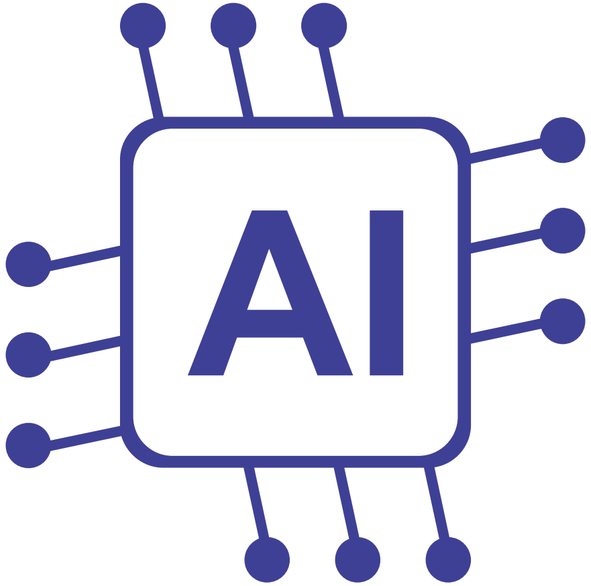
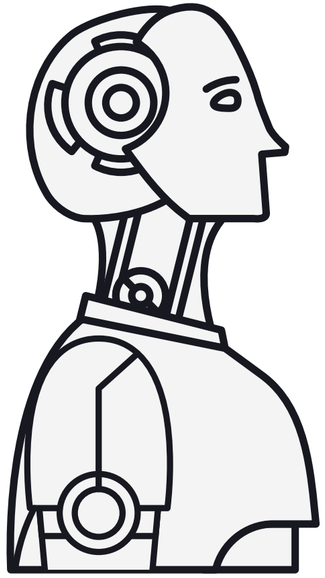




Projekt 1

Norbert Frydrysiak & Bartosz Jezierski

Dane

- **Credit Score** - dane finansowe o klientach.
- Cechy: dług, przychód, oszczędności...
- Cel: wykrycie potencjalnej niewypłacalności
- 1000 wierszy i aż 84 cechy.
- Podstawowe modele słabo dają sobie radę (wynik dla lr 0.66 roc_auc)
- Mocno skolerowane cechy

- **Raisin data** - dane o 2 rodzajach rodzynek.
- Cechy: powierzchnia, szerokość, długość...
- 900 wierszy po 450 na rodzaj rodzynki
- 8 cech
- Podstawowe modele dobrze sobie radzą (wynik dla lr 0.92 roc_auc)

Dane

- **Alzheimer’s Disease** - dane medyczne o przypadkach choroby Alzheimer’a.
- Cechy: dane zdrowotne, niektóre objawy, dane demograficzne
- 2149 wierszy i 32 cechy
- Podstawowe modele dobrze sobie radzą (wynik dla lr 0.89 roc_auc)

- **Salary data** - dane demograficzne o osobach z różnych grup społecznych
- Cechy: wiek, płeć, zawód, edukacja, rasa...
- Cel: Czy zarabiają więcej czy mniej niż 50 000$ rocznie
- 1000 wierszy i 14 cech
- Wynik dla lr_default 0.88 roc_auc

