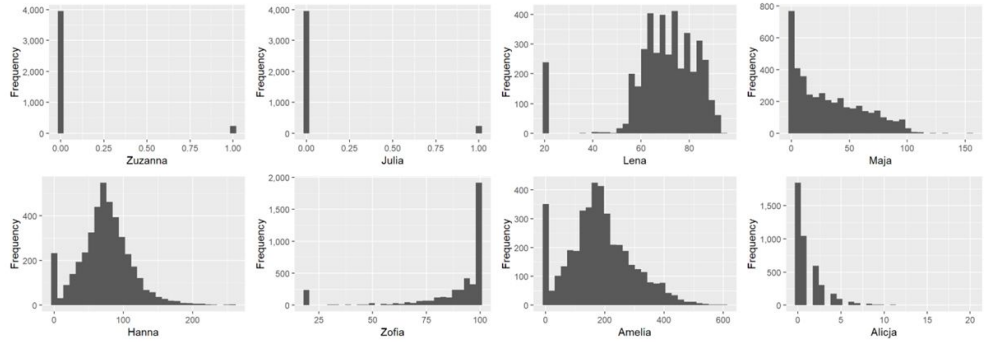
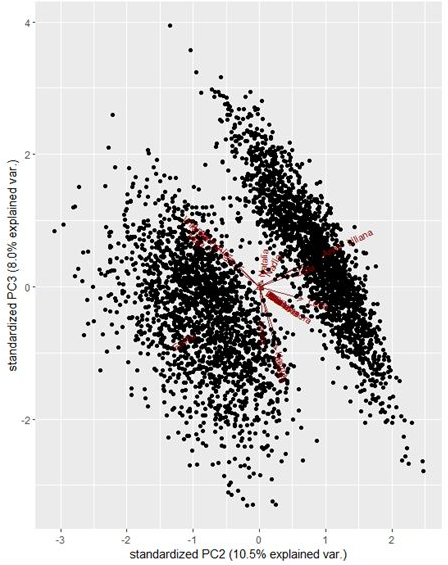
Warsztaty badawcze
Projekt

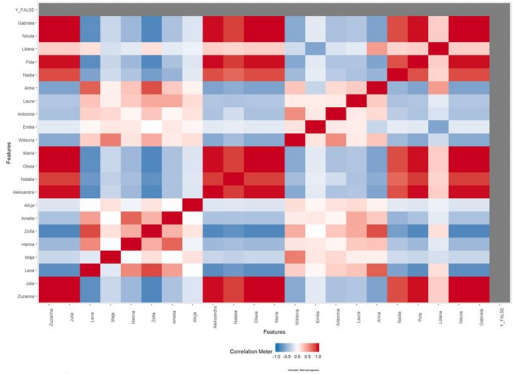
Bogdan Jastrzębski

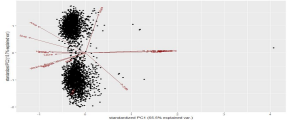
Przypomnienie

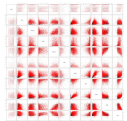
- Analiza naszego zbioru modeli
- Analiza dostępnego zbioru
  - Analiza rozkładów
  - Analiza korelacji
- PCA
- Próby klaseryzacji

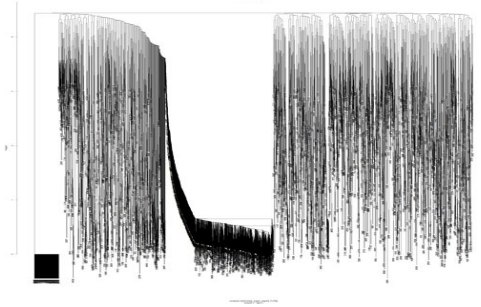












Predykcja

- Usunięcie niepotrzebnych zmiennych
- Przekształcenia logarytmiczne
- One-hot encoding
- Dodanie zmiennych o wystąpieniu zer w różnych kolumnach

