Politechnika Krakowska im. Tadeusza Kościuszki

Karta przedmiotu

obowiązuje studentów rozpoczynających studia w roku akademickim 2023/2024

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 23/24
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 nowymi, 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.

Cel 3 Celem będzie nabycie umiejętności zespołowej (parami) realizacji projektów, pomagającej w utrwaleniu praktycznych umiejętności, aby na podstawie zdobytej wiedzy, w sposób staranny i terminowy realizować te projekty, być gotowym do rozwiązywania problemów ze wspomnianego zakresu wiedzy, oraz utrwalić umiejętność pisania odpowiedniej dokumentacji zrealizowanego projektu, a także zwiększenia swych umiejętności klarownego przedstawienia i wytłumaczenia swoich ukończonych projektów wobec całej klasy, wraz z odpowiadaniem na zadawane im przez pozostałe grupy pytań dotyczących realizacji tych projektów.

4 Wymagania wstępne w zakresie wiedzy, umiejętności i innych kompetencji

1 Podstawowa wiedza ze statystyki i z rachunku prawdopodobieństwa.

2 Podstawowa wiedza z algebry liniowej.

5 Efekty kształcenia

EK1 Wiedza Student będzie potrafił wytłumaczyć działanie oraz zinterpretować wyniki ważniejszych algorytmów stosowanych przy przetwarzaniu, analizie i wizualizacji danych.

EK2 Umiejętności Student utrwali swe umiejętności 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 utrwali swe umiejętności samodzielnego programowania w języku R w środowisku RStudio oraz łączenia się internetowo poprzez RStudio zarówno z uniwersalnym silnikiem dla Big data (tj. ze SPARKiem), jak i z darmowymi serwerami RStudio, wspomagającymi przetwarzanie dużych zbiorów danych w Chmurze.

EK4 Kompetencje społeczne Student utrwali swe umiejętności wzajemnej współpracy w małych zespołach, w samokształceniu, oraz we właściwej komunikacji z innymi grupami i z nauczycielem.

6 Treści programowe

Projekt		
Lp	Tematyka zajęć Opis szczegółowy bloków tematycznych	Liczba godzin
P1	Treści programowe 1. Projekt pierwszy. Dokonanie analizy porównawczej wybranych (dla każdego zespołu innych) modeli predykcyjnych dostępnych w pakietach tidymodels oraz parsnip, w języku R. Każdy zespół będzie miał inny zbiór danych. Analiza porównawcza będzie opierać się o kluczowe czynniki, takie jak: wskazania wybranych metryk oceny modelu, np.: RMSE, MAPE, MASE, oraz MAE, najczęściej dążąc do minimalizacji wartości każdej z nich. Dane prognozowane będą porównywane z danymi rzeczywistymi, dostępnymi w zbiorze testującym. Zespoły przygotują sprawozdania z projektu nr 1 oraz przedstawią je wszystkim pozostałym grupom.	8
P2	Treści programowe 2. Projekt drugi. Rozbudowa poprzednio wykonanej analizy o zrealizowanie w niej kroku dostrajania (tuning) posiadanych przez dany model hiperparametrów, oraz porównanie otrzymanych nowych wartości czterech metryk. Zespoły przygotują sprawozdania z projektu nr 2 oraz przedstawią je wszystkim pozostałym grupom.	7

Projekt		
Lp	Tematyka zajęć Opis szczegółowy bloków tematycznych	Liczba godzin
P3	Treści programowe 3. Projekt trzeci. Rozbudowa poprzednio wykonanych analiz o zrealizowanie w swej analizie kroku walidacji krzyżowej i wybranie modelu najbardziej dostrojonego o najbardziej optymalnych nowych wartościach czterech metryk. Zespoły przygotują sprawozdania z projektu nr 3 oraz przedstawią je wszystkim pozostałym grupom.	7
P4	Treści programowe 4. Projekt czwarty. Modyfikacja poprzednio wykonanych kroków i zastąpienie ich funkcjami automatyzującymi cały wspomniany w poprzednich trzech analizach proces generacji modeli, ich dostrajania i walidacji krzyżowej, wykorzystując w tym celu dostępne w rodzinie pakietów tidymodels funkcje scalające poszczególne kroki wspomnianych trzech analiz. Zespoły przygotują sprawozdania z projektu nr 4 oraz przedstawią je wszystkim pozostałym grupom.	8

