Analiza tunowalności algorytmów uczenia maszynowego

Bartosz Maj Bartosz Olszewski Jan Kruszewski

20.11.2024

Spis treści

1 Abstrakt

Celem niniejszego raportu jest analiza tunowalności poszczególnych hiper-parametrów w modelach uczenia maszynowego. Przeprowadziliśmy szereg eksperymentów, aby ocenić wpływ różnych wartości hiper-parametrów na wydajność modeli. Wyniki pozwalają zidentyfikować kluczowe hiper-parametry, oraz dostarczają wskazówek dotyczących ich efektywnego doboru.

2 Wstęp

Niniejszy raport koncentruje się na badaniu wpływu indywidualnych hiper-parametrów na proces uczenia maszynowego następujących modeli:

1. Las losowy - biblioteka sklearn
2. Regresja logistyczna - biblioteka sklearn
3. XGBoost - biblioteka xgboost

W celu przeprowadzenia dokładnej analizy, każdy z wyżej wymienionych modeli został przetestowany na czterech różnych zbiorach danych.

- Wine - zawiera cechy charakterystycznych dla danego wina wraz z jego ogólną oceną
- Drug - posiada dane medyczne dla pacjentów oraz lek, który został podany
- Iris - zawiera dane opisujące wymiary irysów oraz nazwę konkretnego gatunku
- Titanic - posiada dane opisujące osoby znajdujące się na statku Tytanic razem z informacją, czy dana osoba przeżyła

3 Wizyalizacje wyników

4 Wybrane hiper-parametry wraz z modelami

W poniższych podrozdziałach znajdują się opisy wykorzystanych hiper-parametrów dla poszczególnych modeli.

4.1 Las losowy

Las losowy to algorytm uczenia maszynowego oparty na drzewach decyzyjnych, gdzie każde drzewo jest trenowane na losowo wybranym podzbiorze danych i cech. Wyniki pojedynczych drzew są następnie łączone aby uzyskać dokładniejsze i odporne na przeuczenie predykcje.

Wartości hiper-parametrów dla metod Grid Search oraz Random Search

- n_estimators - [100, 200, 300]
- max_depth - [None, 5, 10]
- min_samples_split - [2, 5]
- min_samples_leaf - [1, 2]

Wartości hiper-parametrów dla metody Bayes Search

- n_estimators - (100, 300)
- max_depth - (1, 15)
- min_samples_split - (2, 10)
- min_samples_leaf - (1, 5)

4.2 Regresja logistyczna

Regresja logistyczna to algorytm uczenia maszynowego służący do modelowania prawdopodobieństwa wystąpienia określonego zdarzenia poprzez zastosowanie funkcji logistycznej.

Wartości hiper-parametrów dla metod Grid Search oraz Random Search

- C - [0.01, 0.1, 1, 10]
- penalty - ['l2']
- solver - ['lbfgs', 'saga']

Wartości hiper-parametrów dla metody Bayes Search

- C - (0.01, 10) z rozmieszczeniem logarytmicznym
- penalty - ['l2']
- solver - ['lbfgs', 'saga']

4.3 XGBoost

XGBoost to zoptymalizowany algorytm uczenia maszynowego oparty na technice gradientowego wzmocnienia (gradient boosting), który wykorzystuje drzewa decyzyjne jako modele bazowe.

Wartości hiper-parametrów dla metod Grid Search oraz Random Search

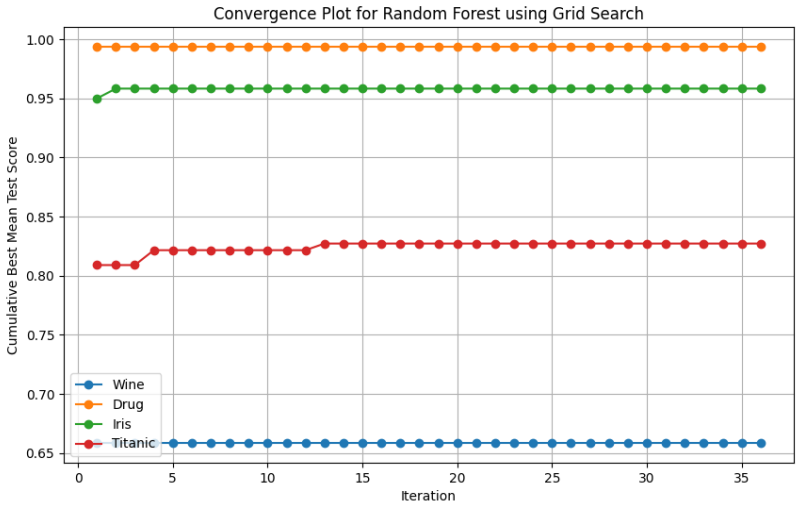
- n_estimators - [100, 300]
- max_depth - [3, 5, 7]
- learning_rate - [0.01, 0.2]
- subsample - [0.6, 1.0]

