# Raport EWD

Znajdą Państwo dane GUS-owskie:

* ludnosc.csv: wyniki ankiet w gospodarstwach domowych, zagregowane po województwach
* gminy.csv: dane statystyczne na temat gmin

Na ich podstawie proszę o przygotowanie w grupach projektu, w którym wykorzystają Państwo co najmniej 4 techniki poznane na wykładzie, w tym:

- analizę statystyczną rozważanego parametru w wybranych populacjach

(tzn. przedział ufności uzyskany z rozkładu Studenta) - np. średnich

dochodów na głowę w gospodarstwach domowych w poszczególnych

województwach w 2020

- zastosowanie któregoś z algorytmów klasyfikacyjnych

Projekt powinien zawierać też raport z analizą i wizualizacją wszystkich

wyników.

W przypadku problemów klasyfikacyjnych, ze względu na zastosowaną już

wcześniej agregację, dużo naturalnych pytań jest niesprawdzalnych.

Jednak dalej pozostaje wiele takich, na które dane mogą odpowiedzieć, np.:

- Można rozważyć klasyfikację gmin wg tego, czy bezrobocie kobiet jest

niższe niż mężczyzn

- Ciekawe może być obcięte drzewo decyzyjne pozwalające wykrywać gminy

rekordzistki (zdefiniowane np. jako górne 10%) względem jakiejś cechy

- Dane można też pokubełkować. Tutaj jednak gorszym pomysłem byłoby

klasyfikowanie w dwóch kubełkach ("powyżej czy poniżej mediany?"),

ponieważ jeśli zmienna ma rozkład normalny, to większość wartości będzie

bliska mediany, zatem klasyfikator musiałby mocno reagować na subtelne

różnice.

Chciałabym, żeby Państwa projekt odpowiadał na rozsądnie brzmiące

pytania. Oznacza to, że zastosowanie modelu regresji liniowej dla

średniej ceny jajek w zależności od średniej ceny garniturów w danym

województwie, jakkolwiek teoretycznie poprawne, nie jest mile widziane.

Podobnie klasyfikacja miasto/wieś (np. wnioskowane z kodu TERYT) na

podstawie liczby mieszkańców jest zbyt oczywista.

Różne techniki powinny zasadniczo odpowiadać na różne pytania; w

przeciwnym razie - jeśli 2 lub więcej techniki będą analizować to samo

pytanie - proszę porównać pomiędzy sobą wyniki otrzymane w każdym z podejść.

--- Uwagi techniczne ---

Dane pochodzą wprost z GUS i nie są oczyszczone. Oznacza to, że wartości

niektórych zmiennych są znane tylko dla niektórych lat i/lub jednostek

terytorialnych. W bazie występują NULLe, wartości '0' w przypadku 'braku

danych', itp.

Jeśli Państwa pomysł na projekt wymaga ściągnięcia dodatkowych

GUS-owskich tabel, można oczywiście to zrobić.

Projekt można napisać w Pythonie, R, C, C++, Javie, natomiast \*nie

można\* w Excelu ani SQL-u.

Proszę o przysyłanie projektów mailem do prowadzącego ćwiczenia.

TERMIN: 15 stycznia.

Proszę o kontakt w przypadku, gdyby chcieli Państwo zrobić coś

niestandardowo, np.:

- pracować na innych zbiorach danych

- użyć języka spoza powyższej listy

- pracować w grupie mniej niż 3 lub ponad 4 osób

- wykonać eksplorację danych na sposób, który Państwa fascynuje, ale nie

mieści się w powyższej treści