Homework 1 raport

Marta Boratyn, Zuzanna Ostas

listopad 2023

Spis treści

1 Wstęp

Celem projektu jest analiza tunowalności wybranych algorytmów oraz ich hiperparametrów. W tym celu wybrane zostały cztery zbiory danych ze strony *OpenML* o kategorycznej zmiennej objaśnianej. Na tych zbiorach badano tunowalność trzech algorytmów klasyfikacyjnych: drzewa klasyfikacyjnego, lasu losowego oraz modelu SVM. Tunowalność badano przy dwóch metodach losowania punktów z przestrzeni hiperparametrów: korzystając z metody *RandomizedSearchCV* oraz z optymalizacji bayesowskiej (przy wykorzystaniu pakietu *SMAC3*). Dodatkowo dla lasu losowego zbadano tunowalność poszczególnych parametrów.

2 Opis eksperymentu

W celu zbadania tunowalności algorytmów w pierwszym kroku należało wybrać zestaw domyślnych hiperparametrów. Powinien być to taki wektor parametrów, który działa dobrze dla różnych zbiorów danych. Optymalny wektor θ^* znaleziono korzystając z metody *RandomizedSearchCV* z pakietu *sklearn*. Jego wybór polegał na przeprowadzeniu tuningu na każdym ze zbiorów danych, a następnie uśrednieniu wyników AUC z historii tuningu. Wówczas wektor θ^* jest tym, który pozwala na otrzymanie najlepszego średniego wyniku ze wszystkich zbiorów. Przy wyborze siatki parametrów dla każdego zbioru korzystano z tabeli 1. w artykule https://jmlr.org/papers/volume20/18-444/18-444.pdf. W celu otrzymania wiarygodnych wyników AUC, przy testowaniu parametrów w funkcji *RandomizedSearchCV* stosowano 5-krotną kroswalidację.

2.1 Wykorzystane dane

1. **Dane Adult** Dane dotyczą prognozowania czy dana osoba zarabia więcej niż 50 tysięcy rocznie. Składa się z 48842 rekordów opisanych przez 15 atrybutów, takich jak na przkład: wiek, kraj zamieszkania, wykonywany zawód, liczba godzin w pracy w tygodniu. Zmienna „Class" jest zmienną binarną określającą czy dana osoba zarabia więcej czy nie.

2. **Dane elevators** Dane dotyczą wind. Jest to zbinaryzowana wersja oryginalnego zestawu danych. 16599 rekordów scharakteryzowanych jest 19 cechami numerycznymi i zaklasykowanych do klasy P lub N.

3. **Dane eeg-eye-state** Dane pochodzą z jednego ciągłego pomiaru EEG przy użyciu zestawu słuchawkowego Emotiv EEG Neuroheadset. Składają się z 14980 rekordów opisanych 15 cechami numerycznymi. Zmienna "Class" jest zmienną binarną określającą czy oko było otwarte, czy zamknięty w trakcie badania.

4. **phoneme** Celem zbioru danych jest scharakteryzowanie samogłosek jako nosowe (klasa 0) lub ustne (klasa 1). W zbiorze znajduję się 5404 samogłosek opisanych 6 cechami numerycznymi.

Wszystkie zbiory danych dotyczą problemy klasyfikacji binarnej.

2.2 Optymalne hiperparametry

2.2.1 Drzewo decyzyjne

Wybrane optymalne parametry:

	Random	Bayes
min samples split	28	56
max samples leaf	20	42
max depth	18	30
ccp alpha	0.0	0.0

Tabela 1: Optymalne parametry wybrane dla drzewa losowego

2.2.2 Las losowy

Wybrane optymalne parametry:

	Random	Bayes
n estimators	16	94
min samples leaf	0.01	0.00
max samples	0.59	0.62
max features	0.62	0.37

Tabela 2: Optymalne parametry dla algorytmu Random Forest

2.2.3 SVM

Uwaga: Przy optymalizacji modelu SVM przyjęto dodatkowe założenie, tzn. ustalono parametr $max\_iter = 100$ ze względu na dużą złożoność obliczeniową tego algorytmu. Takie założenie może spowodować spadek miar AUC zwracanych przez algorytm, jednak celem projektu jest analiza możliwości poprawy jakości modelu przy tuningu parametrów, a nie samo otrzymanie najlepszego możliwego wyniku, dlatego zdecydowano na ograniczenie możliwej liczby iteracji algorytmu.