Wartości hiper-parametrów dla metody Bayes Search

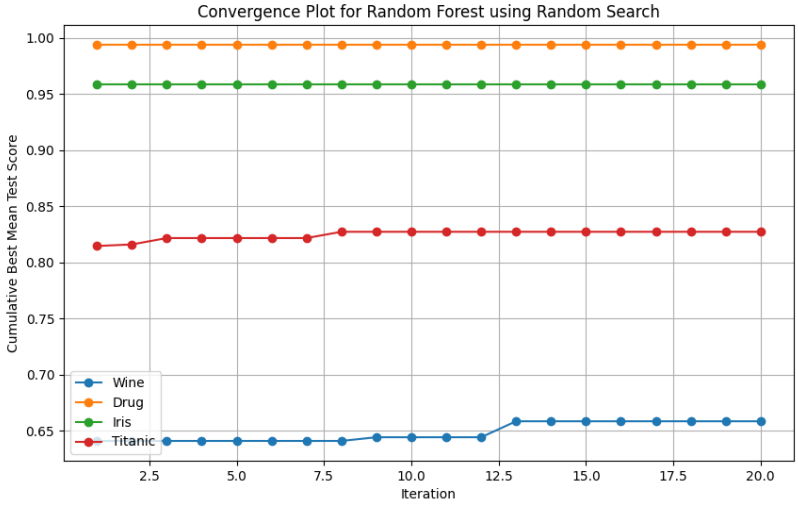
- n_estimators - (100, 300)
- max_depth - (3, 8)
- learning_rate - (0.01, 0.3) z rozmieszczeniem logarytmicznym
- subsample - (0.5, 1.0)

5 Wymagana liczba iteracji

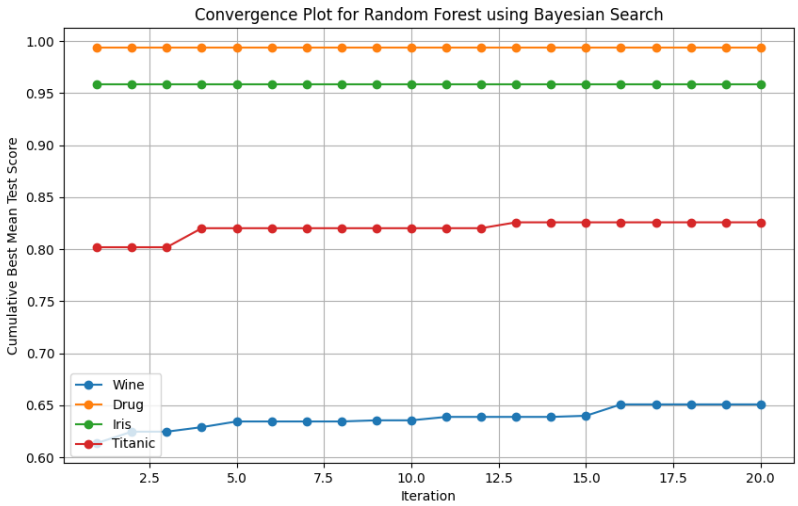
Efektywność danej metody zależy między innymi od ilości iteracji danego algorytmu. Na poniższych wykresach została przedstawiona analiza ilości potrzebnych iteracji w celu uzyskania stabilnych wyników optymalizacji dla każdego modelu oraz metody samplingu.



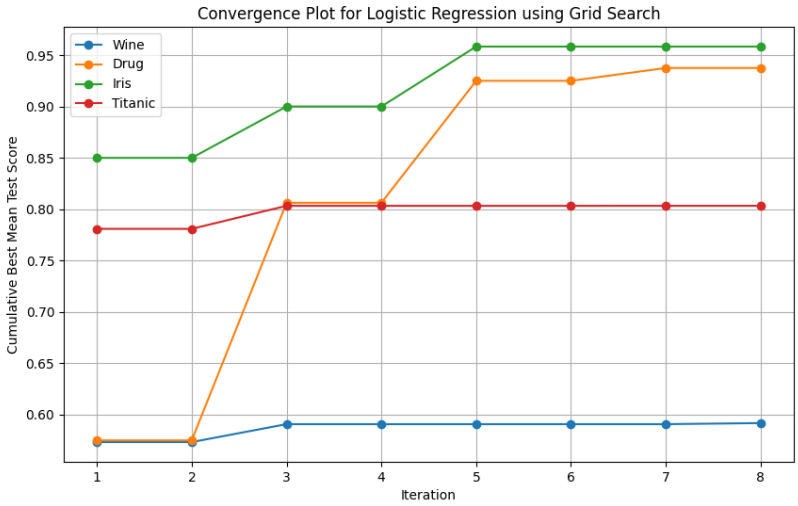
Rysunek 1: Wyniki dla lasu losowego i metody Grid Search



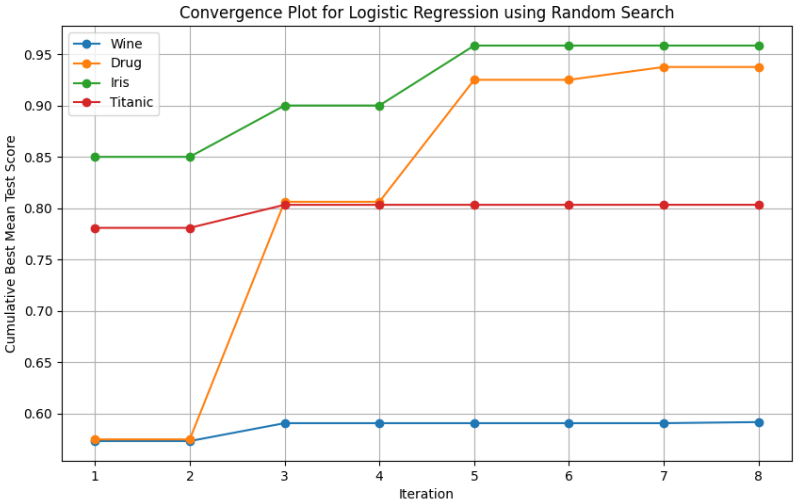
Rysunek 2: Wyniki dla lasu losowego i metody Random Search