Model: ranger

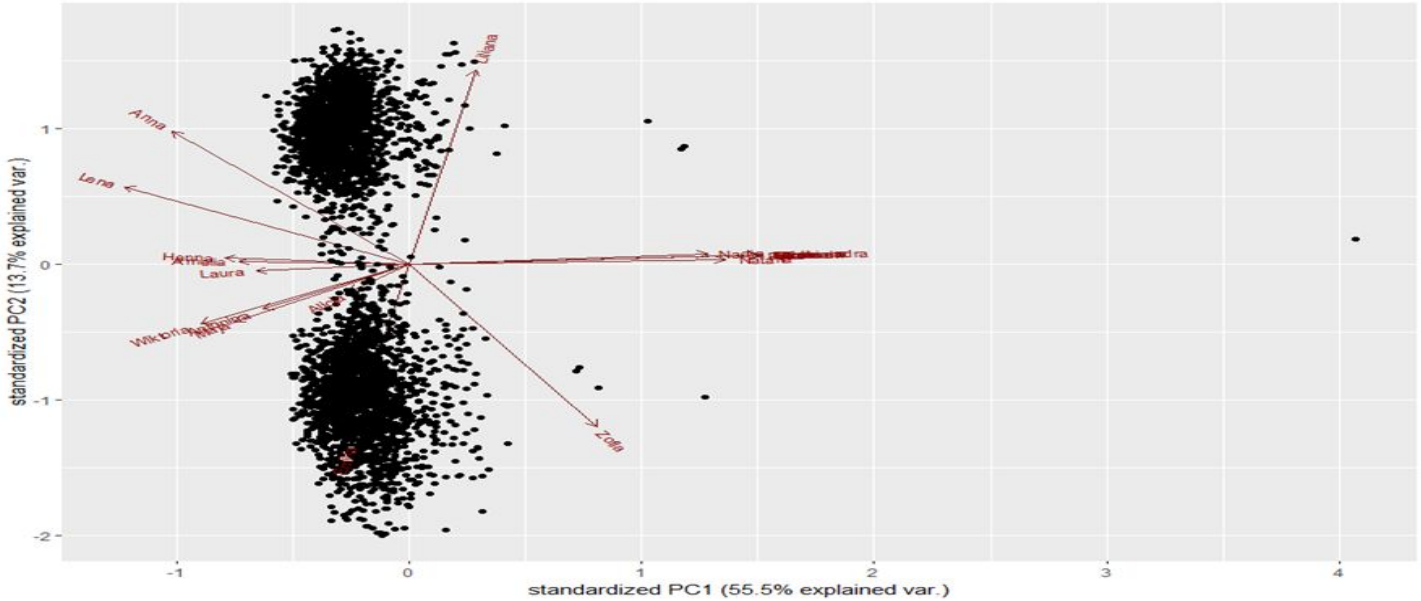
(z parametrami ustawionymi według pracy: https://arxiv.org/pdf/1802.09596.pdf)

Plan prezentacji

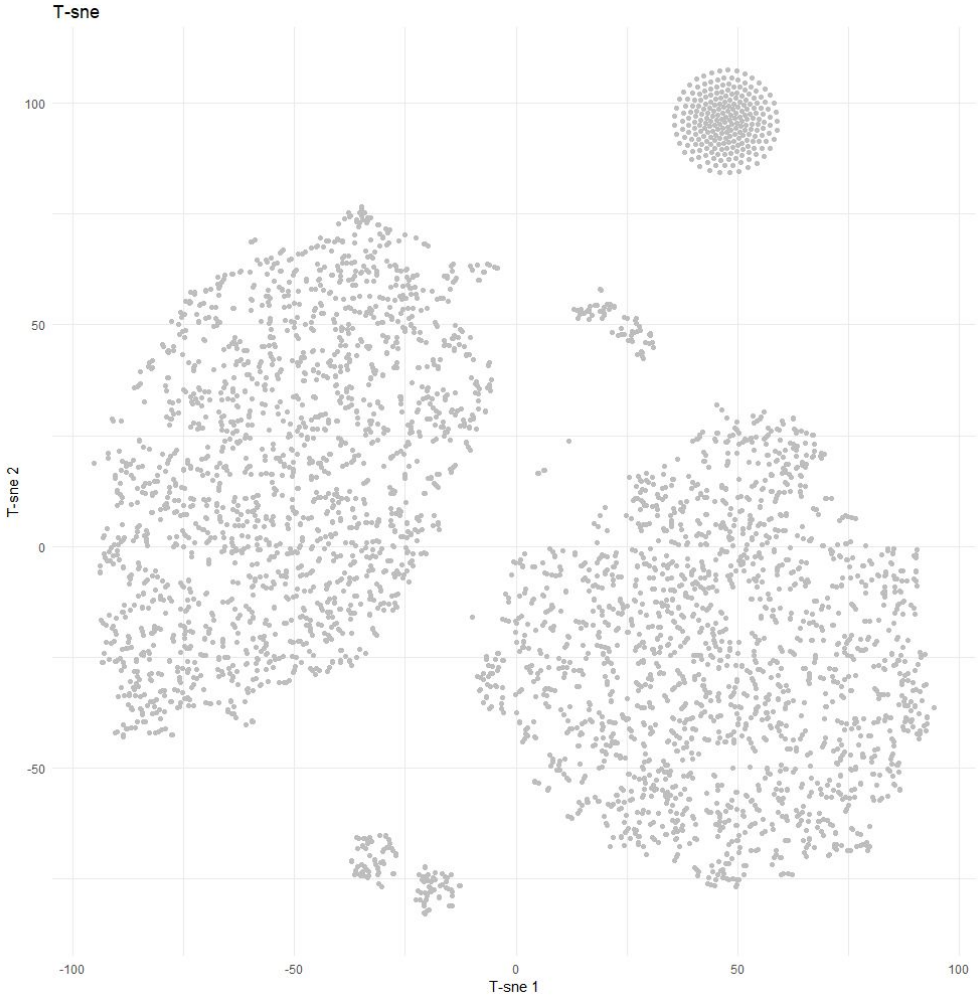
- T-sne zamiast PCA
- Klasteryzacja
- Spacerem z lasu losowego do lasu Bayesa...

