

# Praca domowa nr 1

## Optymalizacja hiperparametrów wybranych modeli.

Antoni Chudy, Maciej Malewicz

29.11.2023

## Rozważane algorytmy:

- Regresja logistyczna,
- drzewo losowe,
- las losowy.

## Kształty zbiorów danych:

- 1000 wierszy, 20 kolumn,
- 5404 wierszy, 5 kolumn,
- 4521 wierszy, 16 kolumn,
- 10885 wierszy, 21 kolumn,

dla których zmienna objaśniana jest binarna.

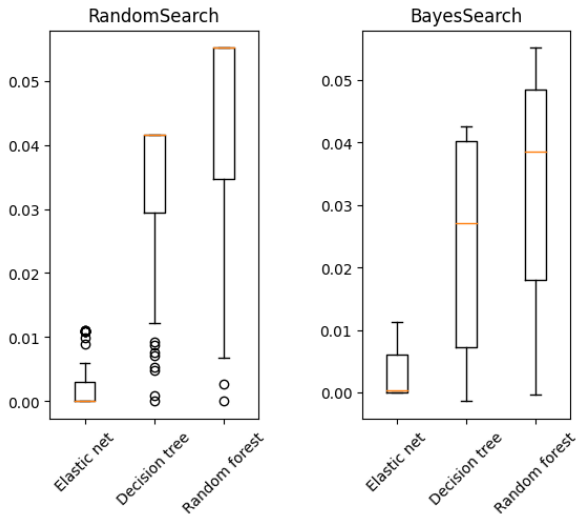
Użyte metody:

- przeszukiwanie za pomocą **RandomizedSearchCV**,
- przeszukiwanie za pomocą **BayesSearchCV**.

Każda metoda wykonuje na każdym zbiorze po 50 iteracji.

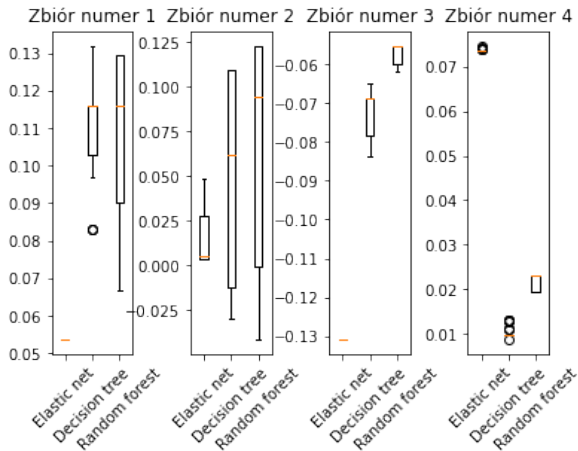
## Uzyskane wyniki

# Uzyskane wyniki - tunowalność



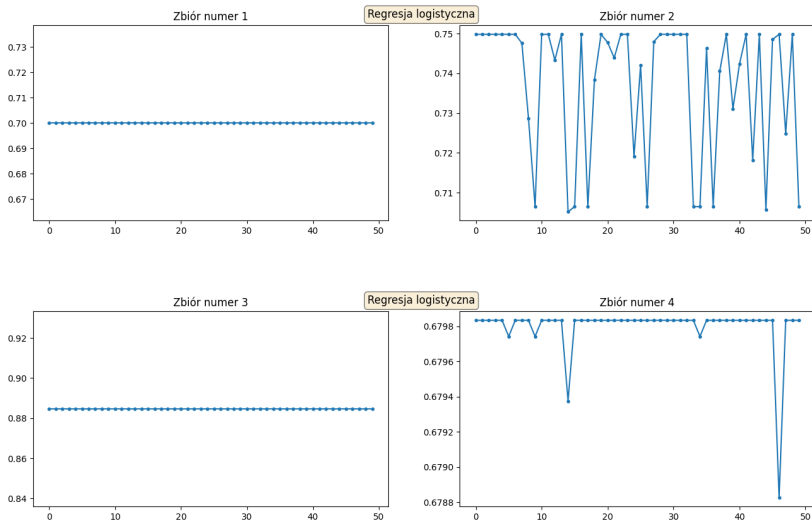
Rysunek: Tunowalność dla poszczególnych metod.