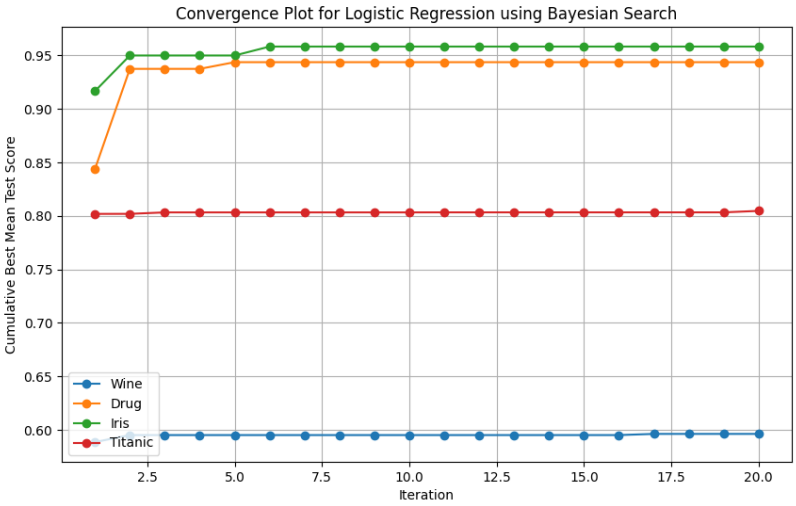
Rysunek 3: Wyniki dla lasu losowego i metody Bayesian Search



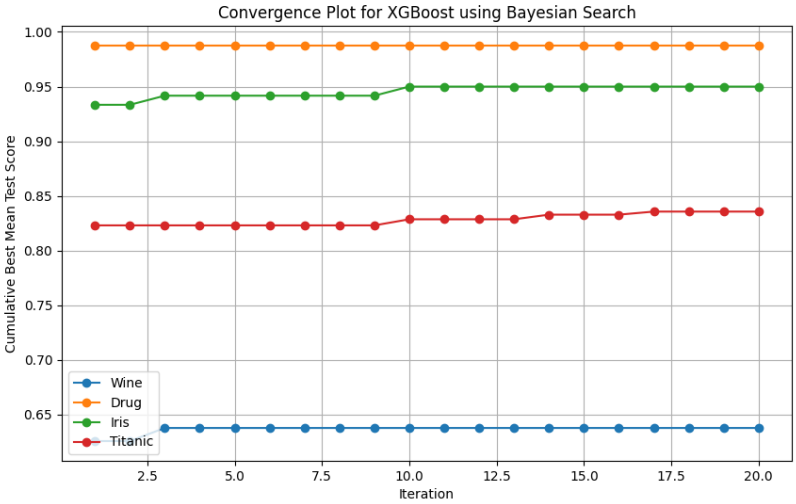
Rysunek 4: Wyniki dla regresji logistycznej i metody Grid Search



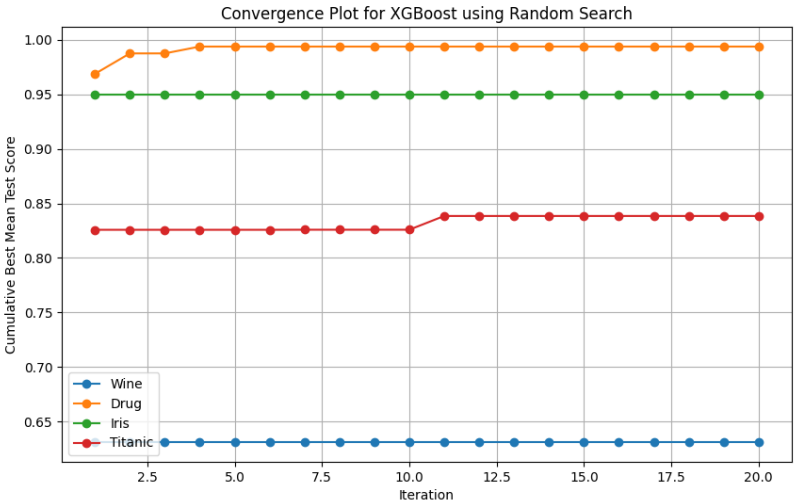
Rysunek 5: Wyniki dla regresji logistycznej i metody Random Search



