Metody Eksploracji Danych – projekt

sem.zimowy 2021/2022

Dr inż. Grzegorz Sarwas

Celem projektu jest opracowanie modelu regresji pokazującego trend w wybranym przez studentów zbiorze danych, wraz z jego analizą. Do zdobycia jest **50 pkt**.

Projekt składa się z 3 części:

1. **Termin - 15 grudnia 2021r. Ilość punktów do zdobycia – 5 pkt.**
   Skompletowanie zespołu projektowego, wybór zbioru danych i zagadnienia.

2. **Termin – 3 stycznia 2022r. Ilość punktów do zdobycia – 20 pkt.**
   Analiza eksploracyjna posiadanego zbioru/wycinka zbioru danych i postawienie tezy/zadania badawczego mającego na celu opracowanie modelu regresji dla opisanych danych.

3. **Termin - 24 stycznia 2022r. Ilość punktów do zdobycia – 25 pkt.**
   Opracowanie modelu regresji i jego prezentacja.

Zasady realizacji projektu i sposób oceniania:

1. Projekt realizujemy w zespołach 2 osobowych (jeśli będzie wymagane stworzenie zespołu 3 osobowego, to jego temat i zwiększony zakres musi być zatwierdzony przez prowadzącego). Każdy zespół ma za zadanie poszukać ogólnodostępnego zbioru danych:
   a. Dominic's Dataset: https://www.chicagobooth.edu/research/kilts/datasets/dominicks
   b. Dane GUSu w Polsce, jak i za granicą: https://stat.gov.pl/podstawowe-dane/
   c. Dane Eurostatu
   d. Dane giełdowe, medyczne, astronomiczne itp.

   Można także poszukać dowolnego innego, sensownego zbioru danych w Kaggle Datasets, jak również wśród różnych danych prezentowanych przez firmy rządowe lub pożytku publicznego. Zależy nam na surowych danych bez postawionego problemu, po to by postawienie jakiegoś zagadnienia wynikało z przeprowadzonej analizy danych.
2. Celem analizy eksploracyjnej danych jest sprawdzenie zależności między posiadanymi danymi (5 pkt.), zbadanie ich zakresów i stopnia zmienności (5 pkt.), analiza stopnia wypełnienia danych (5 pkt.), wizualizacja (5 pkt.). Wynikiem tych analiz, ma być raport zakończony postawieniem hipotezy badawczej mającej na celu znalezienie relacji między zmiennymi objaśniającymi, a zmienną objaśnianą.
3. Ostatnia część projektu związana jest z opracowaniem modelu regresji (5 pkt.), poprzez zastosowanie poznanych na wykładzie i ćwiczeniach laboratoryjnych metodach doboru cech (5 pkt.) i regularyzacji (5 pkt.) oraz opracowanie raportu opisującego przeprowadzone prace, jak i wyciągnięte wnioski z otrzymanych wyników/zależności (5 pkt.). Z przeprowadzonych prac projektowych opracowana ma zostać prezentacja (5 pkt).

Uwaga: przygotowanie raportu i przedstawienie prezentacji jest **warunkiem** zaliczenia tj. bez przedstawienia tych dwóch elementów zaliczenie nie będzie możliwe. Podane punkty w obu tych przypadkach można zdobyć za staranność wykonania, jasność i czytelność wypowiedzi (pisemnej/ustnej), sposób prezentacji problemu i rozwiązania, sensowność sformułowanych na zakończenie wniosków i innych czynników branych zwykle pod uwagę przy ocenianiu tego typu aktywności.

UWAGA: **Brak realizacji powyższych kamieni milowych w podanych terminach powodują utratę 4 pkt. za każdy dzień zwłoki.**