# Uzyskane wyniki - różnorodność

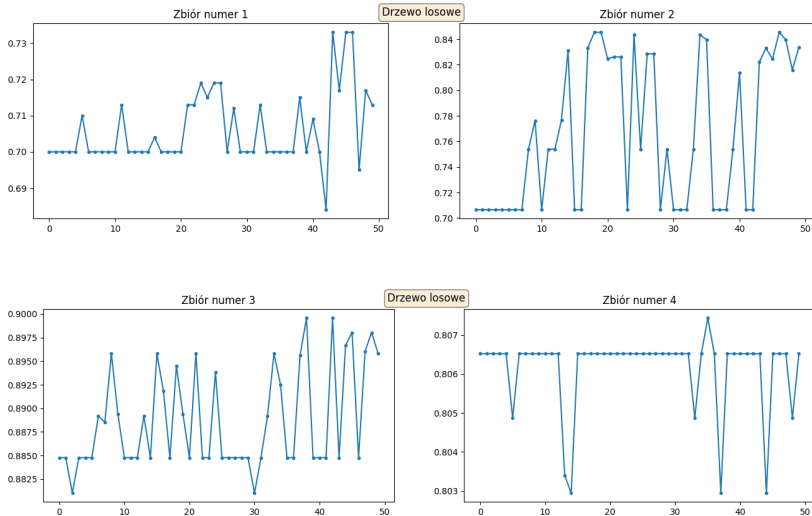


Rysunek: Różnorodność wyników dla każdego zbioru.

# Uzyskane wyniki - zbieżność, regresja logistyczna



# Uzyskane wyniki - zbieżność, regresja logistyczna

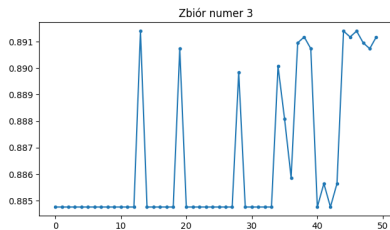
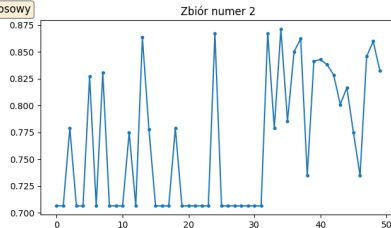




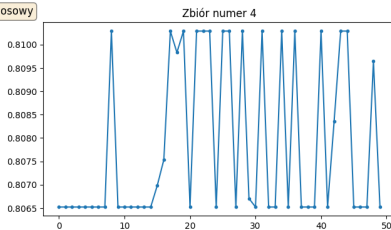
# Uzyskane wyniki - zbieżność, regresja logistyczna



Las losowy



Las losowy



## Testy statystyczne i podsumowanie

$X$  – średnie wyniki uzyskane za pomocą RandomizedGridSearch,  
 $Y$  – średnie wyniki uzyskane za pomocą BayesSearchCV. Do weryfikacji używamy testu Wilcoxon.

## Zadanie weryfikacji hipotez

$$\begin{cases} H_0 : F_X = F_Y, \\ H_1 : F_X \neq F_Y. \end{cases}$$

Dla poszczególnych modeli uzyskujemy następujące  $p$ -wartości:

- Regresja logistyczna: 0,345,
- drzewo decyzyjne: 0,005,
- las losowy: 0,010.

Dla drzew decyzyjnych i lasów losowych przy niewielkiej liczbie iteracji ciężko uzyskać stabilność wyników.

Wyniki dla regresji są mniej różnorodne.

Uzyskane wyniki mogłyby się różnić w zależności od zbiorów przyjętych do analizy.

Potrzebna większa moc obliczeniowa.

- [1] Philipp Probst, Anne-Laure Boulesteix, Bernd Bischl Tunability, Importance of Hyperparameters of Machine Learning Algorithms
- [2] <https://scikit-learn.org/stable/>
- [3] <https://scikit-optimize.github.io/stable/>
- [4] <https://scipy.org/>