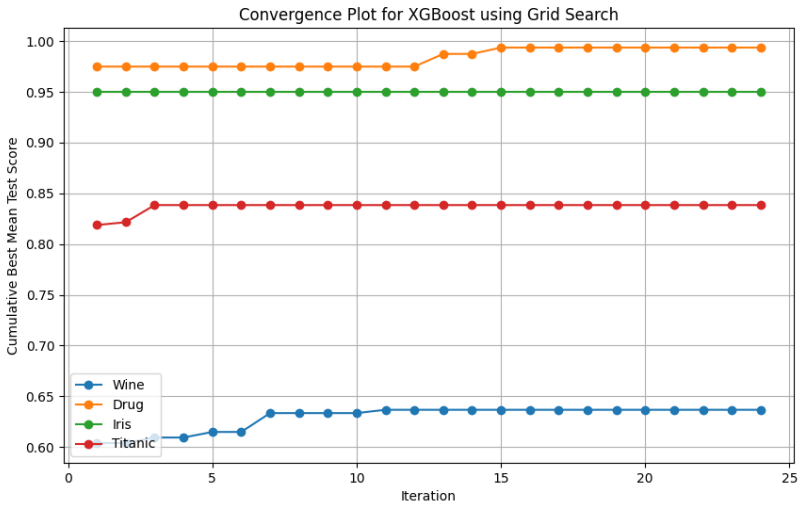
Rysunek 6: Wyniki dla regresji logistycznej i metody Bayesian Search