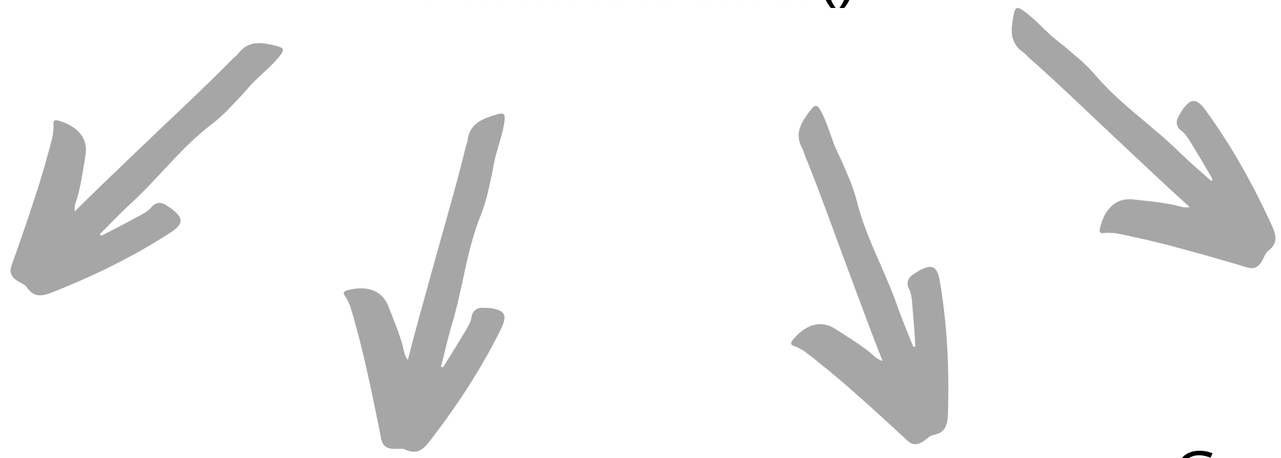
Tunowalność

$$T \stackrel{\text{def}}{=} M(\theta_{opt}) - M(\theta_{default})$$

Metryka to roc auc

Modele

MinMaxScaler()