Wybrane optymalne parametry:

2.3 Analiza tunowalności algorytmów

Po wybraniu optymalnych ("defaultowych") parametrów dla każdego algorytmu zbadano tunowalność tych algorytmów na każdym z czterech zbiorów. Przeprowadzono analizę tunowalności dla dwóch metod losowania wektorów parame-

	Random	Bayes
kernel	'rbf'	'linear'
gamma	2.0	180.32
degree	4	4
C	0.031	890.422

Tabela 3: Optymalne parametry dla algorytmu SVM

trów: metody *RandomizedSearchCV* oraz optymalizacji bayesowskiej. Tunowalność zbadano na dwa sposoby:

- dla każdego zbioru zbadano różnice między wynikami z historii tuningu a wynikiem otrzymanym przy optymalnych parametrach,
- na każdym zbiorze porównano najlepszy otrzymany na nim wynik z wynikiem otrzymanym przy optymalnych parametrach, w celu pomiaru maksymalnego zysku z tuningu. Różnicę między tymi wartościami oznaczono w poniższych tabelach z wynikami jako *tunability*, a procentową zmianę względem wyniku z domyślnymi parametrami oznaczono jako *improvement*.

2.4 Wyniki eksperymentów

W poniższych tabelach wykorzystano oznaczenia:

- RS - losowanie metodą *RandomizedSearchCV*,
- B - optymalizacja bayesowska.

Wartości ujemne oznaczają poprawę wyników względem defaultowych parametrów.

2.4.1 Drzewo decyzyjne

data set	tunability RS	improvement RS	tunability B	improvement B
dane 1	0	0	-0.020	-0.024
dane 2	0	0	0.002	0.002
dane 3	-0.035	-0.076	-0.0353	-0.070
dane 4	0	0	0.011	0.012

Tabela 4: Tunability i improvement dla drzewa decyzyjnego

2.4.2 Las losowy

data set	tunability RS	improvement RS	tunability B	improvement B
dane 1	0	0	-0.0130	-0.015
dane 2	0	0	-0.053	-0.058
dane 3	-0.025	-0.051	-0.042	-0.078
dane 4	0	0	-0.051	-0.053

Tabela 5: Tunability i improvement dla algorytmu Random Forest

2.4.3 SVM

data set	tunability RS	improvement RS	tunability B	improvement B
dane 1	-0.042	-0.061	0.03	0.046
dane 2	-0.070	-0.087	0.007	0.009
dane 3	-0.065	-0.122	-0.084	-0.136
dane 4	-0.018	-0.026	-0.03	-0.042

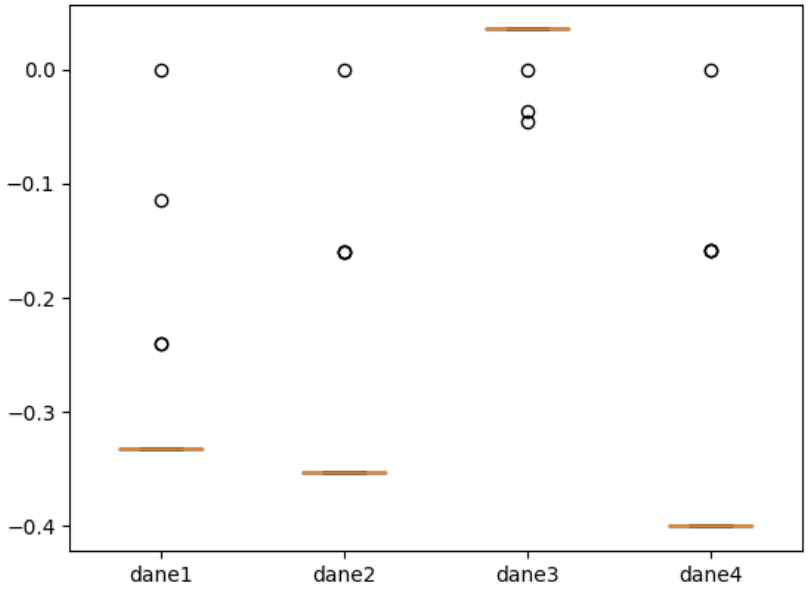
Tabela 6: Tunability i improvement dla algorytmu SVM

Na wykresach w załączniku można również zobaczyć boxploty zawierające rozkład miary tunowalności opartej na podstawie historii tuningu, a także zmiany wartości AUC w zależności od numeru iteracji dla algorytmów *RandomizedSearchCV* oraz optymalizacji bayesowskiej.