Rysunek 7: Wyniki dla XGBoost i metody Grid Search

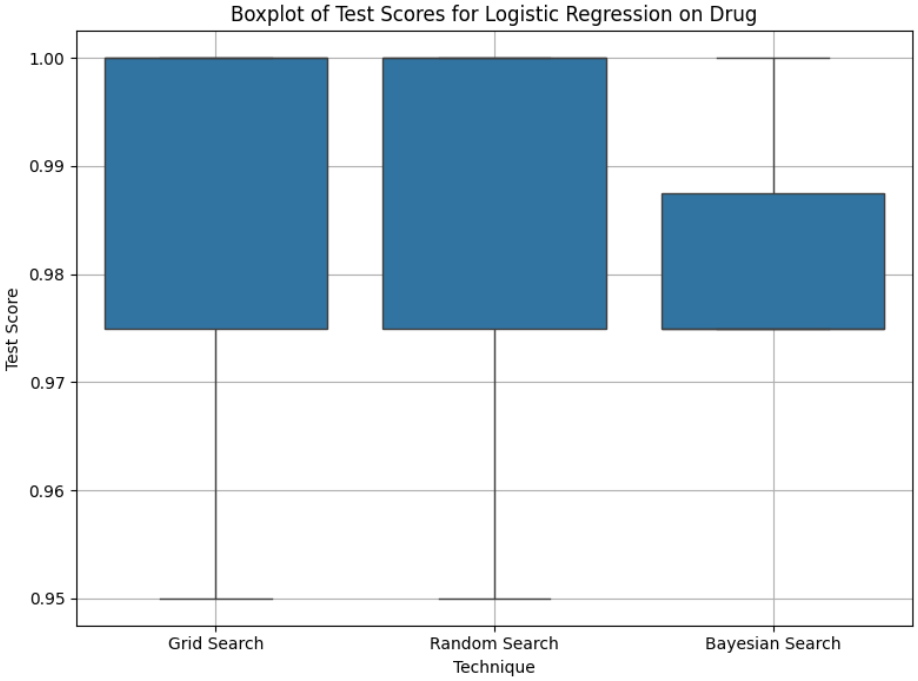


Rysunek 8: Wyniki dla XGBoost i metody Random Search

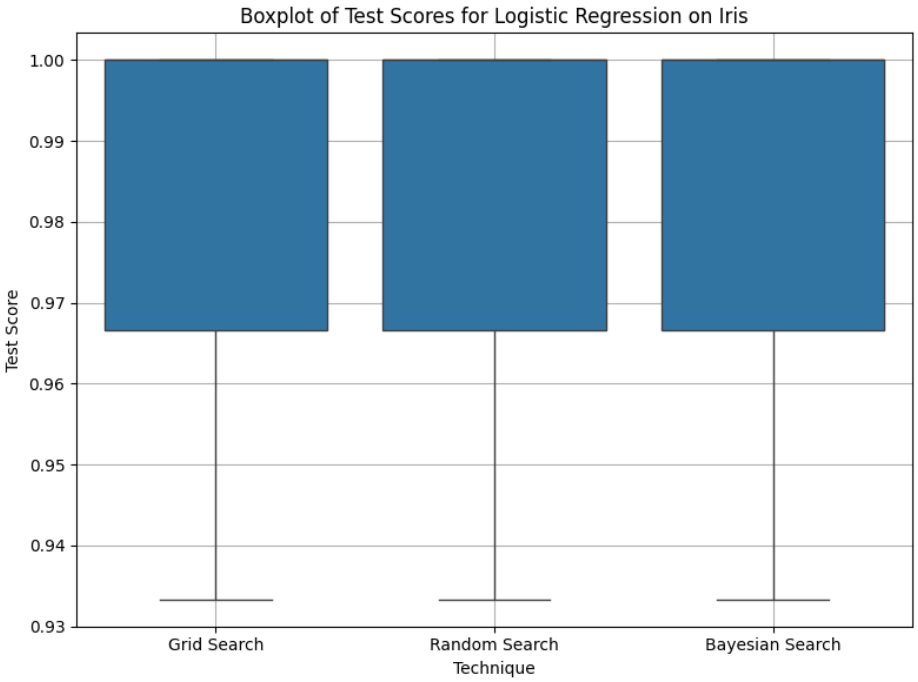


Rysunek 9: Wyniki dla XGBoost i metody Bayesian Search

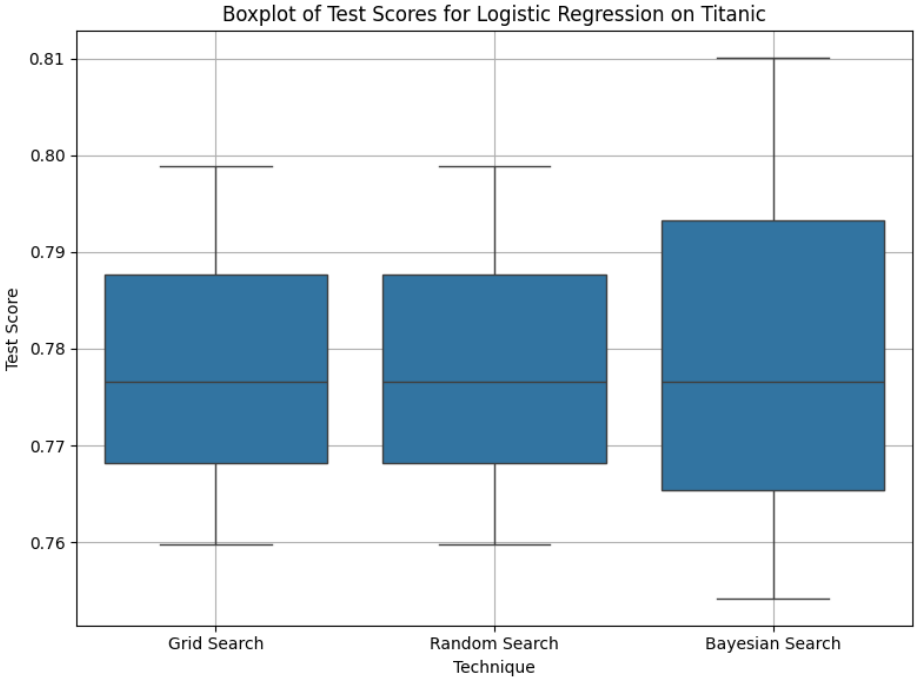
6 Wizualizacja wyników



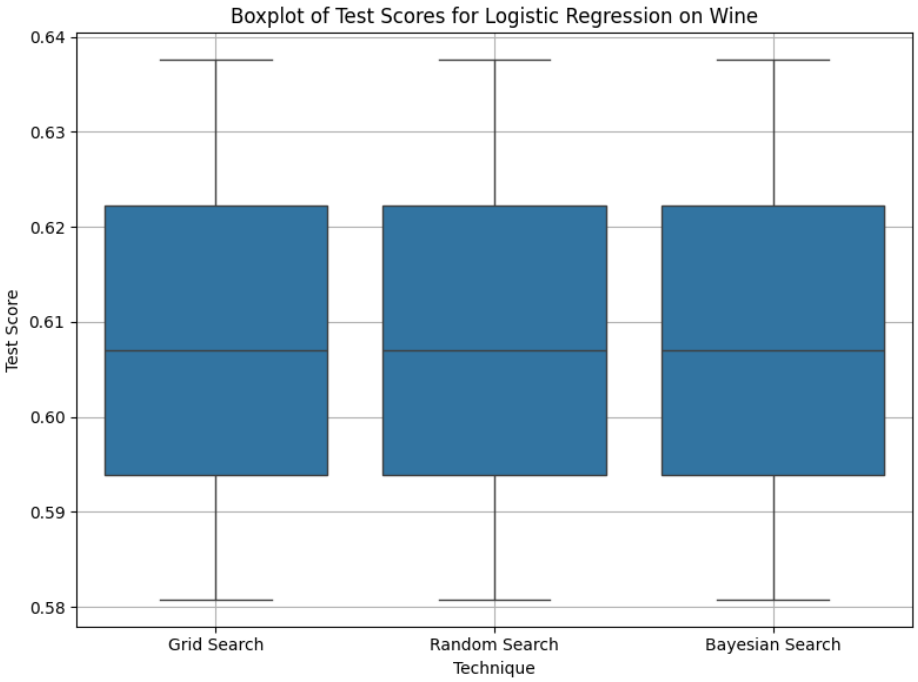
Rysunek 10: Wyniki regresji logistycznej dla zbioru danych Drug



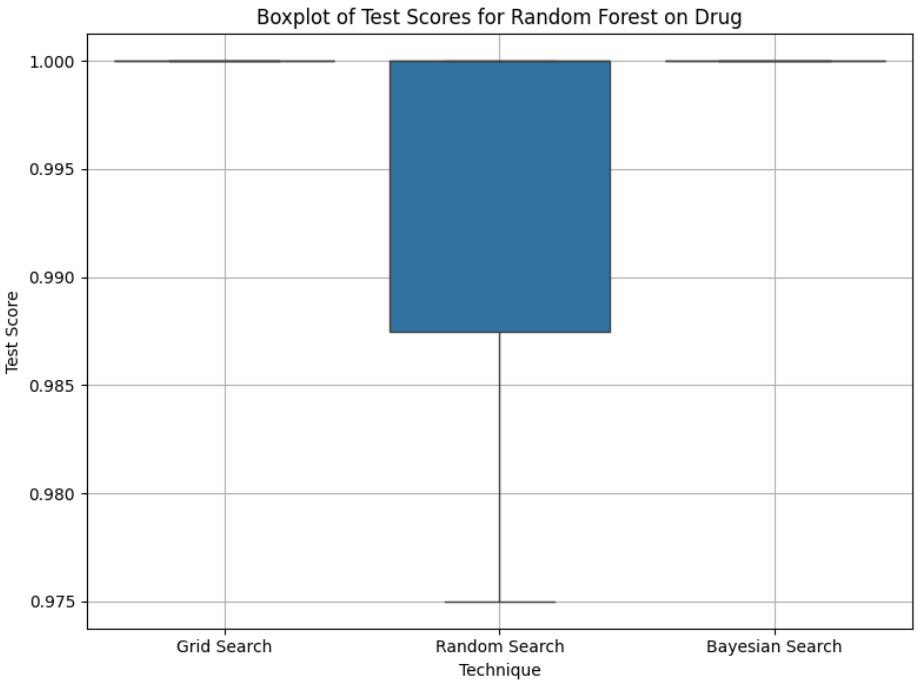
Rysunek 11: Wyniki regresji logistycznej dla zbioru danych Iris