Wykład		
Lp	Tematyka zajęć Opis szczegółowy bloków tematycznych	Liczba godzin
W1	Poznanie nowych elementów architektury i nowych funkcjonalności, dodanych do ekosystemu Apache SPARK i do jego podsystemów. Poznanie nowych funkcji dodanych do pakietu sparklyr (autorstwa twórców środowiska RStudio), pozwalającego na łączenie się ze SPARKiem oraz korzystanie z pakietów SPARKa w środowisku skalowalnym i rozproszonym, będąc w środowisku programistycznym RStudio.	4
W2	Treści programowe 2. Poznanie nowych rozproszonych algorytmów dodanych do bogatego asortymentu biblioteki spark.ml w SPARKu. Poznanie głównych możliwości i funkcjonalności i sposobu wykorzystania tych algorytmów w chmurze oraz możliwości współpracy tych algorytmów ze środowiskiem RStudio.	4
W3	Treści programowe 3. Poznanie nowych funkcjonalności rozszerzonego pakietu rsparkling, dostarczającego możliwości połączenia się z dużą liczbą najnowszych rozproszonych algorytmów uczenia maszynowego w środowisku H20, mocno współpracującym ze SPARKiem. Z platformy H20 korzysta ponad 8000 największych firm na świecie. Wszystko to jest możliwe do sterowania, będąc w środowisku programistycznym RStudio.	8
W4	Treści programowe 4. Poznanie nowych pakietów służących do modelowania tekstu, eksploracji sieci społecznościowych (social networks) oraz nowych metod tworzenia dashboardów, współpracujących z rodziną pakietów tidyverse i tidymodels oraz zaznajomienie się z przykładami ich zastosowania.	4

Wykład		
Lp	Tematyka zajęć Opis szczegółowy bloków tematycznych	Liczba godzin
W5	Treści programowe 8. Poznanie nowych funkcjonalności w drugiej co do ważnosci (obok rodziny tidyverse) rodziny pakietów uczenia maszynowego "Tidymodels", której celem jest uproszczenie i zautomatyzowanie przetwarzania potokowego zadań w uczeniu maszynowym, oraz pomoc w zrównolegleniu całego ciągu procesów, w skład których wchodzi walidacja krzyżowa, wstępne przetwarzanie danych (dzięki pakietowi recipe, należącemu do tej rodziny) oraz dostrajanie (tuning) algorytmów uczenia maszynowego, polegające na przetwarzaniu olbrzymiej ilości wygenerowanych modeli dla różnych kombinacji wartości hiperparametrów tych algorytmów. Rodzina pakietów "Tidymodels" pozwala na wybranie pewnej liczby miar, służących do badania jakości wybranych modeli oraz na zautomatyzowanie procesu znalezienia najlepszego modelu o najlepszej jakości, gdy otrzymanie tego wyniku wymaga generacji olbrzymiej ilości modeli, a każdy z nich musi podlegać większości procesom wspomnianego przetwarzania potokowego.	10

7 Narzędzia dydaktyczne

N1 Wykłady (w przypadku realizacji zajęć w trybie zdalnym z wykorzystaniem stosownych narzędzi teleinformatycznych, np. MS TEAMS)

N2 Praca w 2-3 osobowych grupkach (w przypadku realizacji zajęć w trybie zdalnym z wykorzystaniem stosownych narzędzi teleinformatycznych, np. MS TEAMS)

N3 Prezentacje multimedialne (w przypadku realizacji zajęć w trybie zdalnym z wykorzystaniem stosownych narzędzi teleinformatycznych, np. MS TEAMS)

N4 Konsultacje (w przypadku realizacji zajęć w trybie zdalnym z wykorzystaniem stosownych narzędzi teleinformatycznych, np. MS TEAMS)

N5 Dyskusja (w przypadku realizacji zajęć w trybie zdalnym z wykorzystaniem stosownych narzędzi teleinformatycznych, np. MS TEAMS)

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	15
Egzaminy i zaliczenia w sesji	5
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	20
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 Wysłanie każdego z czterech sprawozdań, dotyczących czterech projektów do prowadzącego zajęcia projektowe. .

F2 Prezentowanie każdego z czterech projektów przed pozostałymi zespołami w klasie.

Ocena podsumowująca

P1 Średnia ważona ocen formujących

Warunki zaliczenia przedmiotu

W1 Uzyskanie odpowiedniej liczby punktów z czterech sprawozdań oraz z czterech prezentacji.

W2 Spełnienie warunku obecności na obowiązkowych formach zajęć (dopuszczalna jedna nieobecność nieusprawiedliwiona na każdej z obowiązkowych form zajęć).