LogisticRegression

RandomForest

KNeighbors

GradientBoosting

Algorytmy przeszukiwania

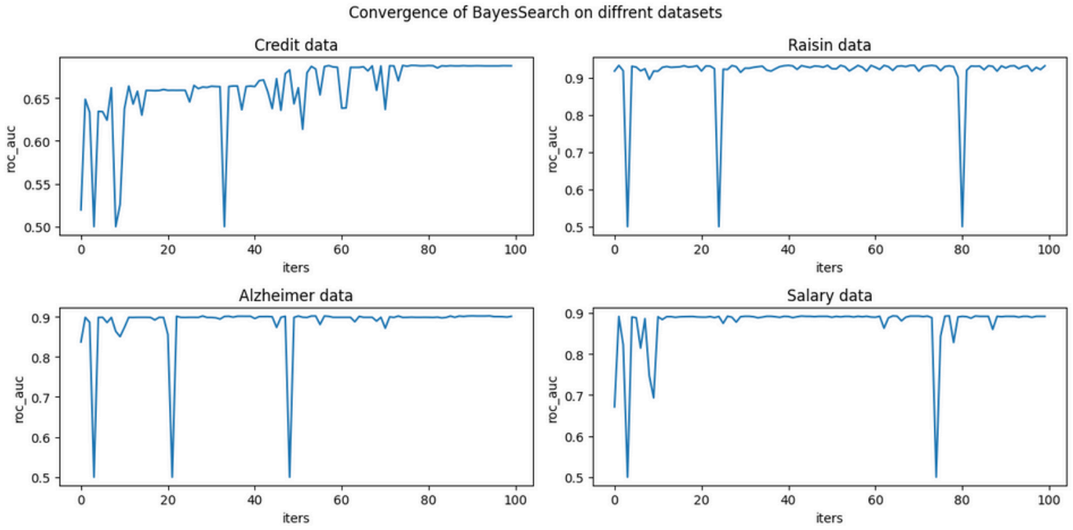
BayesSearch

RandomSearch

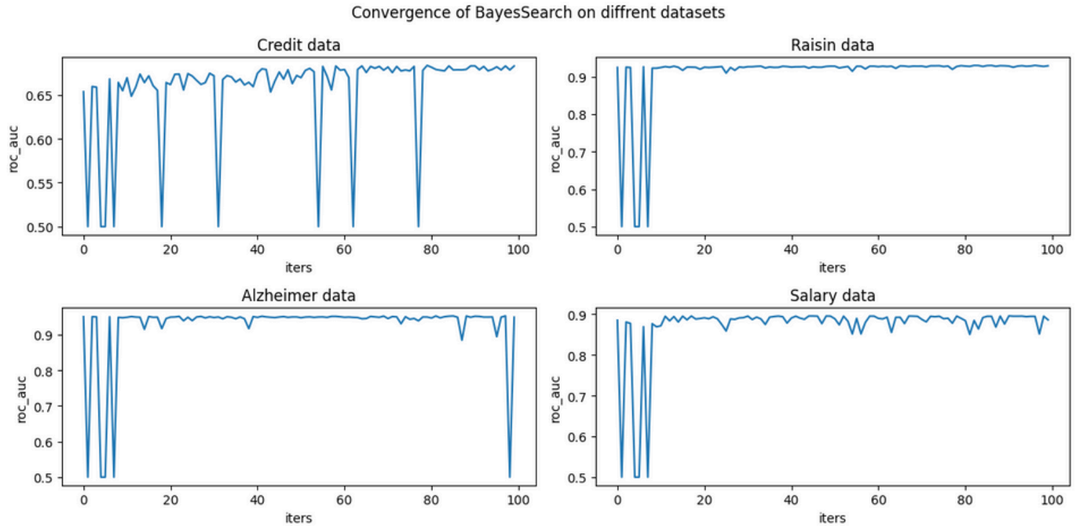
Wyniki