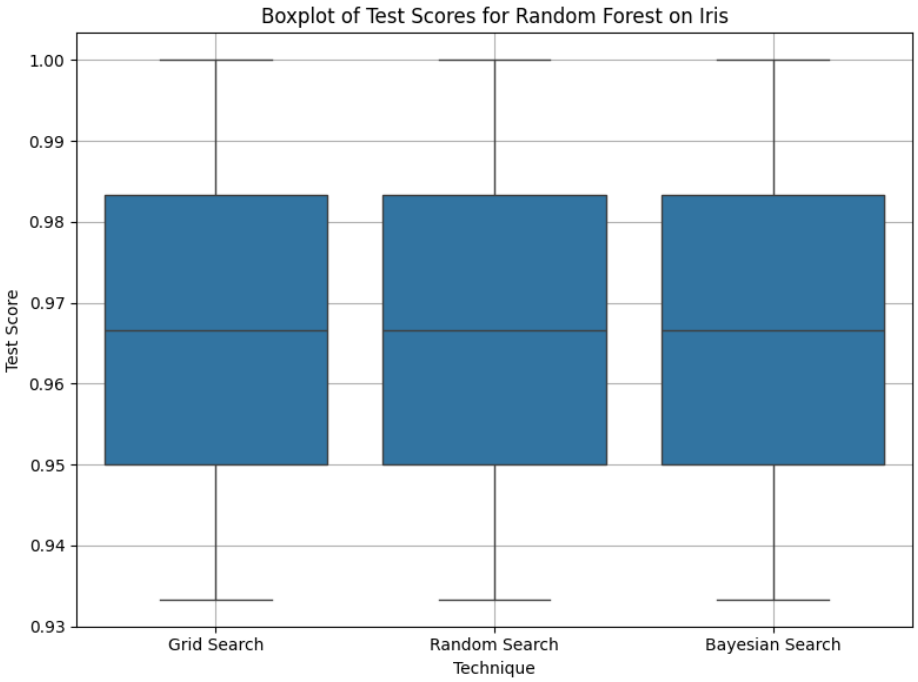
Rysunek 12: Wyniki regresji logistycznej dla zbioru danych Titanic



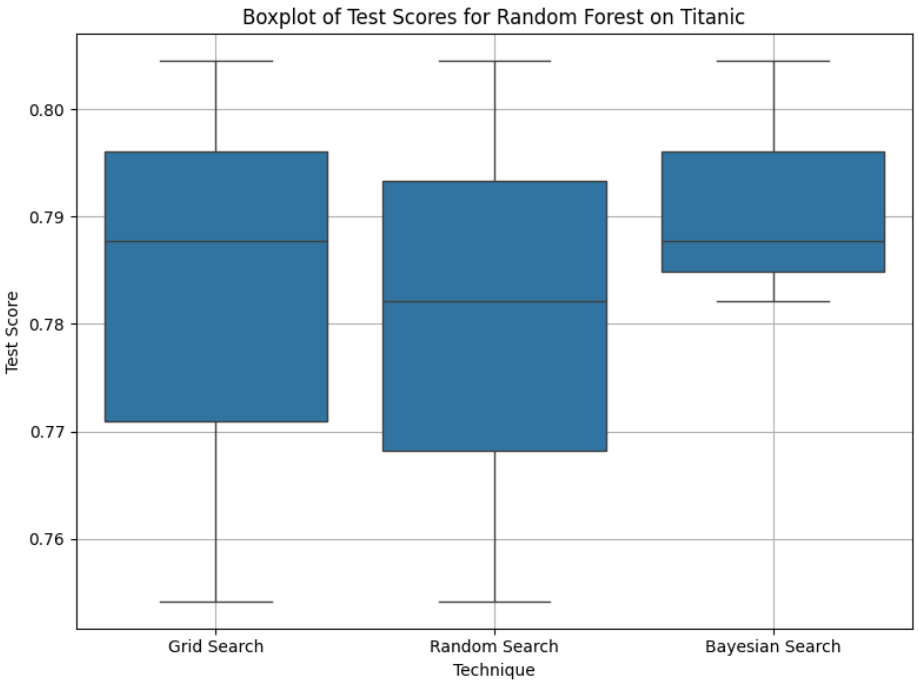
Rysunek 13: Wyniki regresji logistycznej dla zbioru danych Wine



