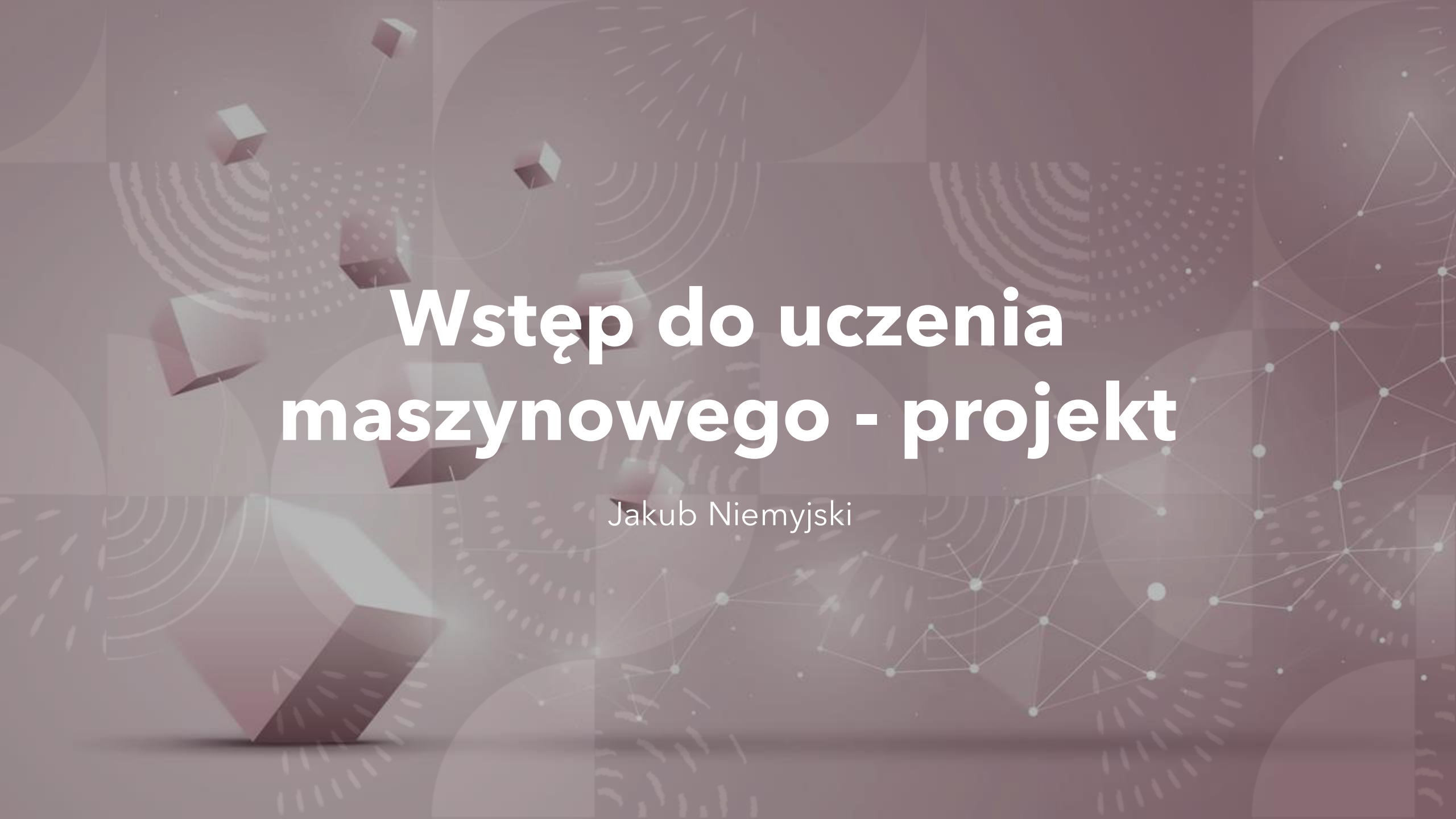
