziedzinie, także docenienia wiedzy z tego zakresu w kształtowaniu współczesnej matematyki, nabędzie także umiejętność, aby w sposób staranny i terminowy realizować powierzone sobie zadania, oraz do bycia gotowym do rozwiązywania problemów ze wspomnianego zakresu, zarówno w ramach pracy indywidualnej, jak i grupowej, a także nabędzie umiejętność poszukiwania niezbędnej w tym zakresie wiedzy oraz umiejętność pracy w małych zespołach. Student zauważy też potrzebę samokształcenia i potrzebę ciągłego uaktualniania swej wiedzy.

6 Treści programowe

Projekt		
Lp	Tematyka zajęć Opis szczegółowy bloków tematycznych	Liczba godzin

Projekt		
Lp	Tematyka zajęć Opis szczegółowy bloków tematycznych	Liczba godzin
P1	Treści programowe 1. Instalacja RStudio, R, ekosystemu Apache SPARK, podstawowych pakietów oferowanych przez RStudio oraz przez R. Proste eksperymenty w środowisku RStudio z wykorzystaniem języka R.	2
P2	Treści programowe 2. Instalacja wybranych pakietów sposród zbioru ponad 18 tysięcy, omówienie ich oraz wywoływanie w RStudio gotowych kodów zródłowych dostepnych online na podanej studentom stronie, a takze proste eksperymenty w srodowisku RStudio z wykorzystaniem zainstalowanych pakietów.	2
P3	Treści programowe 3. Instalacja bogatej rodziny "tidyverse" złożonej z kilkunastu nowoczesnych pakietów RStudio, omówienie ich oraz eksperymentowanie z tymi pakietami wzorując się na przykładach i tutoriałach dostępnych online na stronach: http://r4ds.had.co.nz/ ; https://github.com/tidyverse ; http://tidyverse.org ; https://github.com/hadley/r4ds ; https://github.com/hadley/dplyr ; https://github.com/tidyverse/ggplot2	4
P4	Treści programowe 4. Instalacja pakietu sparklyr i dplyr oraz eksperymentowanie z funkcjami tych pakietów: select(), filter(), arrange(), rename(), mutate(), group_by() oraz z operatorem pipeline, tj. "%>%".	2
P5	Treści programowe 5. Instalacja i eksperymentowanie z pakietami RStudio i języka R - pozwalającymi na tworzenie ważniejszych modeli dla dużych zbiorów danych (takich jak: modele liniowe, modele drzew regresyjnych i klasyfikacyjnych, modele analizy skupień, i innych modeli). Instalowane pakiety - to: rpart, maptree, rattle, party, randomForest, SVM, naiveBayes, boosting, cluster i kilka innych dodatkowych.	6
P6	Treści programowe 6. Instalacja i eksperymentowanie z pakietem ISLR. Eksperymenty będą bazowały na licznych przykładach zawartych w licznych online laboratoryjnych przykładach i tutoriałach dostępnych na stronie, gdzie są wszystkie kody źródłowe z książki [5]: http://www-bcf.usc.edu/gareth/ISL/code.html ; Książka dostępna jest pod adresem: http://www-bcf.usc.edu/gareth/ISL Dodatkowo, dużo przykładów autora książki jest na jego stronie: http://web.stanford.edu/hastie/StatLearnSparsity/ (Jest to światowej sławy autor współpracujący przy tworzeniu oprogramowania dla ekosystemu Apache SPARK (i platformy H20), uniwersalnego silnika dla Big Data).	4
P7	Treści programowe 7. Instalacja i eksperymentowanie z pakietami języka R - pozwalającymi na tworzenie internetowych notebooków. Instalowane pakiety - to: magrittr, Rmarkdown i kilka innych mniejszych.	2
P8	Treści programowe 8. Instalacja i eksperymentowanie z pakietami języka R - pozwalającymi na tworzenie dashboardów. Instalowany pakiet, to bogaty w swych funkcjach pakiet shiny. Eksperymenty będą bazowały na licznych przykładach zawartych w online tutoriałach dostępnych na stronach: http://www.rstudio.com/shiny/ ; http://rstudio.github.io/shiny/tutorial/ ; http://www.rstudio.com/shiny/lessons/Intro/	4