Zbieżność BayesSearch

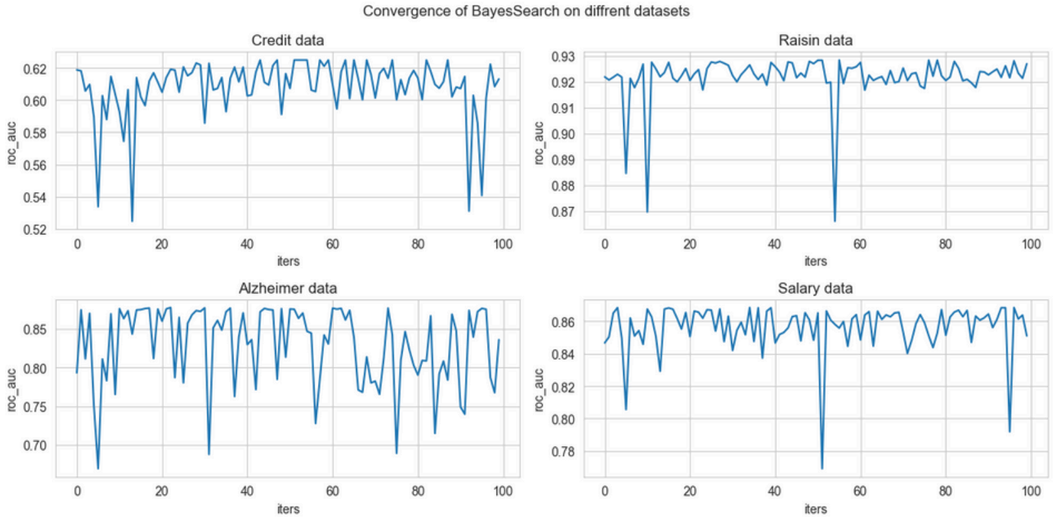
LR



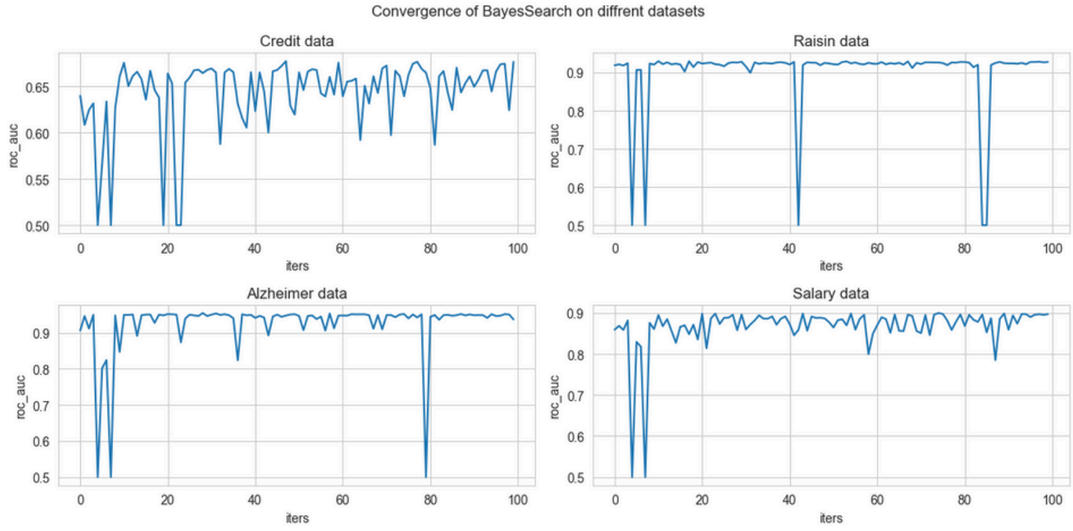
RF



KNN

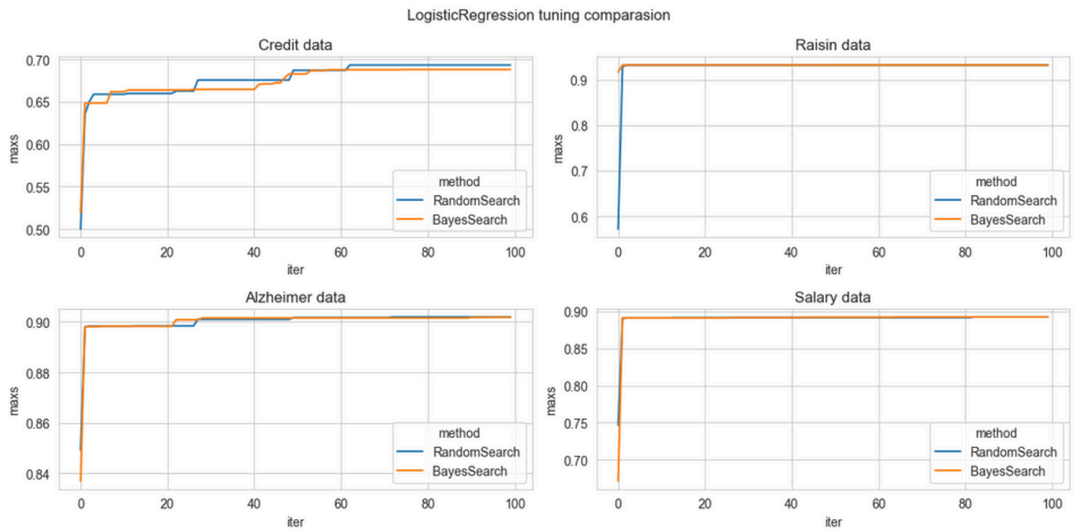


GBC

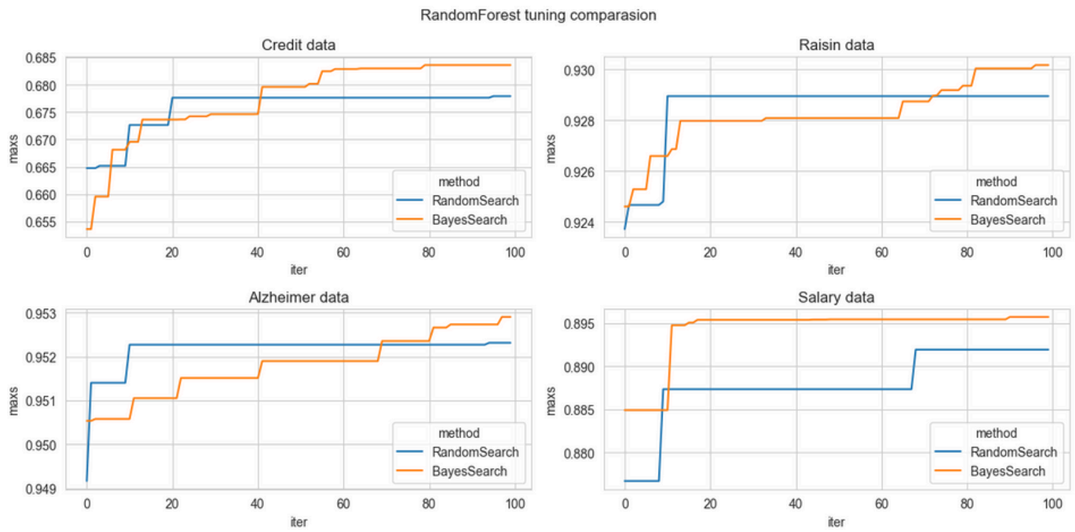


Ile iteracji?