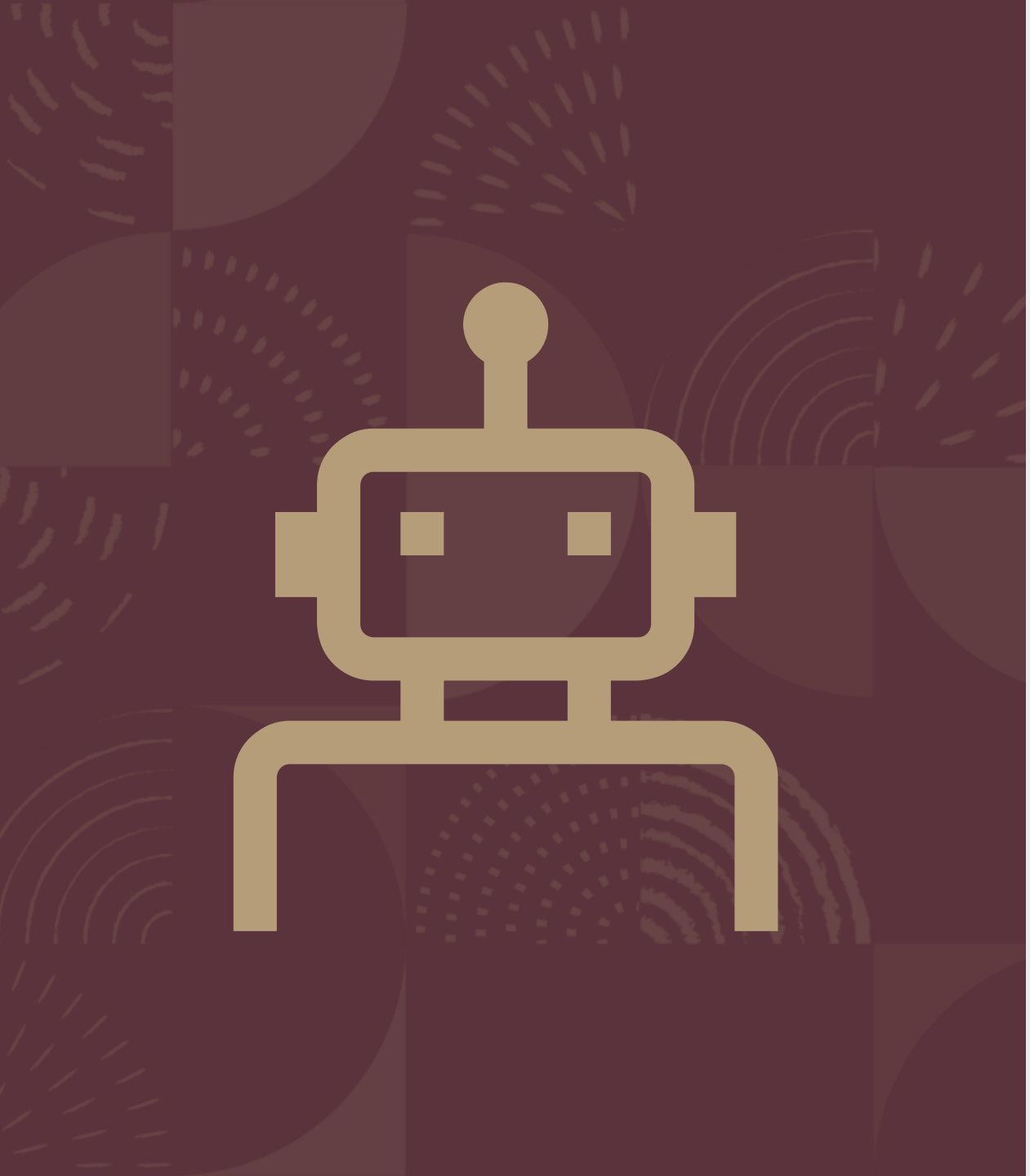