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 zagadnień w stopniu powyżej 50%.
Na ocenę 3.5	Opanowanie zagadnień w stopniu powyżej 60%.
Na ocenę 4.0	Opanowanie zagadnień w stopniu powyżej 70%.
Na ocenę 4.5	Opanowanie zagadnień w stopniu powyżej 80%.
Na ocenę 5.0	Opanowanie zagadnień 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 zagadnień w stopniu powyżej 50%.
Na ocenę 3.5	Opanowanie zagadnień w stopniu powyżej 60%.
Na ocenę 4.0	Opanowanie zagadnień w stopniu powyżej 70%.
Na ocenę 4.5	Opanowanie zagadnień w stopniu powyżej 80%.
Na ocenę 5.0	Opanowanie zagadnień 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 zagadnień w stopniu powyżej 50%.
Na ocenę 3.5	Opanowanie zagadnień w stopniu powyżej 60%.
Na ocenę 4.0	Opanowanie zagadnień w stopniu powyżej 70%.
Na ocenę 4.5	Opanowanie zagadnień w stopniu powyżej 80%..
Na ocenę 5.0	Opanowanie zagadnień 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 zagadnień w stopniu powyżej 50%.
Na ocenę 3.5	Opanowanie zagadnień w stopniu powyżej 60%.
Na ocenę 4.0	Opanowanie zagadnień w stopniu powyżej 70%.
Na ocenę 4.5	Opanowanie zagadnień w stopniu powyżej 80%.
Na ocenę 5.0	Opanowanie zagadnień 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	P1 P2 P3 P4 W1 W2 W3 W4 W5	N1 N2 N3 N4 N5	F1 F2 P1
EK2	K_U02 K_U10 K_U11 K_U12 K_U13 K_U16 K_U20 K_U22	Cel 2	P1 P2 P3 P4 W1 W2 W3 W4 W5	N1 N2 N3 N4 N5	F1 F2 P1
EK3	K_U02 K_U10 K_U11 K_U12 K_U13 K_U15 K_U16 K_U20 K_U21 K_U22	Cel 2	P1 P2 P3 P4 W1 W2 W3 W4 W5	N1 N2 N3 N4 N5	F1 F2 P1
EK4	K_K01 K_K02 K_K03 K_K04 K_K05 K_K06 K_K07	Cel 3	P1 P2 P3 P4	N1 N2 N3 N4 N5	F1 F2 P1

11 Wykaz literatury

Literatura podstawowa

[1] Podstawowa dokumentacja wraz z bardzo licznymi przykładami: https://spark.rstudio.com/guides/ ; Liczne przykłady: https://stat545.com/index.html ; https://ggplot2-book.org/ ; https://spark.apache.org/docs/latest/ml-guide.html ; Tysiące przykładowych notebooków: https://rpubs.com ; https://www.kaggle.com/code?language=R ; https://www.kaggle.com/

[2] Rozszerzenie książki ISLR (T. Hastie, R. Tibshirani) o rodzinę pakietów Tidymodels: https://emilhvitfeldt.github.io/ISLR-tidymodels-labs/index.html ; Książki online: https://bookdown.org/ ; https://www.tmwr.org/ ; https://smltar.com/ ; https://www.tidytextmining.com/ ; https://therinspark.com/ ; http://r4ds.had.co.nz/

[3] Darmowy serwer online z zainstalowanym RStudio i z pakietami: https://rdrr.io/snippets/ ; Środowisko darmowe w Chmurze: https://rstudio.cloud/ ; https://databricks.com/try

[4] Dwie najważniejsze rodziny pakietów w RStudio: tidyverse i Tidymodels; tutoriały: https://rpubs.com/cliex159/885971 ; https://jhudatascience.org/tidyversecourse/model.html ; https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/beginners-tutorial-on-xgboost-parameter-tuning-r/tutorial/ ; https://www.startupe

/https://www.startupengineer.io/_repos/_transfer/data_science/ ; https://jhudatascience.org/tidyversecourse/model.ht ; https://juliasilge.com/blog/xgboost-tune-volleyball/ ; https://www.alexbaecher.com/post/ml-tidymodels/ ; https://cran.rstudio.com/web/packages/workboots/vignettes/Getting-Started-with-workboots.html ; https://tanthiamhu ecosystem/ ; https://scienceofdata.org/2021/07/07/look-what-the-cat-dragged-in-catboost-with-tidymodels/ ; https://dnield.com/posts/tidymodels-intro/ ; https://www.r-bloggers.com/2020/05/using-xgboost-with-tidymodels/

Literatura uzupełniająca

[1] Dokumentacja SPARK'a: http://spark.apache.org/docs/latest/ ; Uczenie maszynowe https://spark.apache.org/docs/latest/ guide.html ; https://docs.databricks.com/ ; https://sparkhub.databricks.com/resources/

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

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: `barbara.borowik@pk.edu.pl`)

13 Zatwierdzenie karty przedmiotu do realizacji

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

Przyjmuję do realizacji (data i podpisy osób prowadzących przedmiot)

. . . . . .