Projekt		
Lp	Tematyka zajęć Opis szczegółowy bloków tematycznych	Liczba godzin
P9	Treści programowe 9. Instalacja i eksperymentowanie z pakietami języka R - pozwalającymi na tworzenie grafów zależności istniejących w dużych zbiorach danych. Instalowane pakiety - to: graphframe, igraph, rgl, snowfall, network, tmap i kilka innych dodatkowych.	2
P10	Treści programowe 10. Instalacja i eksperymentowanie z pakietami języka R - pozwalającymi na dokonywanie wzajemnych porównań (podobieństwa i odróżnialności) zbiorów tekstowych zawierających olbrzymie ilości danych. Instalowane pakiety - to: tm, lda, topicmodels, RTextTools, wordcloud, i kilka innych dodatkowych.	2

Wykład		
Lp	Tematyka zajęć Opis szczegółowy bloków tematycznych	Liczba godzin
W1	Treści programowe 1. Charakterystyka pojęcia "Big data" i sumaryczne omówienie zagadnień związanych z tym pojęciem. Wprowadzenie do paradygmatu MapReduce. Przedstawienie i opisanie architektury i funkcji, jakie spełnia ekosystem Apache SPARK i jego podsystemy. Wprowadzenie do środowiska RStudio.Opisanie pakietu sparklyr (autorstwa twórców środowiska RStudio), pozwalającego na połączenie się z programu w języku R ze SPARKiem oraz pozwalającego na wykonywanie dowolnego pakietu Ra w środowisku skalowalnym i rozproszonym. Omówienie pojęcia ramki danych (tj. DataFrame) oraz ważniejszych operacji na dużych zbiorach danych. Przedstawienie i opisanie przykładów skryptów napisanych z wykorzystaniem pakietu sparklyr.	4
W2	Treści programowe 2. Opisanie środowiska RStudio, jego ważniejszych pakietów, ich funkcji i operacji, oraz zalet i metody pracy w tym środowisku. Omówienie metod klasyfikacji i regresji oraz Systemów uczących, korzystając z bogatych materiałów dostępnych na stronach SPARK'a.	4
W3	Treści programowe 3. Omówienie metod tworzenia notebooków - korzystając z pakietów RStudio: magrittr, markdown, Rmarkdown oraz dplyr. Omówienie metod tworzenia dashboardów - korzystając z pakietów RStudio: magrittr, markdown, R markdown oraz shiny.	6
W4	Treści programowe 4. Omówienie ważniejszych operacji bazodanowych udostępnionych przez funkcje ważniejszych pakietów bazo-danowych środowiska RStudio, np. operacji select, filter, aggregate, operacji na kolumnach, funkcji collect() oraz subset().	4
W5	Treści programowe 5. Opisanie ważniejszych pakietów Ra służących do eksploracji danych tekstowych (text mining): tm, lda, topicmodels, RTextTools, tau, wordcloud. Przedstawienie i opisanie przykładów skryptów napisanych z wykorzystaniem tych pakietów.	4

Wykład		
Lp	Tematyka zajęć Opis szczegółowy bloków tematycznych	Liczba godzin
W6	Treści programowe 6. Opisanie ważniejszych pakietów Ra służących do eksploracji sieci społecznościowych (social networks). Przedstawienie i opisanie przykładów skryptów napisanych z wykorzystaniem pakietów: sna, network, igraph, SocialNetworks, tmap, spnet.	4
W7	Treści programowe 7. Omówienie zasad i reguł potrzebnych do eksperymentowania w Chmurze na dużych zbiorach danych.Omówienie metody tworzenia i importowania notebooków w Chmurze.	4

7 Narzędzia dydaktyczne

N1 Wykłady (w przypadku realizacji zajęć w trybie zdalnym z wykorzystaniem stosownych narzędzi teleinformatycznych)

N2 Cwiczenia laboratoryjne

N3 Prezentacje multimedialne (w przypadku realizacji zajęć w trybie zdalnym z wykorzystaniem stosownych narzędzi teleinformatycznych)

N4 Konsultacje (w przypadku realizacji zajęć w trybie zdalnym z wykorzystaniem stosownych narzędzi teleinformatycznych)

N5 Dyskusja (w przypadku realizacji zajęć w trybie zdalnym z wykorzystaniem stosownych narzędzi teleinformatycznych)