Wstęp do uczenia maszynowego - projekt

Jakub Niemyjski



Cel projektu

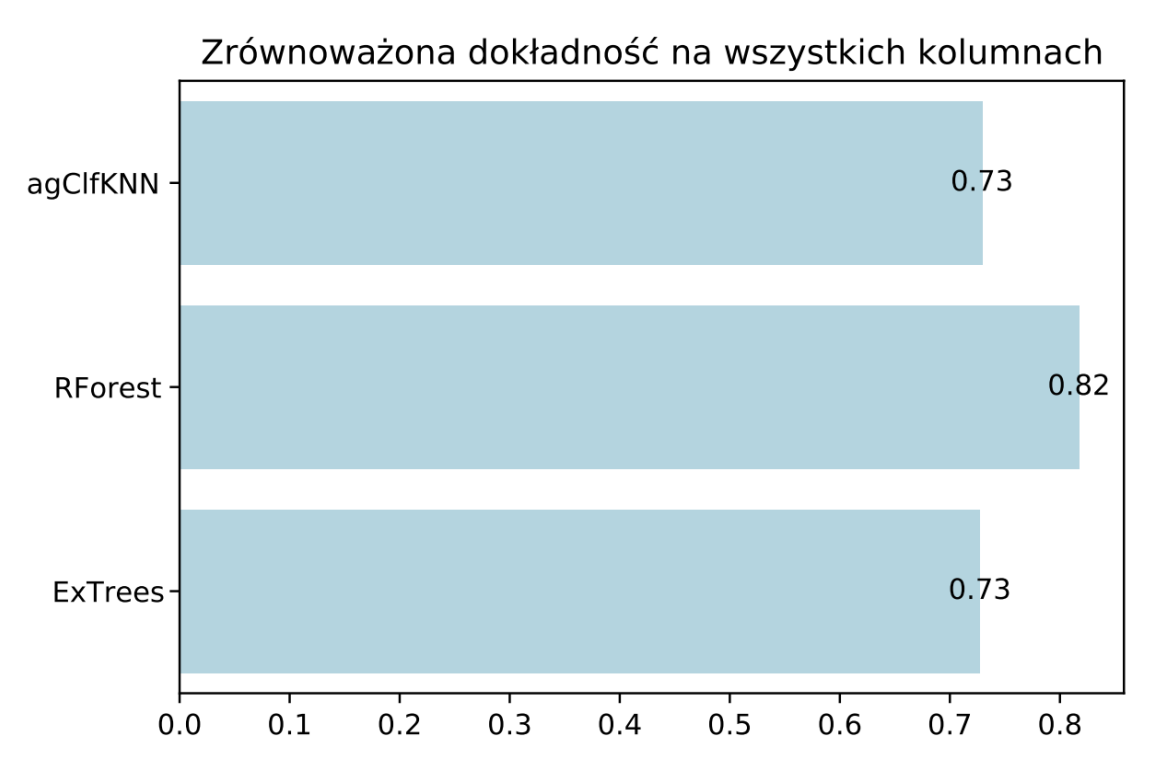
- Stworzenie modelu klasyfikacji binarnej o jak największej mocy predykcyjnej na przykładzie sztucznie wygenerowanych danych – maksymalizacja miary zrównoważonej dokładności

Dane

- 2000 obserwacji w zbiorze treningowym, 600 w zbiorze testowym; dysponujemy etykietami tylko dla zbioru treningowego
- 30 zmiennych objaśniających (część okazała się zbędna)

Dane

- Zbiór treningowy podzielony na część treningową i testową w proporcji 9:1
- Nie ma braków danych
- Niezauważalne obserwacje odstające
- Histogramy większości zmiennych przypominają rozkład normalny



Wstępna budowa modelu

- Tworzymy kilka modeli: *BaggingClassifier100KNN*, *RandomForestClassifier* oraz *ExtraTrees*
- Każdemu z nich za pomocą kroswalidacji dostosowujemy parametry
- Budujemy modele korzystając ze wszystkich zmiennych
- Wyniki dokładności na zbiorze testowym zostały przedstawione na wykresie