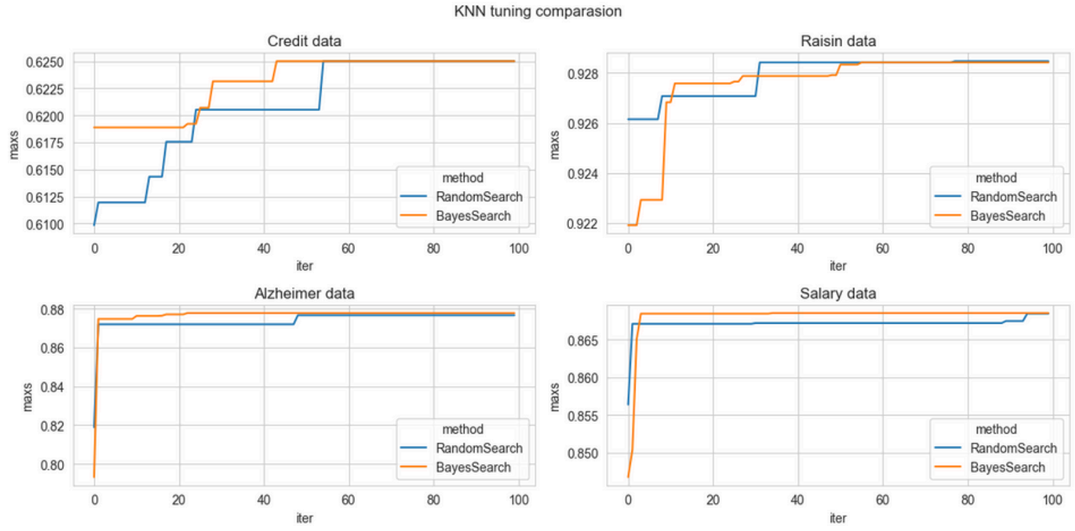
LR



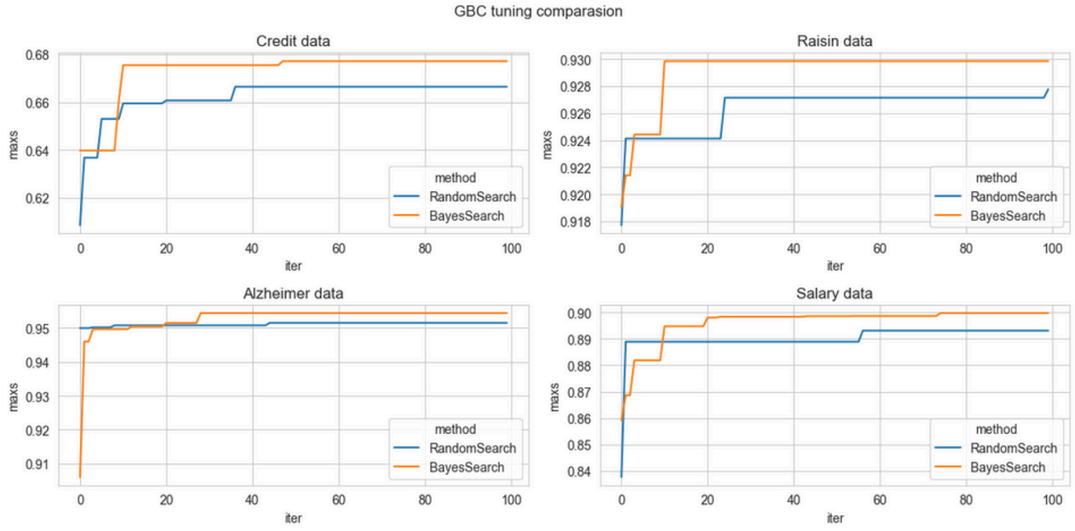
RF



KNN