PCA (poprzednia prezentacja)

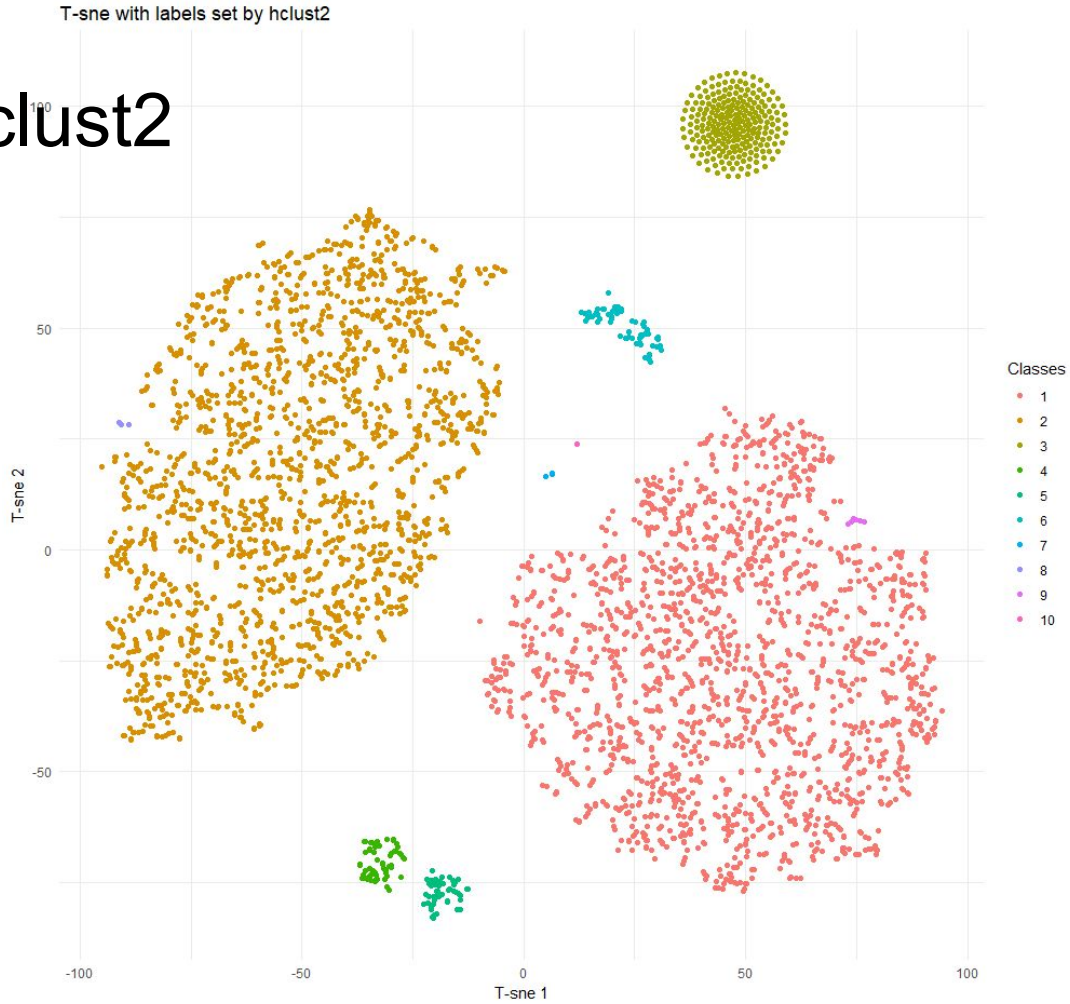
PCA2 to druga Liliana



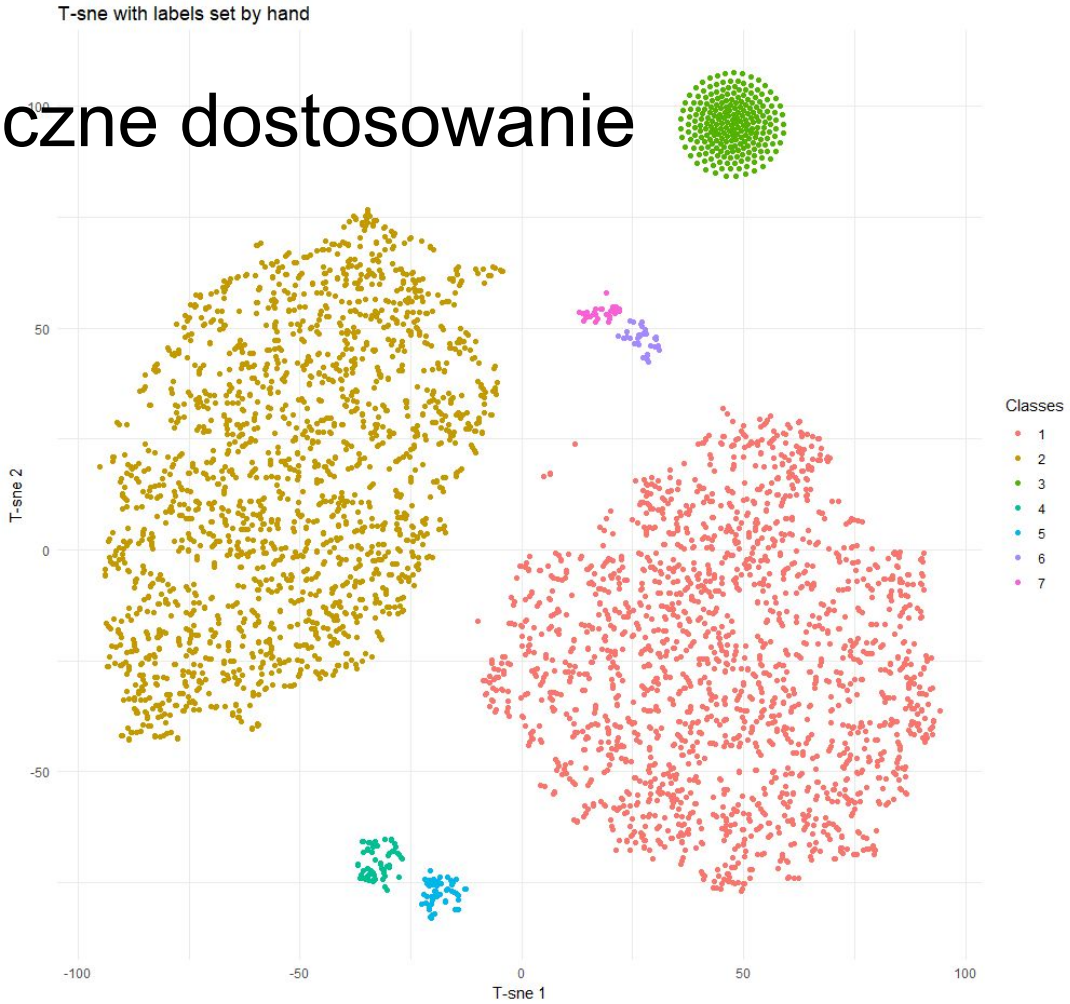
T-sne



T-sne + hclust2



T-sne + ręczne dostosowanie

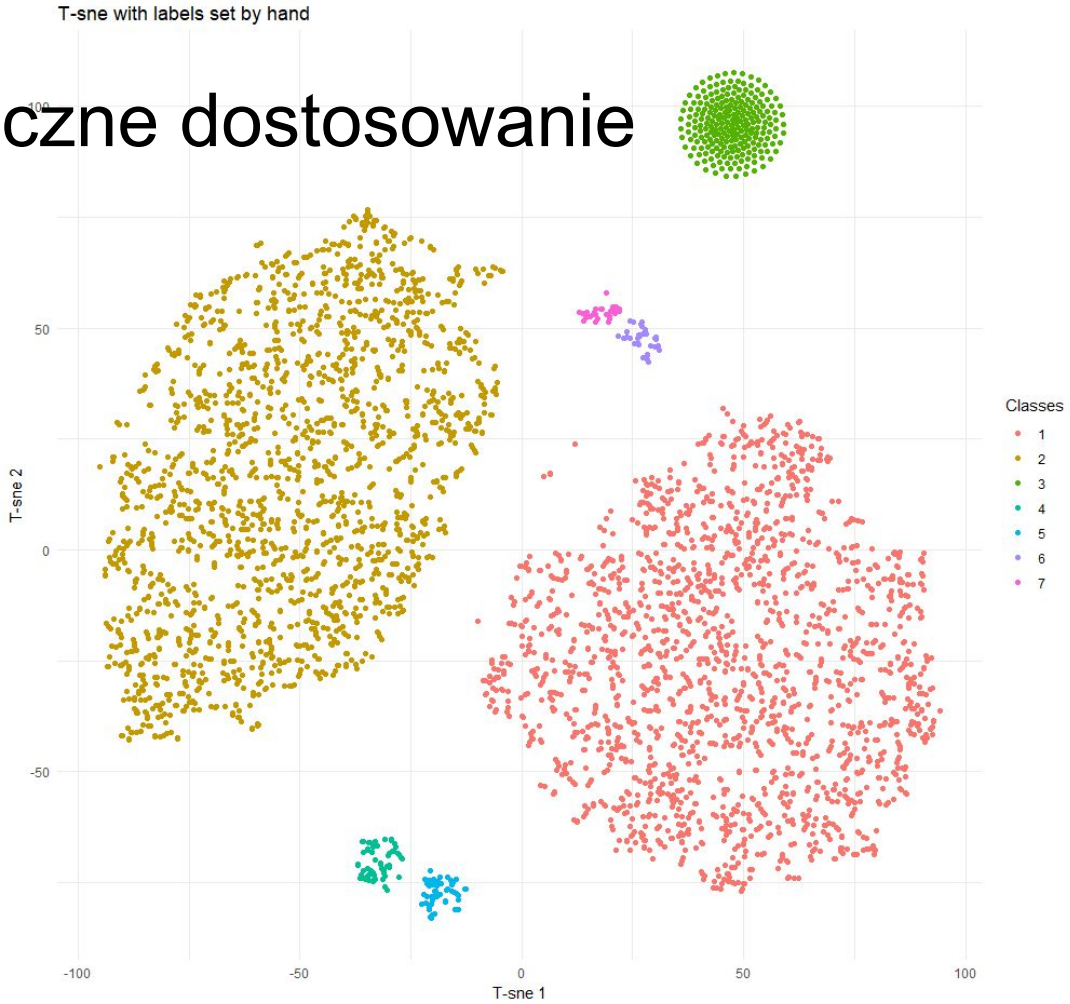


Centroidy

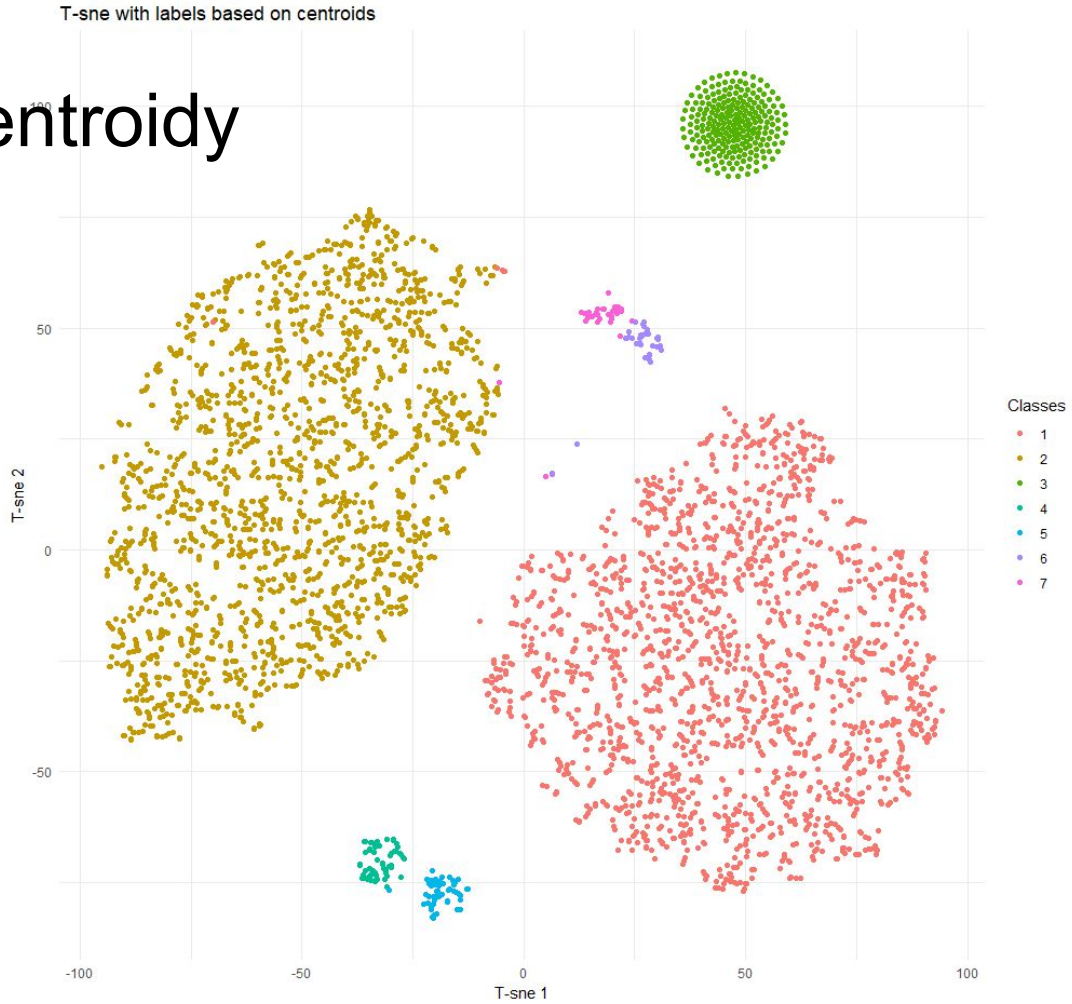
1. Podział na podzbiory **zbioru oryginalnego** według naszych etykiet
2. Centroid = średnia obserwacja w podzbiorze
3. Odległości od centroidów

Czy to ma sens?

T-sne + ręczne dostosowanie



T-sne + centroidy