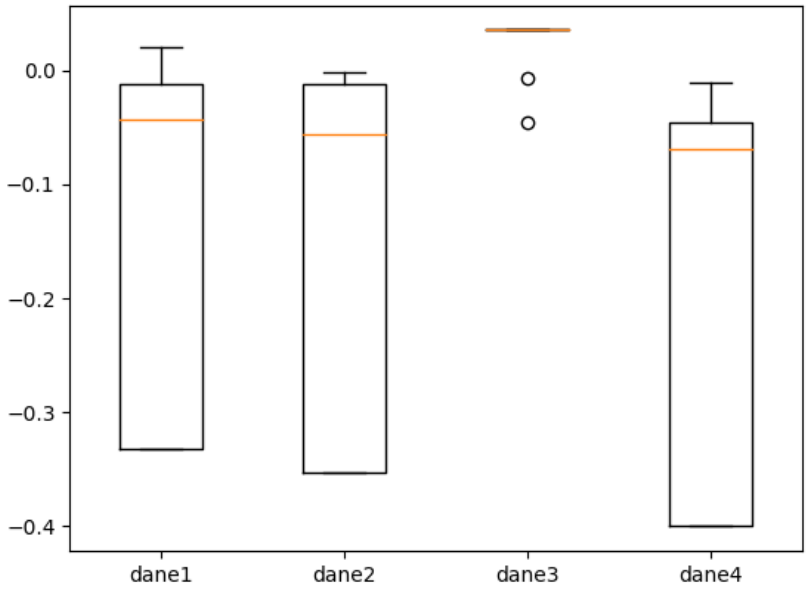
3 Wnioski

Dla drzewa decyzyjnego i lasu losowego przy metodzie *RandomizedSearchCV* widać, że dla trzech zbiorów optymalne hiperparametry okazały się najlepszymi, a dla jednego odpowiedni dobór hiperparametrów daje nieznaczną poprawę wyników. Przy modelu SVM na każdym zbiorze danych udało się nieznacznie poprawić wyniki modelu, a na trzecim zbiorze danych nastąpiła poprawa nawet o 12%. W przypadku optymalizacji bayesowskiej obserwujemy zarówno polepszenie jak i pogorszenie wyników zależnie od zbioru i modelu. Jedynie tuning na lesie losowym zapewnia poprawę wyników niezależnie od zbioru danych. Należy jednak zauważyć, że poprawa wyników następuje maksymalnie o 13%, natomiast pogorszenie maksymalnie o 5%. Dodatkowo optymalizacja bayesowska została ograniczona ze względów obliczeniowych do 50 iteracji, zatem prawdopodobnie możliwe jest nieznaczne poprawienie otrzymanych wyników.