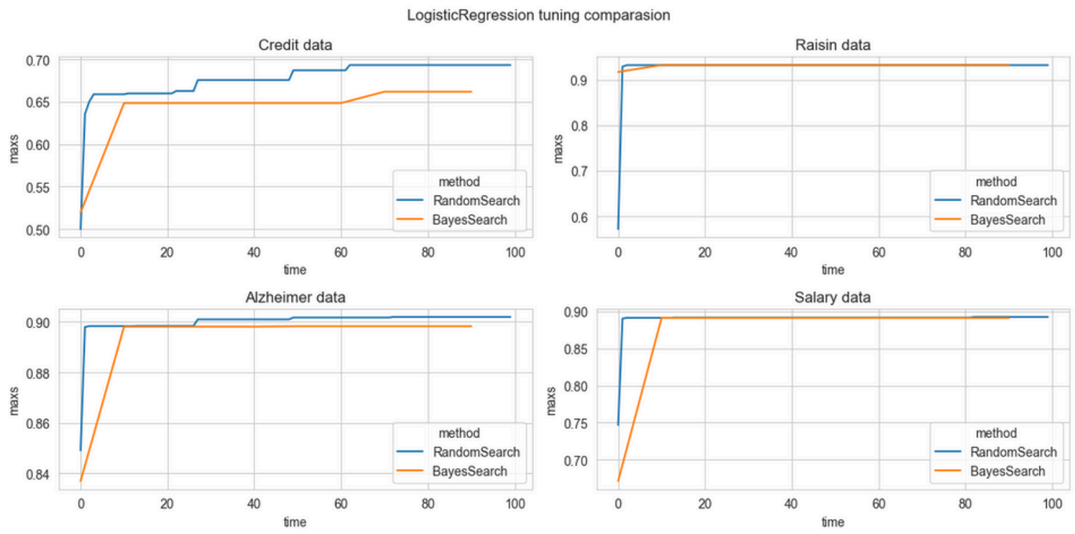


GBC

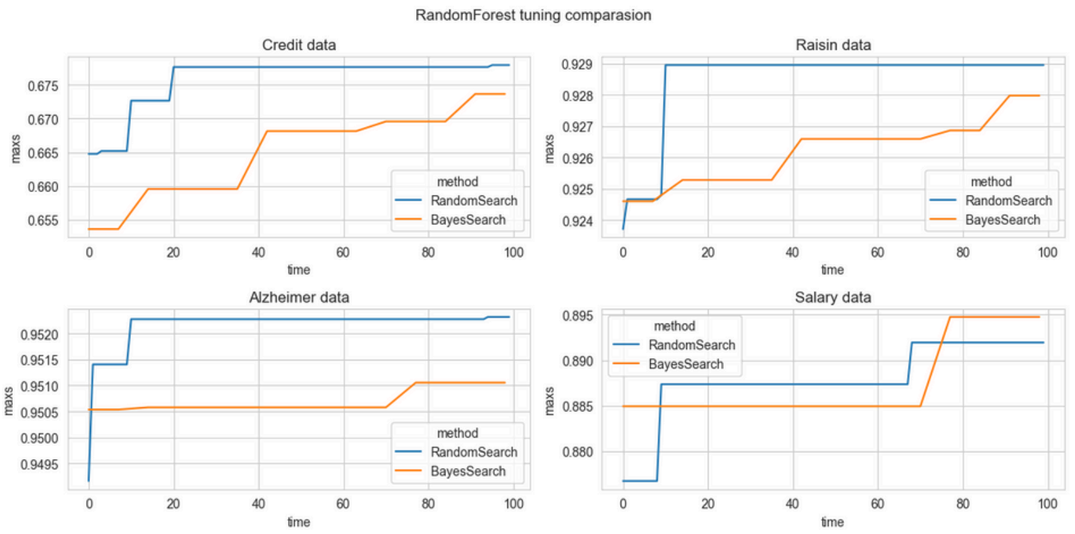


A ile czasu?