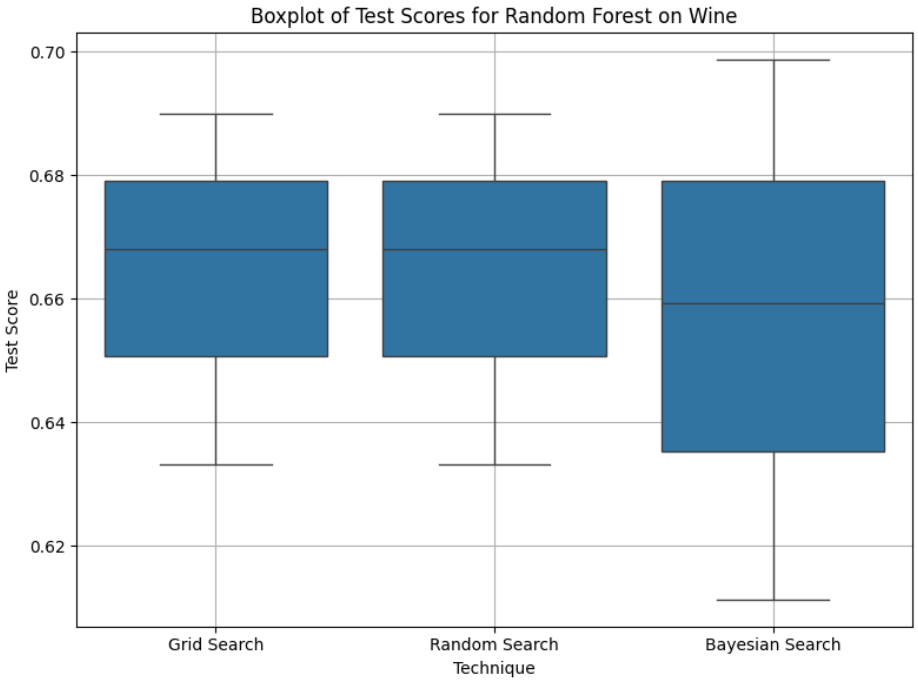
Rysunek 14: Wyniki lasu losowego dla zbioru danych Drug



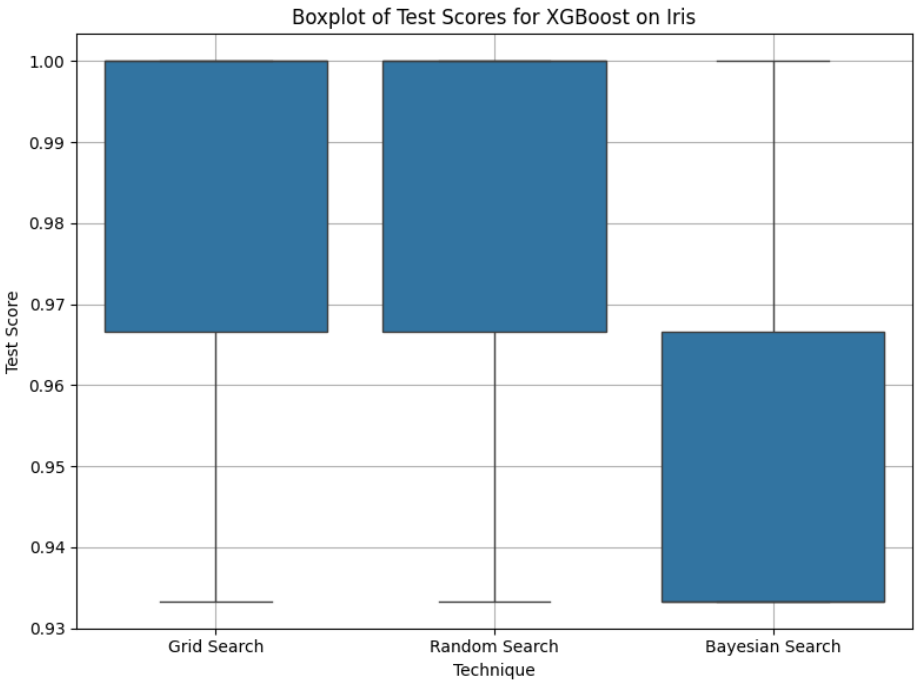
Rysunek 15: Wyniki lasu losowego dla zbioru danych Iris



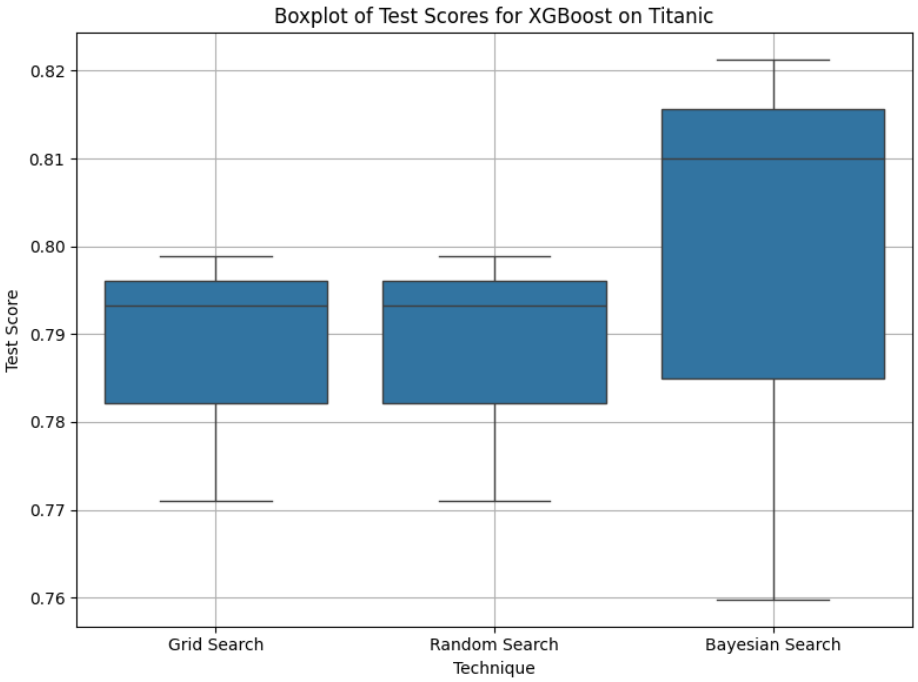
Rysunek 16: Wyniki lasu losowego dla zbioru danych Titanic