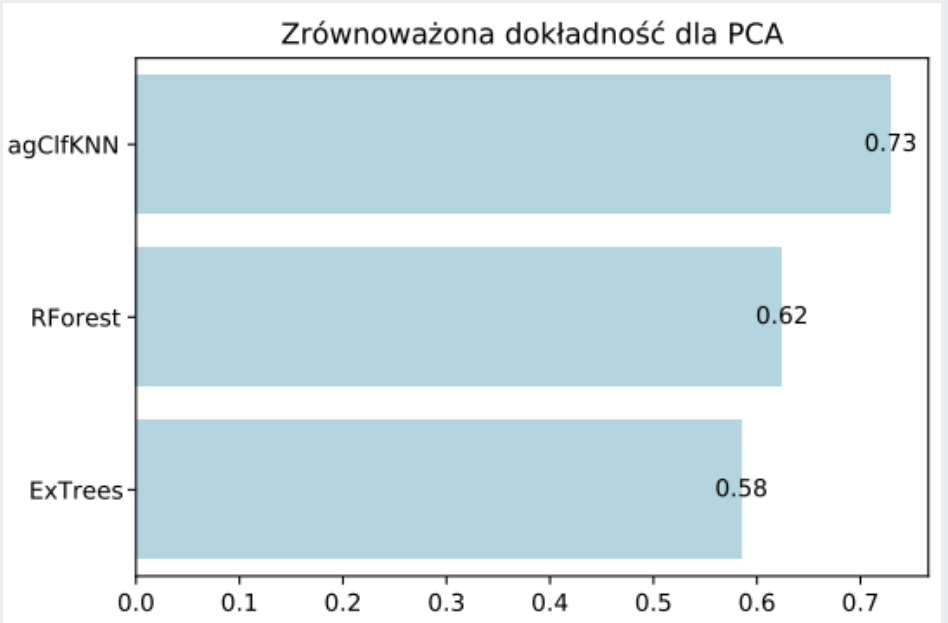
Redukcja wymiarowości

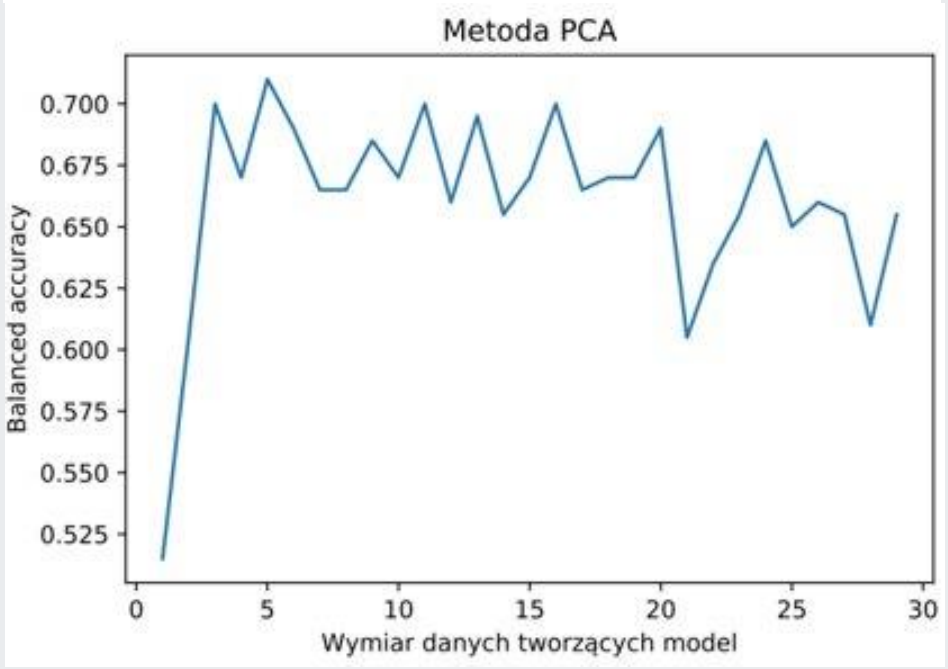
Chcemy sprawdzić, które zmienne są nieistotne dla modeli

Testujemy cztery metody redukcji wymiarowości: PCA, współczynnik informacji wzajemnej, Fisher score i regularyzację LASSO