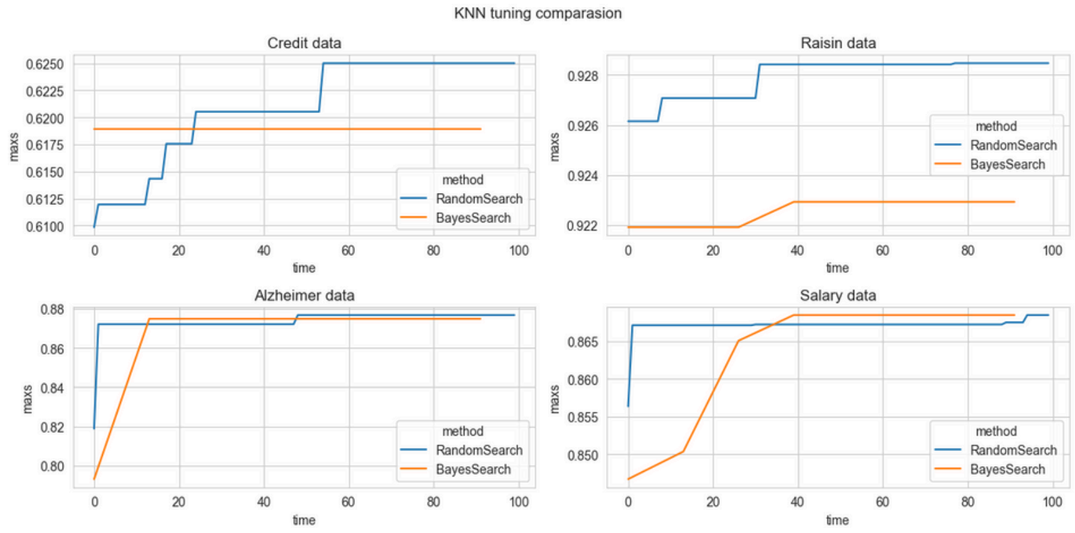
LR



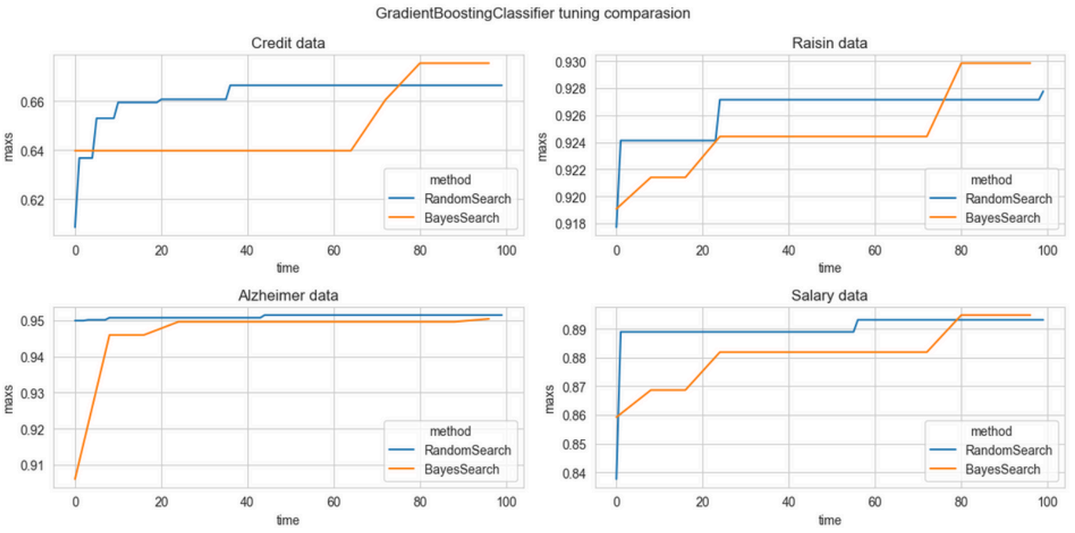
RF



KNN

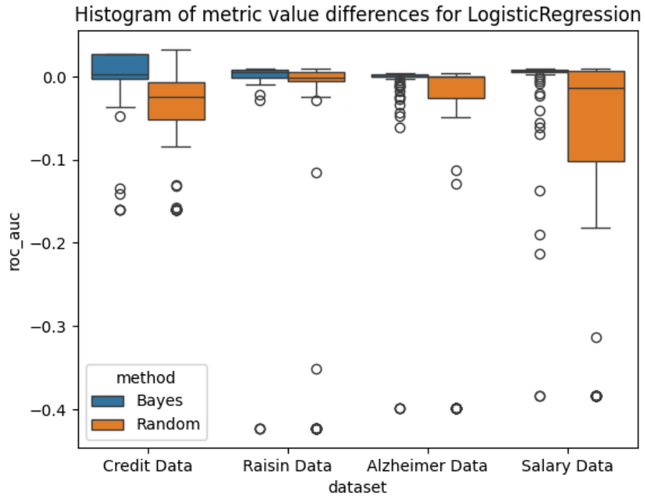


GBC