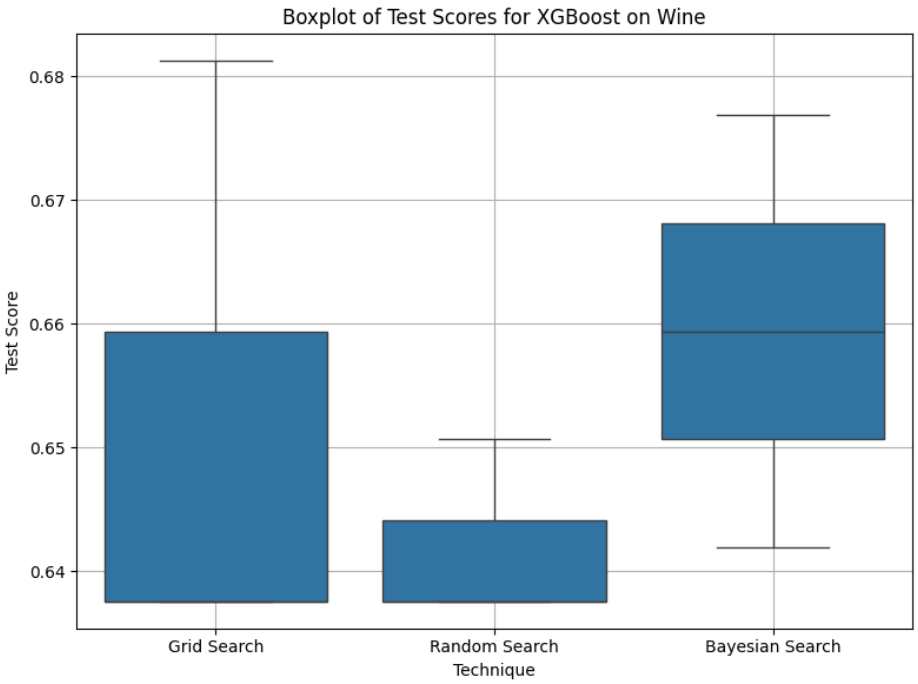
Rysunek 17: Wyniki lasu losowego dla zbioru danych Wine



Rysunek 18: Wyniki XGBoost dla zbioru danych Iris



Rysunek 19: Wyniki XGBoost dla zbioru danych Titanic



Rysunek 20: Wyniki XGBoost dla zbioru danych Wine

7 Testy statystyczne

W celu zbadania różnicy wyników dla różnych technik wyboru hiper-parametrów użyto testu Wilcoxona. Porównane zostały ze sobą każde dwie metody, dla każdego zbioru danych i algorytmu. W Tabeli 1 znajdują się przykładowe wyniki testu. Pozostałe wyniki znajdują się w pliku statistical_test_results.csv.

Algorytm	Zbiór danych	Technika 1	Technika 2	p-value
Random Forest	Wine	Grid Search	Bayesian Search	0.5
XGBoost	Wine	Grid Search	Random Search	0.32
Random Forest	Titanic	Grid Search	Bayesian Search	0.65
Logistic Regression	Titanic	Random Search	Bayesian Search	0.65
Logistic Regression	Drug	Random Search	Bayesian Search	1.0

Tabela 1: Tabela z przykładowymi wynikami testu Wilcoxona

8 Wnioski

Celem pracy było przeanalizowanie tunowalności hiperparametrów dla Lasu Losowego, Regresji logistycznej oraz XGBoost z wykorzystaniem trzech metod samplingu. Analiza różnych zbiorów danych wykazała, że charakterystyka danych wpływa na skuteczność poszczególnych metod optymalizacji. Projekt wykazał, że techniki tunowania hiperparametrów mają istotny wpływ na efektywność optymalizacji modeli uczenia maszynowego.

9 Bibliografia

- Zbior danych Wine - `https://www.kaggle.com/datasets/yasserh/wine-quality-dataset/data`
- Zbior danych Drug - `https://www.kaggle.com/datasets/ammaraahmad/top-10-machine-learning-data`
- Zbior danych Iris - `https://scikit-learn.org/1.5/auto_examples/datasets/plot_iris_dataset.html#sphx-glr-auto-examples-datasets-plot-iris-dataset-py`
- Zbior danych Tytanic - `https://www.kaggle.com/c/titanic/overview`