# OPTYMALIZACJA HIPERPARAMETRÓW MODELI UCZENIA MASZYNOWEGO

MACIEJ MOMOT, MIKOŁAJ ROWICKI,
KRZYSZTOF TKACZYK

# ZBIORY DANYCH

W naszym eksperymencie wykorzystaliśmy 4 zbiory danych pochodzące z repozytorium  kaggle.com oraz UC Irvine Machine Learning Repository. Wszystkie zbiory mają dokładnie 5000 wierszy i od 17 do 25 kolumn.

**01**

**Airline Passenger Satisfaction**

**02**

**Rain in Australia**

**03**

**Banking Dataset**

**04**

**Mushroom Dataset**

# ALGORYTMY UCZENIA MASZYNOWEGO

**TYP KLASYFIKACJI:**
**BINARNA**

**METODY OPTYMALIZACJI:**
- **BAYES SEARCH**
- **RANDOM SEARCH**

**METRYKA:**
**ROC-AUC**

**01**

**DECISION TREE**

**02**

**RANDOM FOREST**

**03**

**XGBOOST**

# WYBÓR HIPERPARAMETRÓW

| Model | Parametr | Typ danych | Dolny zakres | Górny zakres |
|---|---|---|---|---|
| Decision Tree | max_depth | integer | 1 | 30 |
| | min_samples_split | integer | 2 | 60 |
| | criterion | discrete | - | - |
| | min_samples_leaf | integer | 1 | 60 |
| Random Forest | n_estimators | integer | 100 | 500 |
| | min_samples_leaf | integer | 1 | 250 |
| | max_samples | numeric | 0.5 | 1 |
| | max_features | numeric | $10^{-6}$ | 1 |
| XGBoost | max_depth | integer | 1 | 19 |
| | min_child_weight | integer | 0 | 19 |
| | eta | numeric | 0.01 | 0.11 |
| | alpha | lognumeric | $10^{-4}$ | 10 |

# ZNALEZIONE OPTYMALNE KONFIGURACJE HIPERPARAMETRÓW

| Model | Parametry | Score |
|---|---|---|
| DecisionTree | criterion: gini<br>max_depth: 17<br>min_samples_leaf: 10<br>min_samples_split: 58 | 0,904 |
| RandomForest | max_features: 0,498<br>max_samples: 0,738<br>min_samples_leaf: 3<br>n_estimators: 478 | 0,94 |
| XGBoost | alpha: 1,248<br>eta: 0,098<br>max_depth: 16<br>min_child_weight: 0 | 0,94 |

