MAGISTERKA

Zaproponować problem z danymi (mam puste miejsca, odstające przypadki)
Znalazłem niespójne formatowanie np.(polska, Polska, spacja na końcu stringi, albo zły format daty.)
Cena napisana jakoś string(100zł) zamiast int, albo kod pocztowy zamiast string to int.
Kiedy nazwa kolumny jest w komórce. (Np. familyymember1age, familymember2age itd.) należy użyć stack, melt, wide to long(funkcje pandas)
Powtórzenia, np raz jest Jan Kowalski, a potem J Kowalski.

Mam 3 na kartce z tematem.

Walidacja potrzebna bo eksperyment: train and test ponad 1000
Cross validation 500-1000
Leave one out 100

Napisać do Bazana o prace odnośnie odstających miejsc oraz klasa do testowania kompatybilna z scikitleasn do walidacji. (Napisane)

Przykłady modeli uczenia dla danych tabelarycznych:
-Random forest
-xg boost

Random forest na początek. Na przykładowym zbiorze danych.

Choroba naczyń - dane najistotniejsze

Znasz dobre biblioteki do czyszczenia danych.
2 metody normalizacji.

3 pary, 3 pojedyncze sposoby, i 3 na raz sposoby czyszczenia.