Podstawy reprezentacji i analizy danych – projekt

sem.letni 18/19

grupa poniedziałkowa 10-12 (prowadzący M.Iwanowski)

Projekt wykonujemy w zespołach **dwuosobowych**. Każdy zespół wybiera zbiór danych spośród podanych przez prowadzącego.

Rozwiązanie otrzymanego zadania należy wykonać w języku **Python** (alternatywnie może być **R** lub **Matlab)** wykorzystując metody i narzędzia analizy, wizualizacji, grupowania oraz klasyfikacji danych. Rozwiązując postawione problemy należy przede wszystkim skupić się na danych wykonując następujące kroki (w zależności od zadania podejmowane kroki mogą się nieco różnić od podanych poniżej):

1. Opisać postawiony problem.
2. Zbadać czy i jakie problemy zostały już rozwiązane dla wybranego zbioru
3. Określić liczbę obiektów, zakresy zmienności poszczególnych atrybutów, ich wartości statystycznych, poziom wypełnienia kolumn, ilość unikalnych danych, atrybut decyzyjny/możliwe atrybuty decyzyjne, liczbę klas itp.
4. Przygotować dane do analizy – jeśli zachodzi, rozwiązać problem brakujących danych
5. Przeanalizować zależności między między atrybutami/klasami przy pomocy poznanych metod (korelacja, grupowanie).
6. Dobrać optymalny klasyfikator do postawionego zadania
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?

Raport powinien zostać przygotowany w pliku **jupyter notebook** w terminie najpóźniej do końca semestru tj. do 12.06.2019. Skrypt powinien zawierać obszerny opis zadania, kolejno wykonywanych kroków, wyników, wszelkie niezbędne komentarze oraz stosowne wstawki kodu. Plik należy wgrać na iSOD, a następnie zapisać się na najbliższy termin konsultacji projektowych (terminy i zapisy w ISOD), a następnie przyjść na te konsultacje na jego obronę.

Za projekt można otrzymać do **30 pkt**. Liczba punktów może być zwiększona o kolejnych 20 po rozszerzeniu zakresu zadania. Rozszerzenie zakresu można zrealizować w następujący sposób:

- zespół przygotowuje rozwiązanie zadania w wariancie podstawowym i zalicza je przed końcem semestru.

- w trakcie obrony ustala z prowadzącym zakres rozszerzenia zadania

- w terminie najpóźniej do końca sesji letniej (28.06.2019) przygotowuje raport z zakresu rozszerzonego, a następnie omawia go z prowadzącym na zaliczeniu/obronie.

Zbiory danych (jeden zestaw na nie więcej niż 3 zespoły !):

1. piłkarze (https://www.kaggle.com/karangadiya/fifa19)
2. owoce (https://www.kaggle.com/moltean/fruits)
3. korzystanie z telefonu (https://www.kaggle.com/mboaglio/simplifiedhuarus)
4. karty kredytowe (https://www.kaggle.com/mlg-ulb/creditcardfraud)
5. cztery kształty (https://www.kaggle.com/smeschke/four-shapes)
6. nieprawidłowości w kręgosłupie (www.ee.pw.edu.pl/~sarwasg/ED/Projekt_10.zip)
7. oceny studentów (www.ee.pw.edu.pl/~sarwasg/ED/Projekt_12.zip)
8. rozpoznawanie płci właściciela profilu na Tweeterze (www.ee.pw.edu.pl/~sarwasg/ED/Projekt_5.zip)
9. choroby serca (https://archive.ics.uci.edu/ml/datasets/Heart+Disease)
10. dowolny inny zbiór danych (np. z serwisu Kaggle o więcej niż 200 obiektach, na którym można zrealizować zadania analizy eksploracyjnej, wizualizacji, (opcjonalnie) grupowania i (obowiązkowo) klasyfikacji – po akceptacji przez prowadzącego