LABORATORIUM 4

Braki w danych

Zestawy danych

https://data.gov/ to portal otwartych danych rządu Stanów Zjednoczonych.

https://www.ncdc.noaa.gov/cag/ - zapewnia portal klimatyczny NOAA

https://www.esrl.noaa.gov/psd/data/timeseries/ - Portal badań Ziemi administracja laboratoriów (ESRL, Earth System Research Laboratory) NOAA dostarcza miesięczne i sezonowe z danymi klimatycznymi.

https://www.quandl.com/search - Quandl udostępnia setki bezpłatnych szeregów czasowych z danymi finansowymi, a także zestawy danych z płatnym dostępem.

https://datamarket.com/data/list/?q=provider:tsdl - biblioteka danych TSDL (biblioteka danych szeregów czasowych) zawiera linki do setek tymczasowych zbiorów danych rzędy w wielu obszarach przemysłowych.

http://archive.ics.uci.edu/ml/datasets.html - repozytorium uczenia maszynowego Uniwersytetu Kalifornijskiego w Irvine (UCI) zawiera dziesiątki zestawów danych szeregów czasowych z różnych obszary.

ZADANIE

1. Należy napisać program (za pomocą biblioteki **scikit-learn**), który usuwa **Braki w danych**.
   Można postępować na kilka sposobów:
   - usuwać wiersze z brakującymi danymi,
   - usuwać kolumny z brakującymi danymi,
   - dokonać imputacji brakujących danych,
   - zastępować brakujące wartości średnią wyliczoną na podstawie całej kolumny cech

   Sprawdzić, jak wpłynęli powyższe sposoby na wyniki algorytmów klasyfikacji (otrzymany w laboratorium 3):

- regresji logistycznej
- maszyny wektorów nośnych
- jądra SVM
- Uczenia drzew decyzyjnych
- Algorytm k-najbliższych sąsiadów

Pracujemy z zestawom danych, który Państwo wybrali na laboratorium 3 i sprawdzamy na nim algorytmy klasyfikacji z laboratorium 3.

Na ocenę 3 proszę o usunięcie braków danych (za pomocą różnych sposobów) i zaprogramowanie dowolnego (jednego!) algorytmu,

na 4 proszę o usunięcie braków danych (za pomocą różnych sposobów) i zaprogramowanie dowolne dwa algorytmy,

na 5 proszę o usunięcie braków danych(za pomocą różnych sposobów) i zaprogramowanie dowolnych trzech algorytmów

Zrób podsumowanie: jak zmienili się wyniki w klasyfikacji, dlaczego