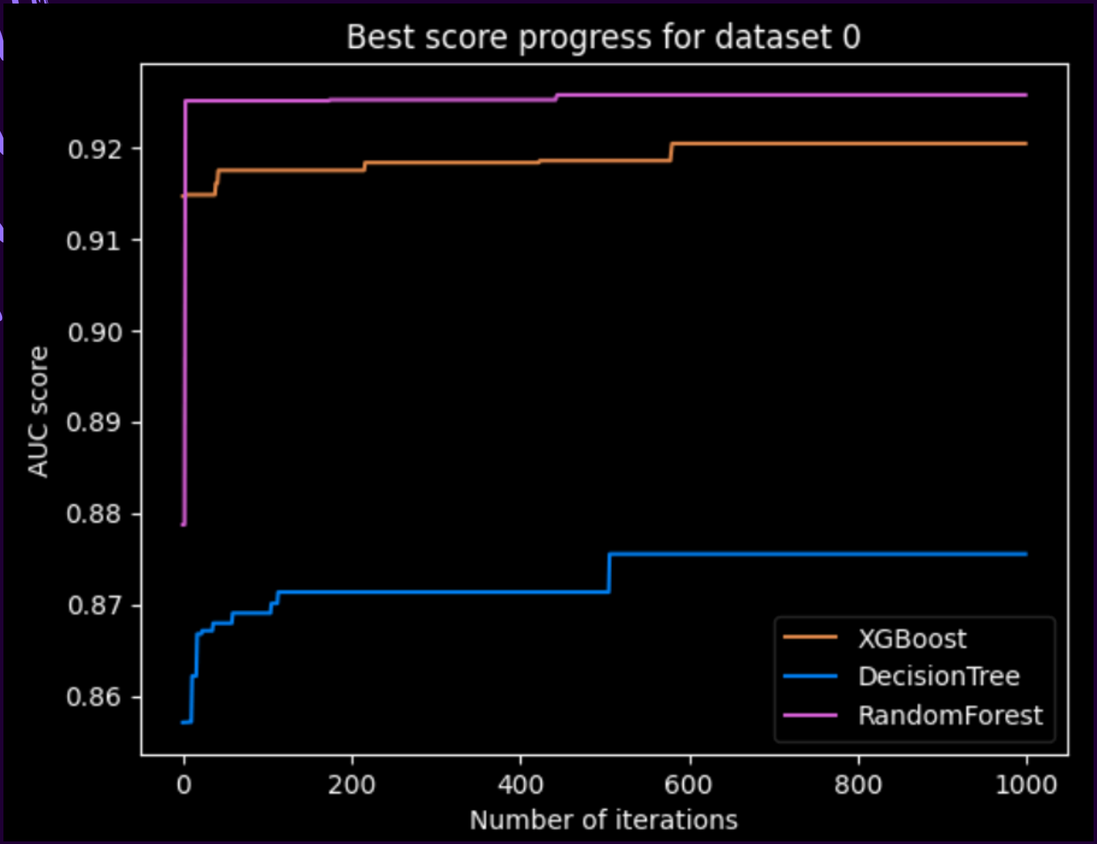
# TUNOWALNOŚĆ ALGORYTMÓW

$$tunability = \frac{best\_score - default\_score}{default\_score} \cdot 100\%$$

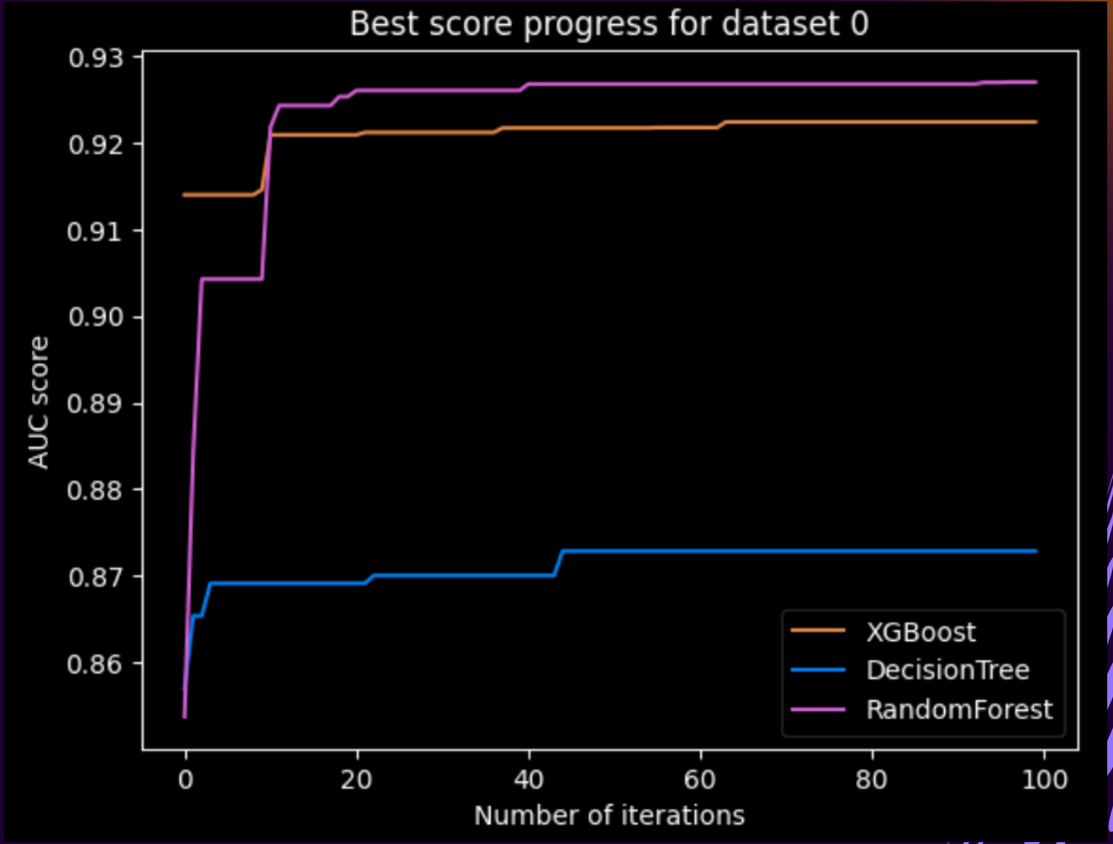
| Zbiór | Decision Tree | | Random Forest | | XGBoost | |
|---|---|---|---|---|---|---|
| | RS | BO | RS | BO | RS | BO |
| weather | 0,649% | 0,372% | 0,058% | 0,197% | 0% | 0,219% |
| flights | 0,491% | 0,632% | 0,174% | 0,206% | 0,027% | 0,012% |
| banking | 1,166% | 1,260% | 0,003% | 0,003% | 0,001% | 0,001% |
| mushrooms | 2,490% | 2,307% | 0,160% | 0,095% | 0,301% | 0,297% |