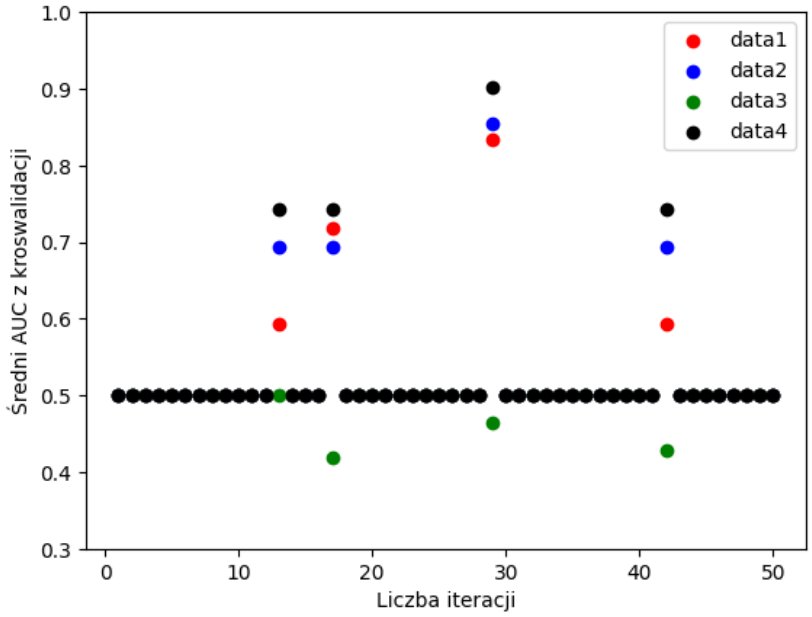
4 Załączniki



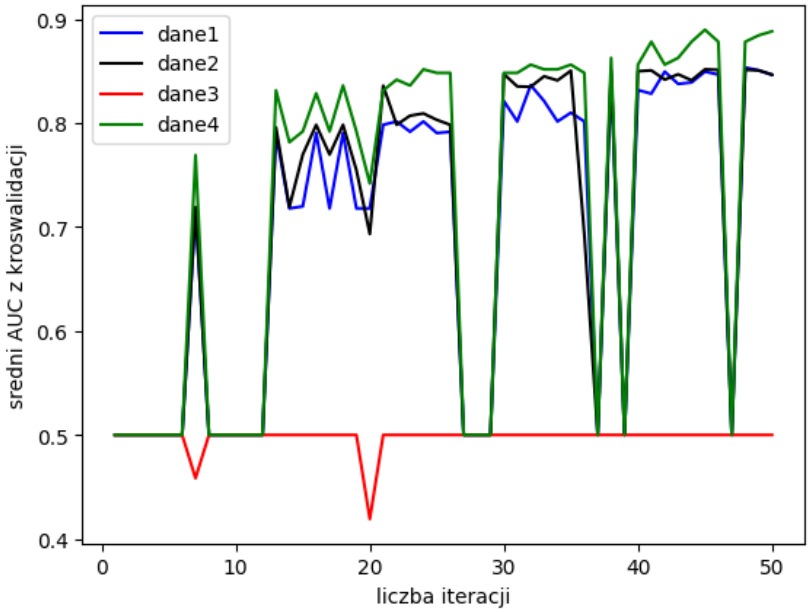
Rysunek 1: Tunability drzewa na podstawie historii tuningu - RandomizedSearchCV



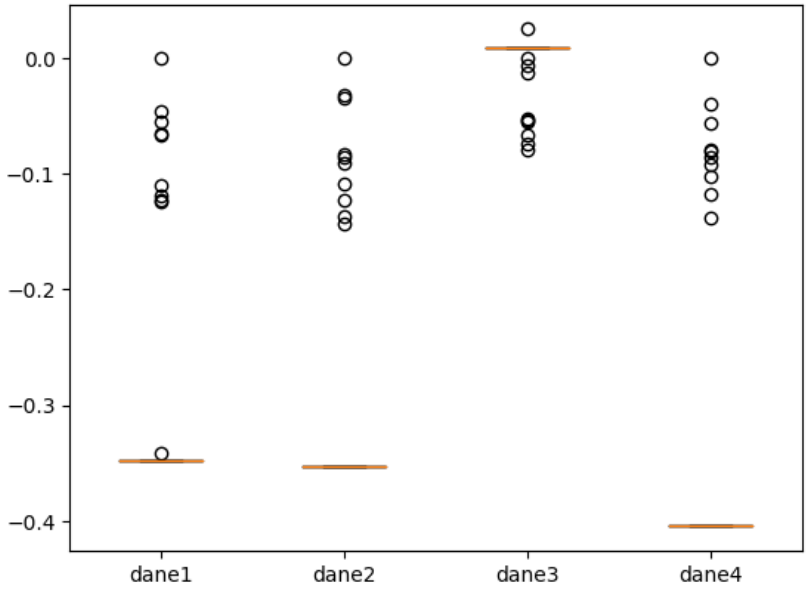
Rysunek 2: Zmiany AUC drzewa na podstawie historii tuningu - Bayes