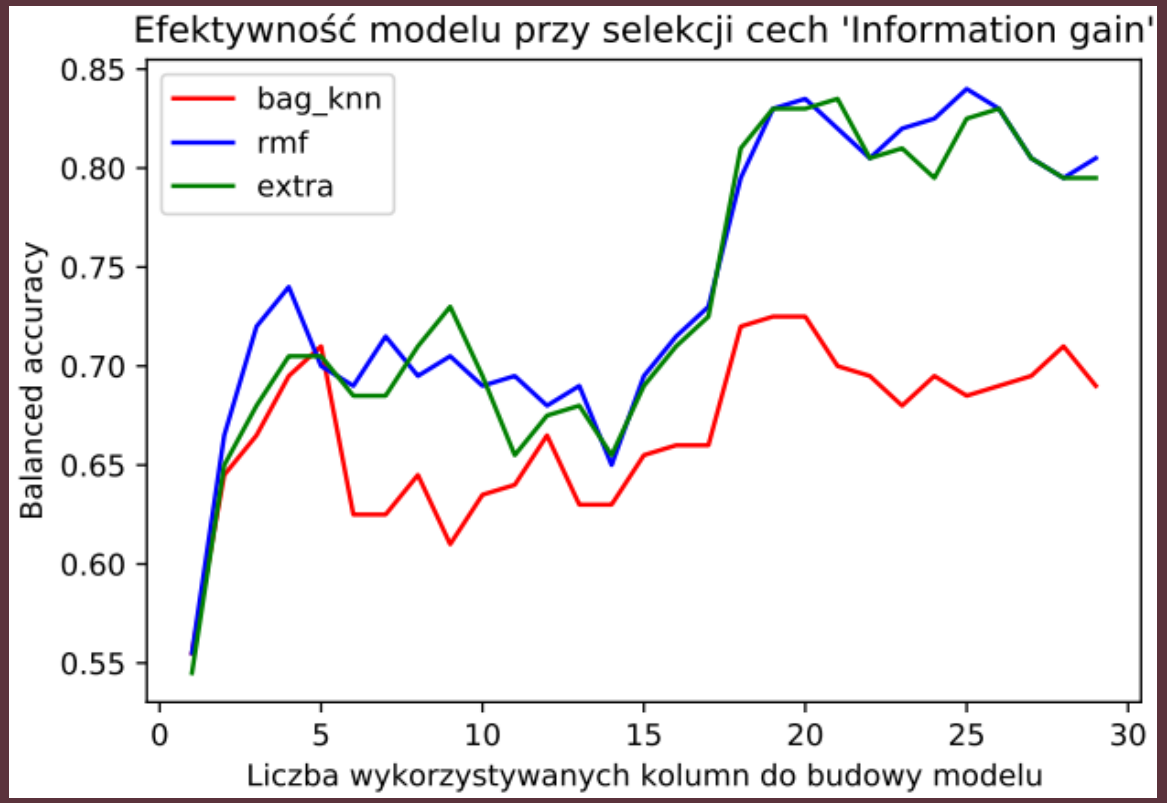
PCA

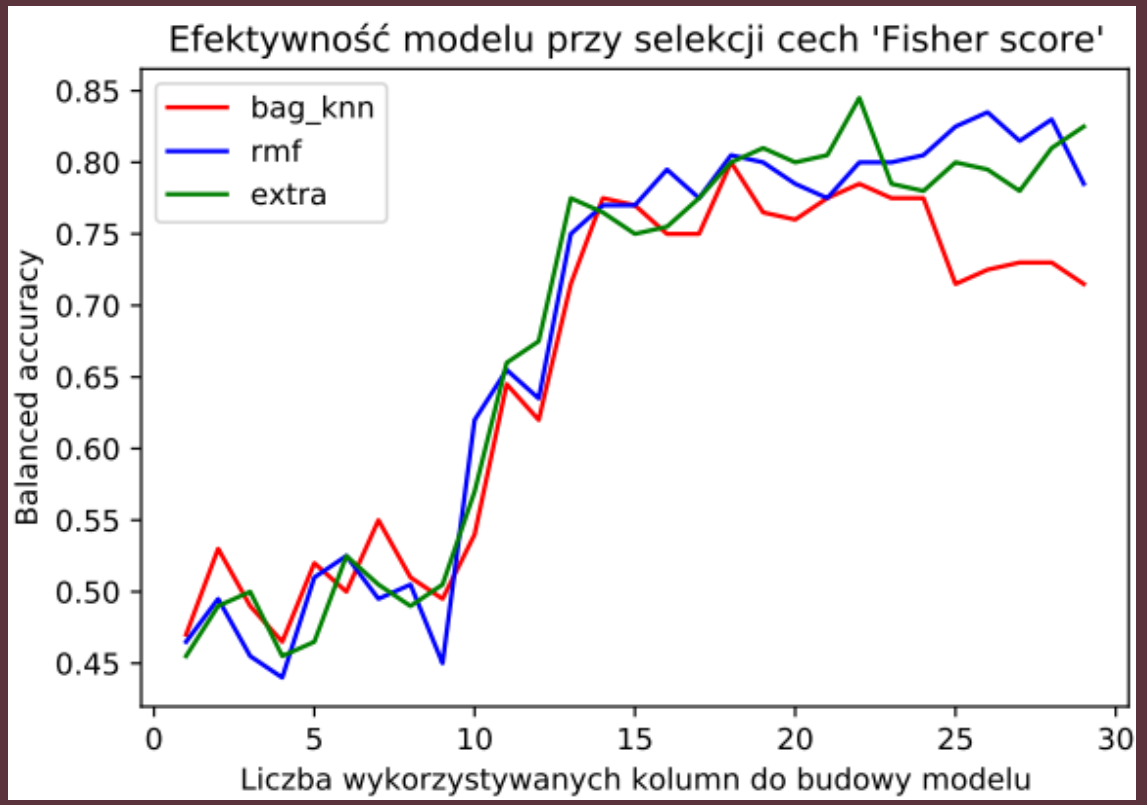
- Bez wcześniej przeskalowanych danych: eliminujemy 15 zmiennych
- Model zbudowany na pełnych danych jest lepszy na zbiorze testowym - odrzucamy metodę PCA