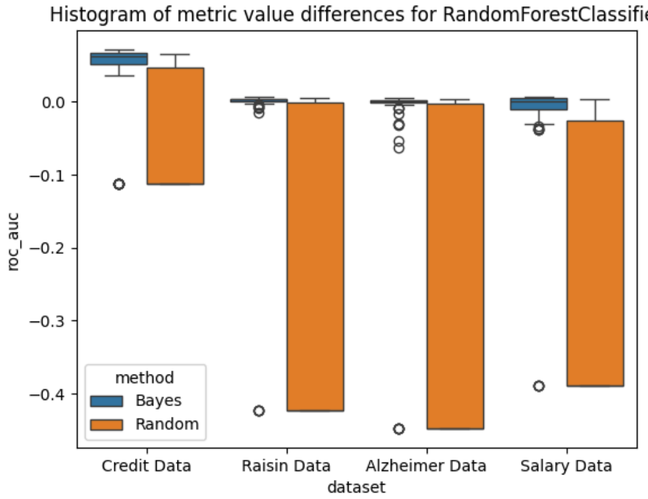


Rozkład wartości metryki

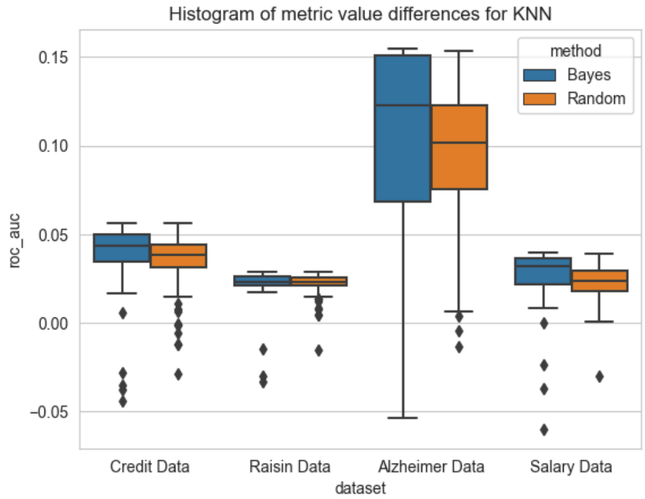
LR



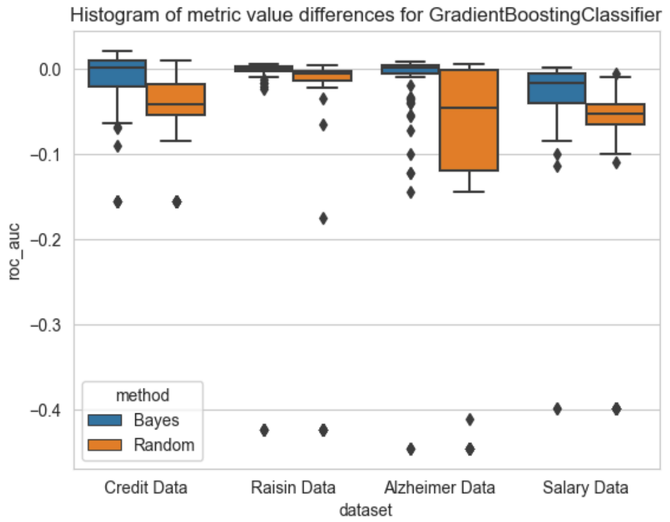
RF



KNN



GBC



Tunowalność