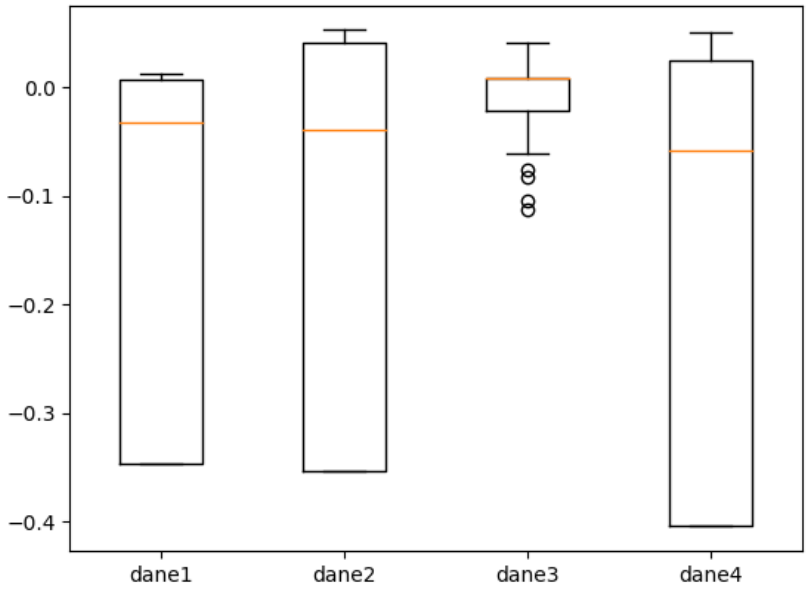
Rysunek 3: Zmiany AUC drzewa w zależności od liczby iteracji - RandomizedSearchCV



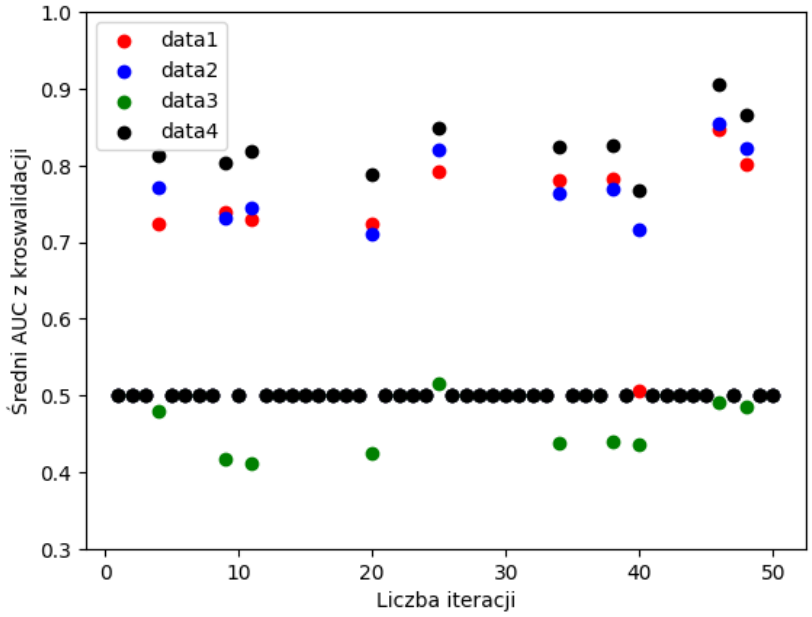
Rysunek 4: Zmiany AUC drzewa w zależności od liczby iteracji - Bayes