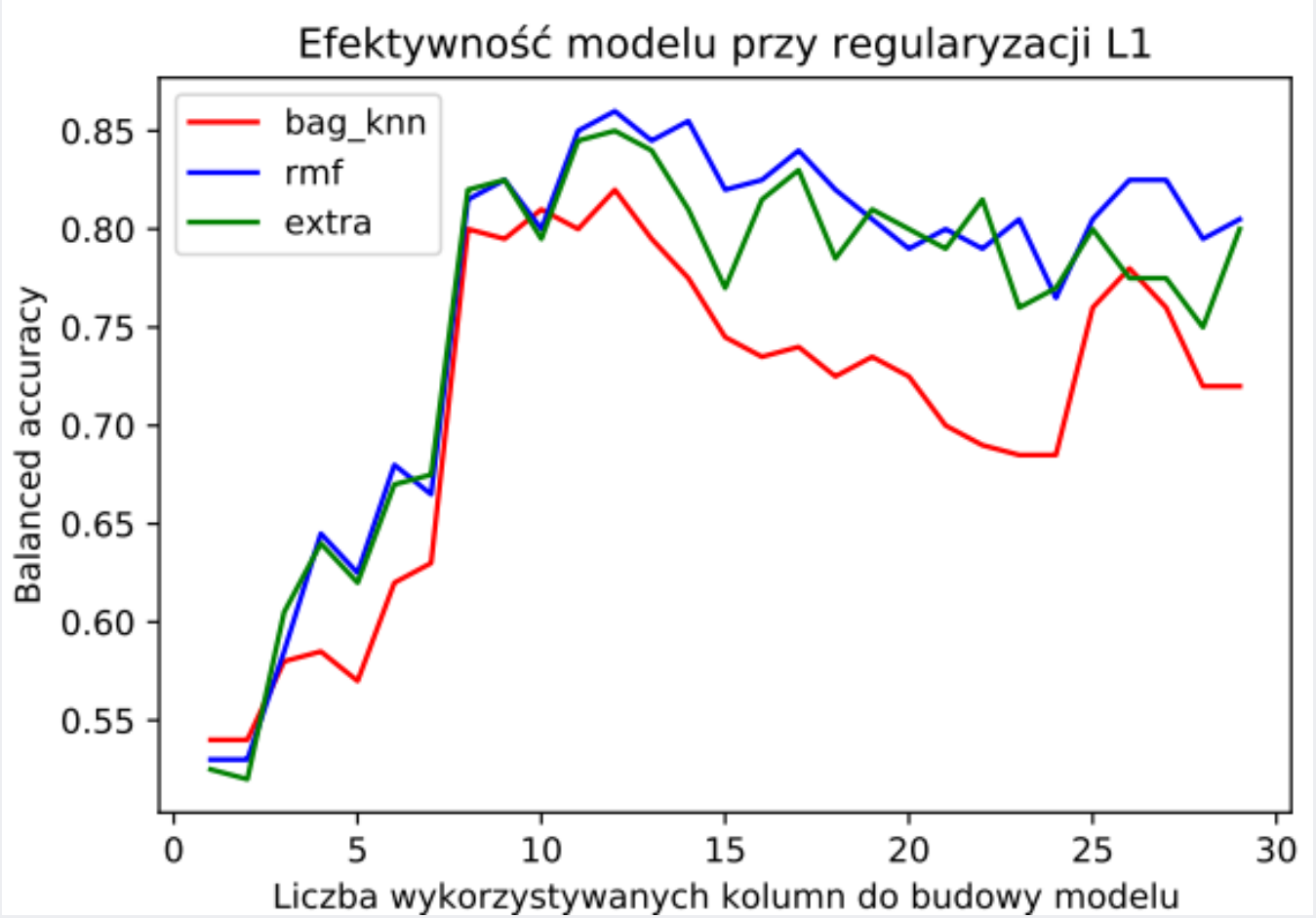
MI & Fisher score





Regularyzacja LASSO

Bardzo dobra moc predykcji dla zestawu 12 kolumn zbioru treningowego



Ostateczny wybór

- LASSO – Czy optymalizować parametry?
- Decyzja podjęta na podstawie miar efektywności dla modelu ze zoptymalizowanymi hiperparametrami i dla modelu z domyślnymi.

- Najlepszą metodą redukcji zmiennych okazało się LASSO
- Las losowy bardzo dobrze poradził sobie z zadaniem klasyfikacji

Podsumowanie


Koniec