# ZBIEŻNOŚĆ OPTYMALIZACJI

## RANDOM SEARCH




## BAYES SEARCH

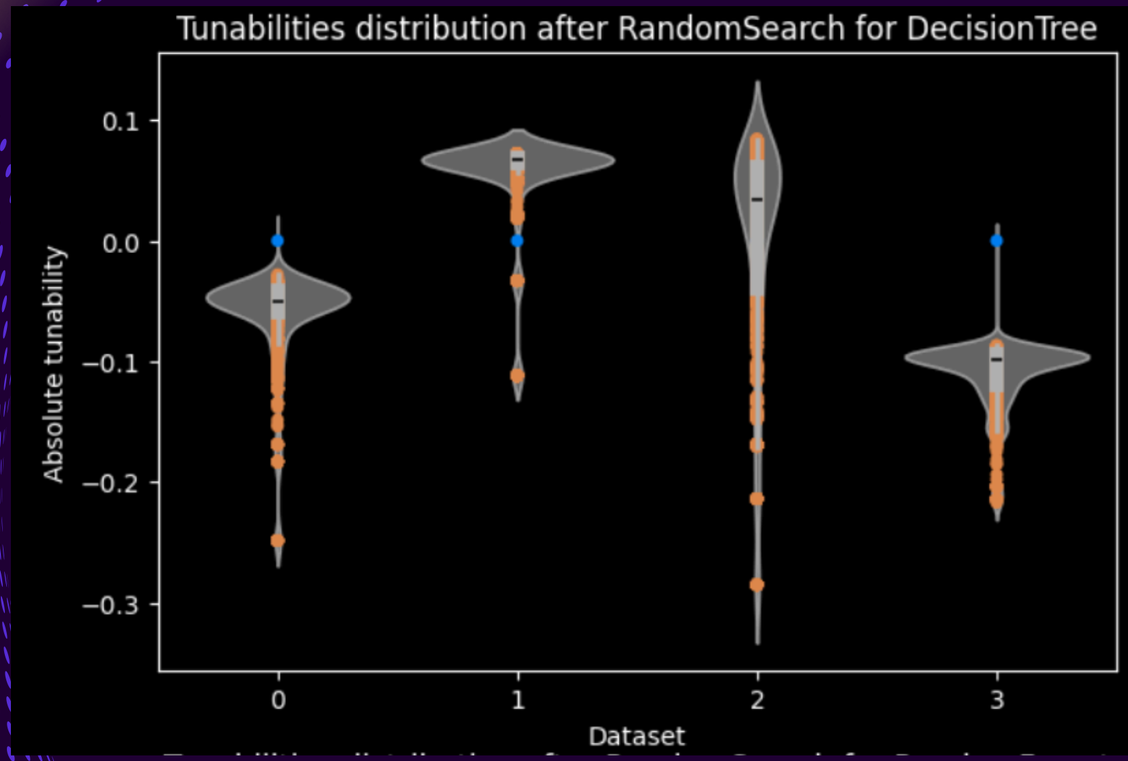


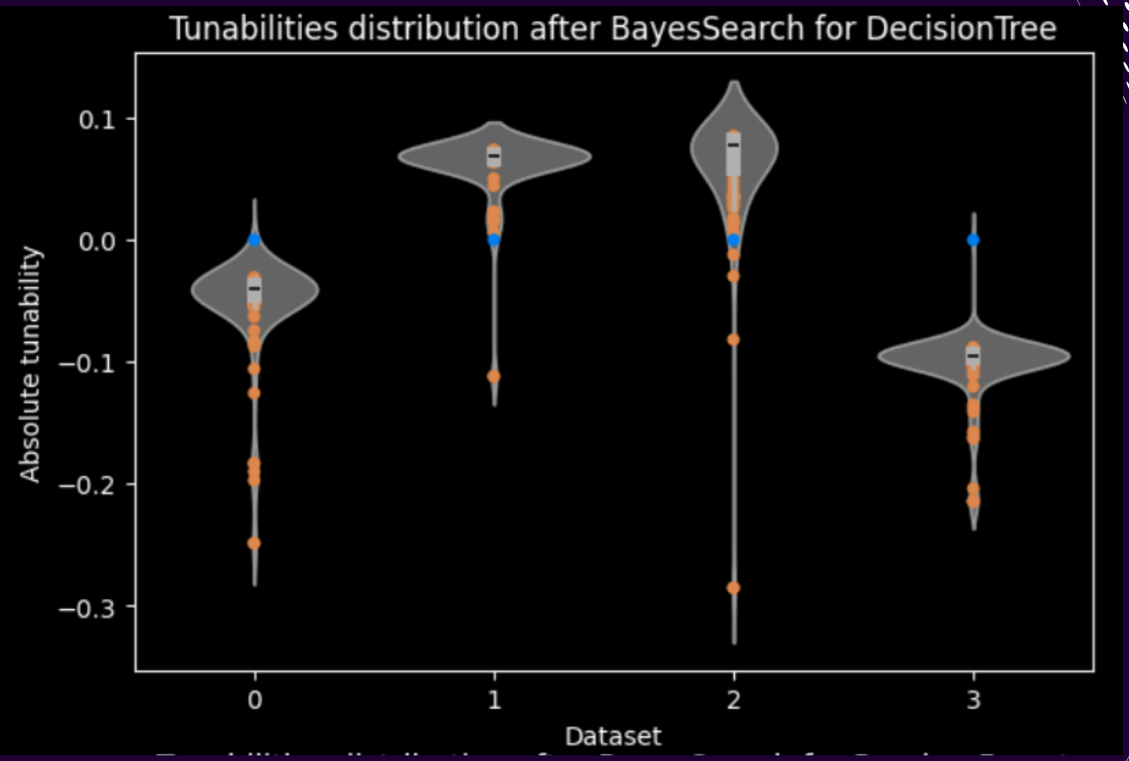
# TUNOWALNOŚĆ HIPERPARAMETRÓW

$$\text{tunability} = \text{mean}\left(\max\left\{\frac{\text{best\_score} - \text{default\_score}}{\text{default\_score}} \cdot 100\%, 0\%\right\}\right)$$

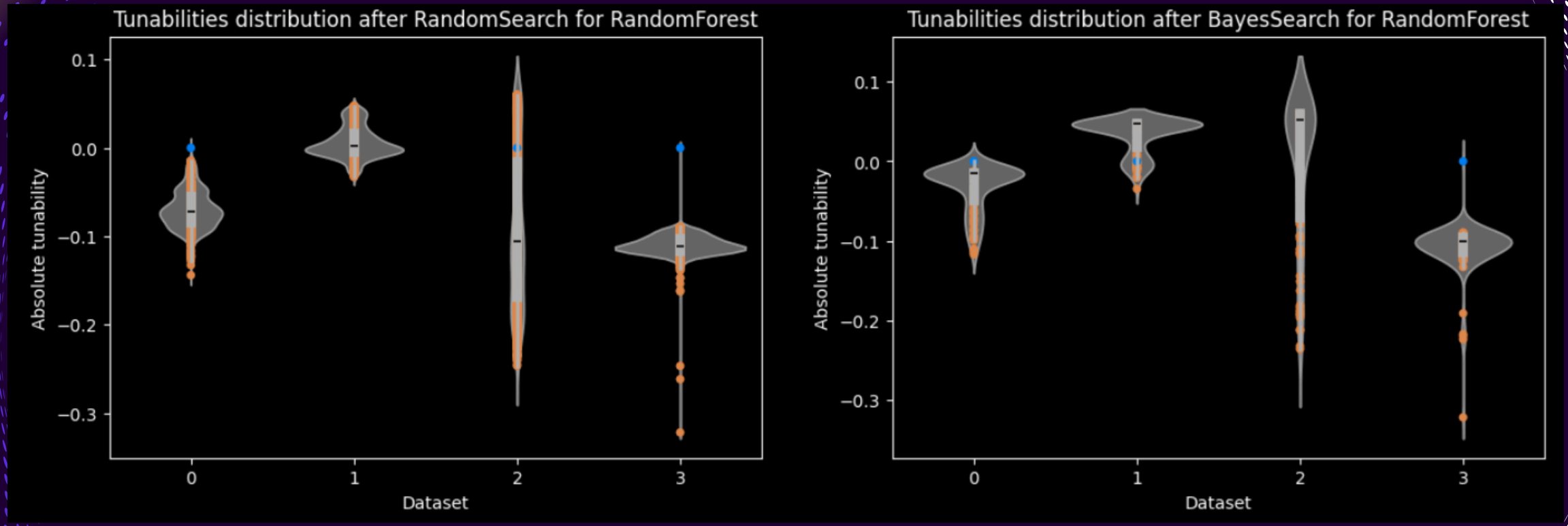
| Model | Parametr | Random Search | Bayes Search |
|---|---|---|---|
| Decision Tree | max_depth | **0.523%** | 0.522% |
| | min_samples_split | 0.215% | **0.217%** |
| | criterion | 0.178% | **0.179%** |
| | min_samples_leaf | 0.831% | **0.841%** |
| Random Forest | n_estimators | **0.073%** | 0.066% |
| | min_samples_leaf | **0.101%** | 0.069% |
| | max_samples | **0.077%** | 0.065% |
| | max_features | **0.078%** | 0.077% |
| XGBoost | max_depth | **0.046%** | 0.022% |
| | min_child_weight | **0.046%** | 0.001% |
| | eta | **0.120%** | 0.052% |
| | alpha | 0.092% | **0.182%** |

