ZMUM, 2019 **Projekt 2**

1. Celem projektu jest zbadanie metod selekcji zmiennych.
2. Należy zaproponować metody selekcji zmiennych oraz klasyfikacji które umożliwiają zbudowanie modelu o dużej mocy predykcyjnej przy użyciu możliwie małej liczby zmiennych.
3. Zbiór danych:
   - Zbiór *artificial* to sztuczny zbiór w którym są ukryte istotne zmienne (pliki: artificial_train.data, artificial_train.labels, artificial_valid.data).
4. Mamy 3 pliki: dane treningowe, etykiety dla danych treningowych, dane walidacyjne. Tabela 1 zawiera informacje o zbiorze.

Dane	Liczba zmiennych	Liczba obserwacji (treningowy)	Liczba obserwacji (walidacyjny)
artificial	500	2000	600

Tabela 1: Charakterystyka zbioru danych.

5. Dane treningowe służą do budowy modelu oraz selekcji zmiennych. Należy dokonać prognozy dla danych walidacyjnych, przypisując każdej obserwacji oszacowane prawdopodobieństwo aposteriori dla klasy "1", t.j. $P(y = 1|x_1, \ldots, x_p)$.
6. Wyniki należy zapisać do plików:
   - *KOD_artificial_prediction.txt*, pstwa dla zbioru walidacyjnego dla danych *artificial*.
   - *KOD_artificial_features.txt*, wybrane zmienne dla danych *artificial*.

   KOD oznacza kod studenta (3 pierwsze litery imienia+ 3 pierwsze litery nazwiska), np. dla Jana Kowalskiego będzie JANKOW. W pierwszej linijce każdego pliku powinien być kod studenta, w kolejnych prawdopodobieństwa lub numery wybranych zmiennych. Przykładowe pliki wynikowe: `JANKOW_artificial_prediction.txt` oraz `JANKOW_artificial_features.txt`.
7. Dane potrzebne do wykonania projektu znajdują się na stronie `https://home.ipipan.waw.pl/p.teisseyre/TEACHING/ZMUM/index.html`.
8. Projekty są wykonywane w zespołach 1 osobowych.
9. Należy przetestować co najmniej 2 metody selekcji zmiennych.
10. Ocena na podstawie:
    - jakości klasyfikacji mierzonej jako balanced accuracy ($BA = \frac{1}{2}(\frac{TP}{P} + \frac{TN}{N})$) oraz liczby zmiennych. Ranking będzie utworzony na podstawie BA (im większe BA tym lepiej). W przypadku nieistotnych różnic w BA o kolejności w rankingu będzie decydować liczba wybranych zmiennych (im mniej tym lepiej) (50 %).
    - prezentacji (5 minut) podsumowującej wyniki (25 %),
    - raportu (maksymalnie 3 strony a4) który zawiera: podsumowanie eksperymentów, uzasadnienie wyboru końcowej metody, opis przetwarzania danych (25 %).
11. Prezentacje odbędą się:
    - 11 czerwca (wtorek) na zajęciach projektowych,

12. Plik z wynikami, prezentacje, raport i kody programu należy wysłać na adres: teisseyrep(at)ipipan.waw.pl. Ostateczny termin wysyłania:

    - 7 czerwca (piątek), godzina 23:59:59.

    Wysłanie po terminie będzie skutkować naliczeniem karnych punktów.