Do mojej ramki dodaję kolumny:

- Odległości od centroidów

  (dla każdej nowej klasyfikowanej obserwacji liczymy najpierw odległości od centroidów)

Analiza naszych modeli - klasyfikacja

Schemat postępowania:

- Ekstrakcja cech modeli
- Klasyfikacja mojego datasetu dla wszystkich testowanych klasyfikatorów
- Analiza wyników

Ekstrakcja cech modeli

- Cechy takie jak liczba zmiennych, liczba zmiennych numerycznych etc.
- Powtórzenie piątej pracy domowej

Klasyfikacja dla różnych klasyfikatorów

- Klasyfikacja mojego datasetu dla wszystkich testowanych klasyfikatorów
- Czyli:
  - Dla każdego klasyfikatora
    - Do wiersza ze statystykami mojego datasetu dodaję kolumnę klasyfikator
    - Wykonuję predykcję

- Każda predykcja wykonana jest na obserwacjach, w których różni się tylko nazwa klasyfikatora

Analiza wyników

- Przy uwzględnieniu tego, że pewne modele nie występowały we wszystkich zbiorach danych…
- Najlepsze modele:
  - Knn / Modele drzewiaste:
    - RRF, rotationForest, h2o.randomForest, cforest, bartMachine...
    - KNeighborsClassifier, IBk
  - Bardzo dobry bilans (przewidywane ACC - średnie ACC) miały metody bayesowskie

Pewnego razu spacerując przez las losowy...

… trafiłem do magicznego lasu Bayesa…

Rozdział pierwszy: Lista algorytmów mlr

- https://mlr.mlr-org.com/articles/tutorial/integrated_learners.html

BART machine

(to nie to) →

- “Bayesian Additive Regression Trees”
- “Gradient Boosting Trees” + wnioskowanie bayesowskie?

Dziękuję za uwagę