N6 Praca w 2-3 osobowych grupkach

8 Obciążenie pracą studenta

Forma aktywności	Średnia liczba godzin na zrealizowanie aktywności
Godziny kontaktowe z nauczycielem akademickim, w tym:	
Godziny wynikające z planu studiów	60
Konsultacje przedmiotowe	10
Egzaminy i zaliczenia w sesji	0
Godziny bez udziału nauczyciela akademickiego wynikające z nakładu pracy studenta, w tym:	
Przygotowanie się do zajęć, w tym studiowanie zalecanej literatury	10
Opracowanie wyników	10
Przygotowanie raportu, projektu, prezentacji, dyskusji	30
Sumaryczna liczba godzin dla przedmiotu wynikająca z całego nakładu pracy studenta	**120**
Sumaryczna liczba punktów ECTS dla przedmiotu	4.00

9 Sposoby oceny

Ocena formująca

F1 Cwiczenia praktyczne

F2 Odpowiedzi ustne

F3 Sprawozdania z umiejetności wykorzystania wybranych bibliotek jezyka R wywoływanych w środowisku RStudio i Apache SPARK.

Ocena podsumowująca

P1 Średnia ważona ocen formujących

Warunki zaliczenia przedmiotu

W1 Warunkiem otrzymania zaliczenia z przedmiotu jest uzyskanie wystarczajacej liczby punktów za aktywnie wykonywane cwiczenia praktyczne, za odpowiedzi ustne podczas zajec laboratoryjnych oraz za oddanie wszystkich zleconych do napisania sprawozdan wykonanych w środowisku programistycznym RStudio i Apache SPARK, z wykorzystaniem wspomnianych bibliotek. Wszystko to będzie punktowane.

Ocena aktywności bez udziału nauczyciela

B1 Oddanie wszystkich zleconych do napisania sprawozdań wykonanych w srodowisku programistycznym RStudio i Apache SPARK. Wszystkie sprawozdania będą punktowane.

B2 Ocena za odpowiedzi ustne podczas zajeć.

B3 Ocena za aktywność podczas wykonywania ćwiczeń praktycznych w klasie.

Kryteria oceny

Efekt kształcenia 1	
Na ocenę 2.0	Student nie spełnia warunków określonych dla oceny 3.0
Na ocenę 3.0	Opanowanie zagadnienia w stopniu powyżej 50%.
Na ocenę 3.5	Opanowanie zagadnienia w stopniu powyżej 60%.
Na ocenę 4.0	Opanowanie zagadnienia w stopniu powyżej 70%.
Na ocenę 4.5	Opanowanie zagadnienia w stopniu powyżej 80%.
Na ocenę 5.0	Opanowanie zagadnienia w stopniu powyżej 90%.
Efekt kształcenia 2	
Na ocenę 2.0	Student nie spełnia warunków określonych dla oceny 3.0
Na ocenę 3.0	Opanowanie zagadnienia w stopniu powyżej 50%.
Na ocenę 3.5	Opanowanie zagadnienia w stopniu powyżej 60%.
Na ocenę 4.0	Opanowanie zagadnienia w stopniu powyżej 70%.
Na ocenę 4.5	Opanowanie zagadnienia w stopniu powyżej 80%.
Na ocenę 5.0	Opanowanie zagadnienia w stopniu powyżej 90%.
Efekt kształcenia 3	
Na ocenę 2.0	Student nie spełnia warunków określonych dla oceny 3.0
Na ocenę 3.0	Opanowanie zagadnienia w stopniu powyżej 50%.
Na ocenę 3.5	Opanowanie zagadnienia w stopniu powyżej 60%.
Na ocenę 4.0	Opanowanie zagadnienia w stopniu powyżej 70%.
Na ocenę 4.5	Opanowanie zagadnienia w stopniu powyżej 80%..
Na ocenę 5.0	Opanowanie zagadnienia w stopniu powyżej 90%.
Efekt kształcenia 4	
Na ocenę 2.0	Student nie spełnia warunków określonych dla oceny 3.0
Na ocenę 3.0	Opanowanie zagadnienia w stopniu powyżej 50%.
Na ocenę 3.5	Opanowanie zagadnienia w stopniu powyżej 60%.
Na ocenę 4.0	Opanowanie zagadnienia w stopniu powyżej 70%.
Na ocenę 4.5	Opanowanie zagadnienia w stopniu powyżej 80%.

