Podstawy reprezentacji i analizy danych
rok akad. 2018/19

prowadzący:
dr inż. Grzegorz Sarwas sarwasg@ee.pw.edu.pl

Zadanie projektowe

Celem zadania jest wykonanie analizy danych w celu rozwiązania problemu postawionego w jednym z 15 niżej podanych tematów. Zadania te dotyczą problemu klasyfikacji. Każdy zbiór danych ma w opisie postawiony dla niego problem/problemy, jednakże, jeśli któryś zespół jest w stanie zaproponować inny problem do rozwiązania/udowodnienia przy pomocy otrzymanych danych to droga jest wolna.

Rozwiązanie otrzymanego zadania należy wykonać w języku **Python** lub **R** wykorzystując metody i narzędzia analizy, wizualizacji, grupowania oraz klasyfikacji danych.

Rozwiązując postawione problemy należy przede wszystkim skupić się na danych wykonując poszczególne kroki:

1. Opisać postawiony problem.
2. Określić liczbę obiektów, liczbę klas, zakresy zmienności poszczególnych atrybutów, ich wartości statystycznych, poziom wypełnienia kolumn, ilość unikalnych danych itp.
3. Przeanalizować korelację między zmiennymi.
4. Przygotować dane do analizy: Imputować brakujące dane lub usunąć rzadko wypełnione kolumny.
5. Przeanalizować podobieństwo między danymi przy pomocy poznanych algorytmów grupowania wraz z analizą ilości grup.
6. Należy przetestować wybrane klasyfikatory pod kątem doboru ich parametrów.
7. Proszę ocenić czy do poprawnej klasyfikacji należy wykorzystać wszystkie atrybuty, czy wystarczy ich podzbiór, a może należy stworzyć jakieś nowe dane w oparciu o istniejące?

Projekt wykonujemy w zespołach **dwuosobowych**. Każde dane można analizować na wiele sposobów, więc proponuję podzielić się pracą, a później zebrać do raportu końcowego wszystkie wyniki, komentarze oraz zrozumiale opisany sposób analizy. Można w oparciu o wcześniej wyuczone klasyfikatory przez poszczególne osoby wykonać próbę złożenia klasyfikatorów (ensembling).

Raport ma zostać dostarczony w pliku **tekstowym** lub **ipython notebook**. Raport posiadający w sobie skrypt należy wgrać na iSOD i następnie przyjść na jego obronę.

Za projekt można otrzymać do **30 pkt**.

Tematy projektów:

1. Heart Disease UCI
   https://archive.ics.uci.edu/ml/datasets/Heart+Disease
2. Credit Card Fraud Detection
   Anonymized credit card transactions labeled as fraudulent or genuine
   https://www.kaggle.com/mlg-ulb/creditcardfraud
3. Classify gestures by reading muscle activity.
   a recording of human hand muscle activity producing four different hand gestures
   https://www.kaggle.com/kyr7plus/emg-4
4. Loan Default Prediction
   Predicting Whether a Customer can pay their first EMI
   https://www.kaggle.com/roshansharma/loan-default-prediction
5. Bank_Loan_Classification
   Universal bank data for classification
   https://www.kaggle.com/sriharipramod/bank-loan-classification
6. Rain in Australia
   Predict rain tomorrow in Australia
   https://www.kaggle.com/jsphyg/weather-dataset-rattle-package
7. Rozpoznawanie nieprawidłowości w kręgosłupie.
8. Rozpoznawanie końcowej oceny studenta (przy pomocy klasyfikacji) zbiór A
9. Rozpoznawanie końcowej oceny studenta (przy pomocy klasyfikacji) zbiór B.
10. Rozpoznawanie typu Pokemona po jego cechach.
11. Określenie czy dana osoba zarabia więcej niż 50tyś dolarów rocznie.
12. Rozpoznawanie płci właściciela profilu na tweeterze.
13. Rozpoznawanie jednego z sześciu stanów aktywności przy pomocy czujników ze smartfonów.
14. Rozpoznawanie kategorii artykułu po tytule i wydawcy.
15. Rozpoznawanie płci na podstawie głosu.