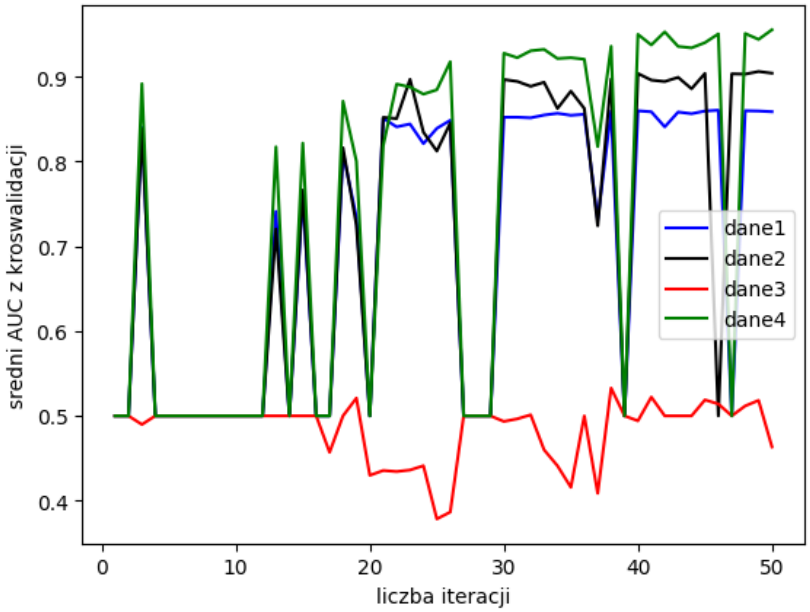
Rysunek 5: Tunability lasu losowego na podstawie historii tuningu - RandomizedSearchCV



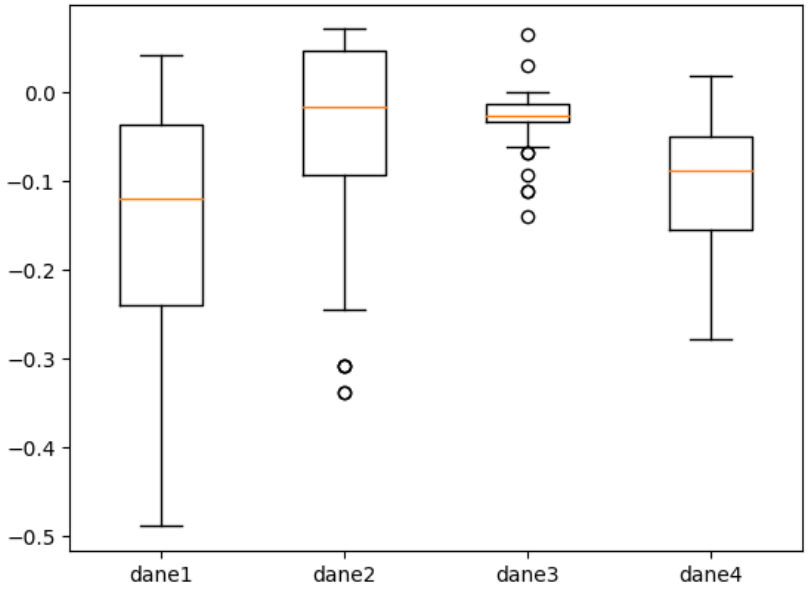
Rysunek 6: Tunability lasu losowego na podstawie historii tuningu - Bayes



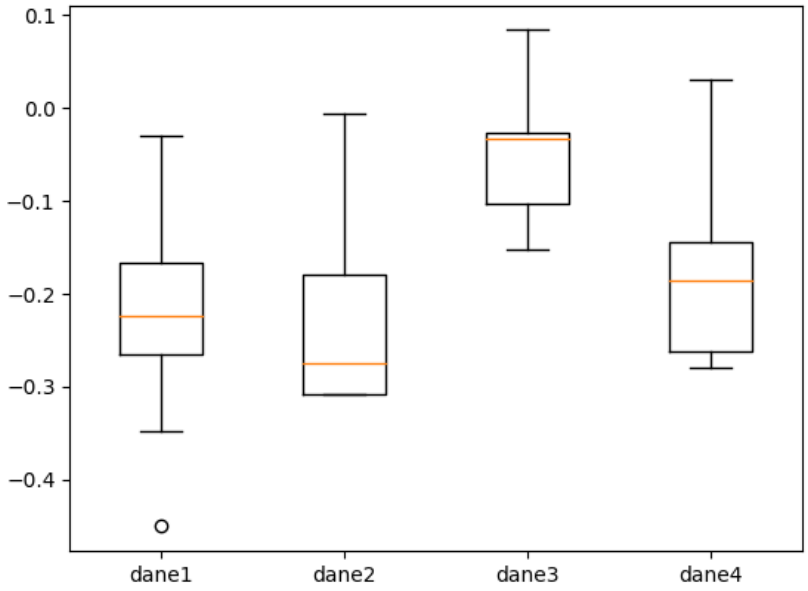
Rysunek 7: Zmiany AUC lasu losowego w zależności od liczby iteracji - RandomizedSearchCV