Na ocenę 5.0	Opanowanie zagadnienia w stopniu powyżej 90%.

10 Macierz realizacji przedmiotu

Efekt kształcenia	Odniesienie danego efektu do szczegółowych efektów zdefiniowanych dla programu	Cele przedmiotu	Treści programowe	Narzędzia dydaktyczne	Sposoby oceny
EK1	K_W01 K_W02 K_W04 K_W06 K_W07 K_W08 K_W10 K_W11 K_W12	Cel 1	W1 W2 W3 W4 W5 W6 W7	N1 N2 N5 N6	F1 F2 F3 P1
EK2	K_U02 K_U04 K_U10 K_U11 K_U12 K_U13 K_U19	Cel 2	P1 P2 P3 P4 P5 P6 P7 P8 P9 P10 W1 W2 W3 W4 W5 W6 W7	N1 N2 N3 N4 N5 N6	F1 F2 F3 P1
EK3	K_U02 K_U03 K_U04 K_U05 K_U10 K_U11 K_U12 K_U13 K_U15 K_U16 K_U19 K_U20 K_U21 K_U22	Cel 2	P1 P2 P3 P4 P5 P6 P7 P8 P9 P10 W1 W2 W3 W4 W5 W6 W7	N1 N2 N3 N4 N5 N6	F1 F2 F3
EK4	K_K01 K_K02 K_K03 K_K04 K_K05 K_K06 K_K07	Cel 3	P1 P2 P3 P4 P5 P6 P7 P8 P9 P10	N1 N2 N3 N4 N5 N6	F1 F2 F3 P1

11 Wykaz literatury

Literatura podstawowa

[1] Liczne przykłady, gotowe Notebooki wraz z tutoriałami i możliwość tworzenia własnych online w Chmurze z użyciem języka R i środowiska RStudio: https://spark.rstudio.com/index.html, https://stat545.com/index.html, https://rstudio.cloud/, i pod adr.: https://www.kaggle.com, też pod adr.: https://datascienceplus.com i pod adr.: http://www.rdatamining.com/docs, pod adr.: https://rpubs.com, pod adr.: https://rnotebook.io/

[2] Roger D. Peng, R programming for data science, książka dostępna online na str.: http://www.cs.upc.edu/-robert/teaching/estadistica/rprogramming.pdf

[3] G. Grolemund, H. Wickham, R for Data Science, książka dostępna online na stronie: http://r4ds.had.co.nz/; tutoriały i kody są dostępne na stronach: https://github.com/hadley/r4ds, https://github.com/hadley/dplyr, https://github.com/tidyverse/ggplot2, http://tidyverse.org, https://github.com/tidyverse.

[9] G. James, D. Witten, Trevor Hastie, Robert Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer Series in Statistics, 2016, Stanford, CA, książka dostępna w Internecie na str.: http://www-bcf.usc.edu/gareth/ISL (kody źródłowe z książki są na str.: http://www-bcf.usc.edu/gareth/ISL/code.html); dużo przykładów autora jest na str.: http://web.stanford.edu/hastie/StatLearnSparsity/

[10] Y. Zhao, "R and Data Mining: Examples and Case Studies", 2014, ksiazka dostepna Online z licznymi innymi materiałami: http://www.rdatamining.com/docs/introduction-to-data-mining-with-r i przykłady w R: http://www.rdatamining.com/examples i wiele inn. plików: http://www.rdatamining.com

Literatura uzupełniająca

[1] Materiały w wersji elektronicznej dostarczone studentom na pierwszych laboratoriach.

12 Informacje o nauczycielach akademickich

Osoba odpowiedzialna za kartę

dr Barbara Borowik (kontakt: `bborowik@pk.edu.pl`)

Osoby prowadzące przedmiot

1 dr Barbara Borowik (kontakt: `bborowik@pk.edu.pl`)

13 Zatwierdzenie karty przedmiotu do realizacji

(miejscowość, data) (odpowiedzialny za przedmiot) (dziekan)

Przyjmuję do realizacji (data i podpisy osób prowadzących przedmiot)

. . . . . .