# ROZKŁAD TUNOWALNOŚCI - DECISION TREE

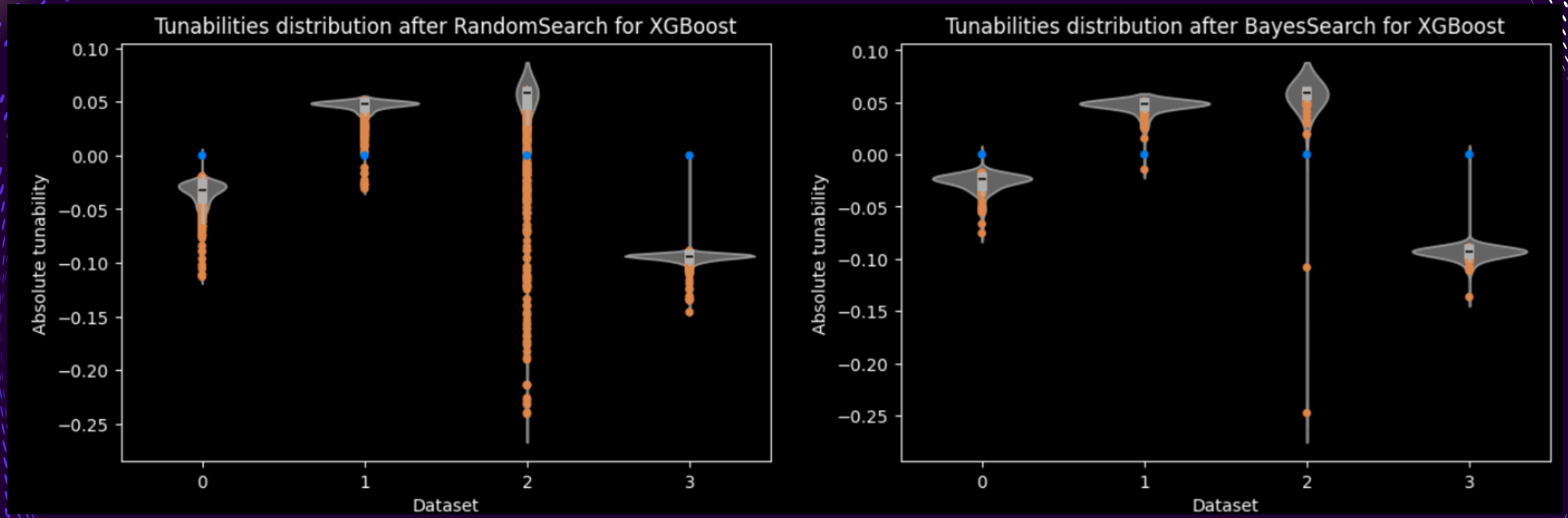


# ROZKŁAD TUNOWALNOŚCI - RANDOM FOREST



# ROZKŁAD TUNOWALNOŚCI - XGBOOST



# WNIOSKI

**01** Niezależnie od wybranej metody optymalizacji najbardziej tunowalny jest algorytm Decision Tree, a w szczególności 3 spośród jego hiperparametrów

**02** Dla 5 par zbiór-model lepsze wyniki osiąga RandomSearch, dla 5 BayesSearch, a w 2 przypadkach jest remis.

**03** Test Wilcoxona na poziomie istotności 0,01 wykazał istotne różnice między wynikami RandomSearcha i BayesSearcha dla 9 z 12 par zbiór-model.

# DZIĘKUJEMY ZA UWAGĘ