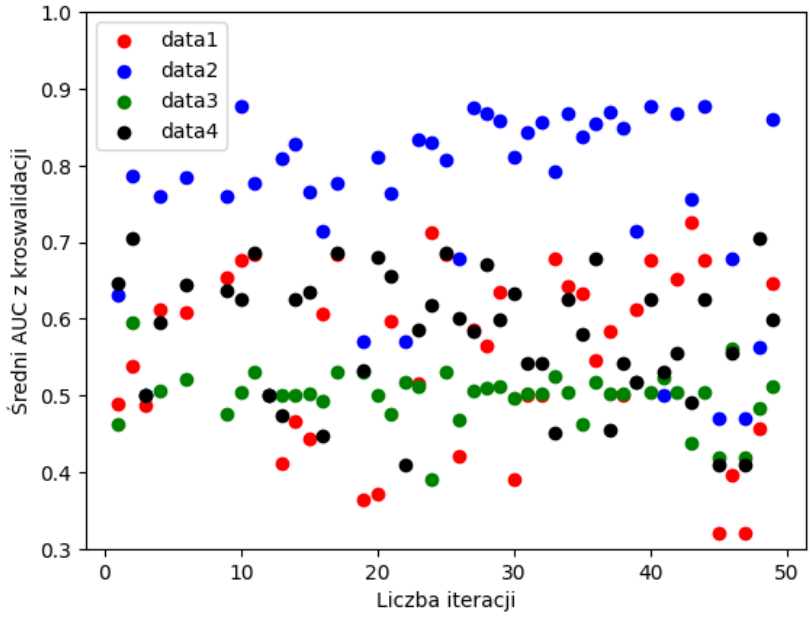
Rysunek 8: Zmiany AUC lasu losowego w zależności od liczby iteracji - Bayes



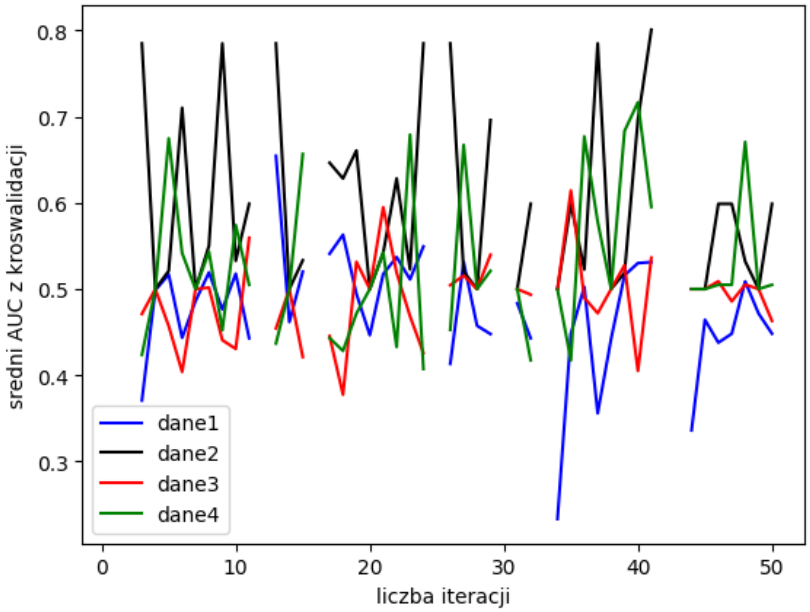
Rysunek 9: Tunability SVM na podstawie historii tuningu - RandomizedSearchCV



Rysunek 10: Tunability SVM na podstawie historii tuningu - Bayes



Rysunek 11: Zmiany AUC SVM w zależności od liczby iteracji - RandomizedSearchCV



Rysunek 12: Zmiany AUC SVM w zależności od liczby iteracji - Bayes