Metody klasyfikacji - projekt 2

Anna Kozioł

Wydział Matematyki i Nauk Informacyjnych
Politechnika Warszawska

Data Mining 2019

9 czerwca 2019

Plan prezentacji

1. Wstępna analiza danych
2. Selekcja zmiennych
3. Wybór i opis użytych metod klasyfikacji
4. Podsumowanie

Cel

Celem projektu było praktyczne sprawdzenie metod klasyfikacji na podstawie 2000 obserwacji, należących do klas, oznaczonych odpowiednio 0 i 1.

Wstępna analiza danych

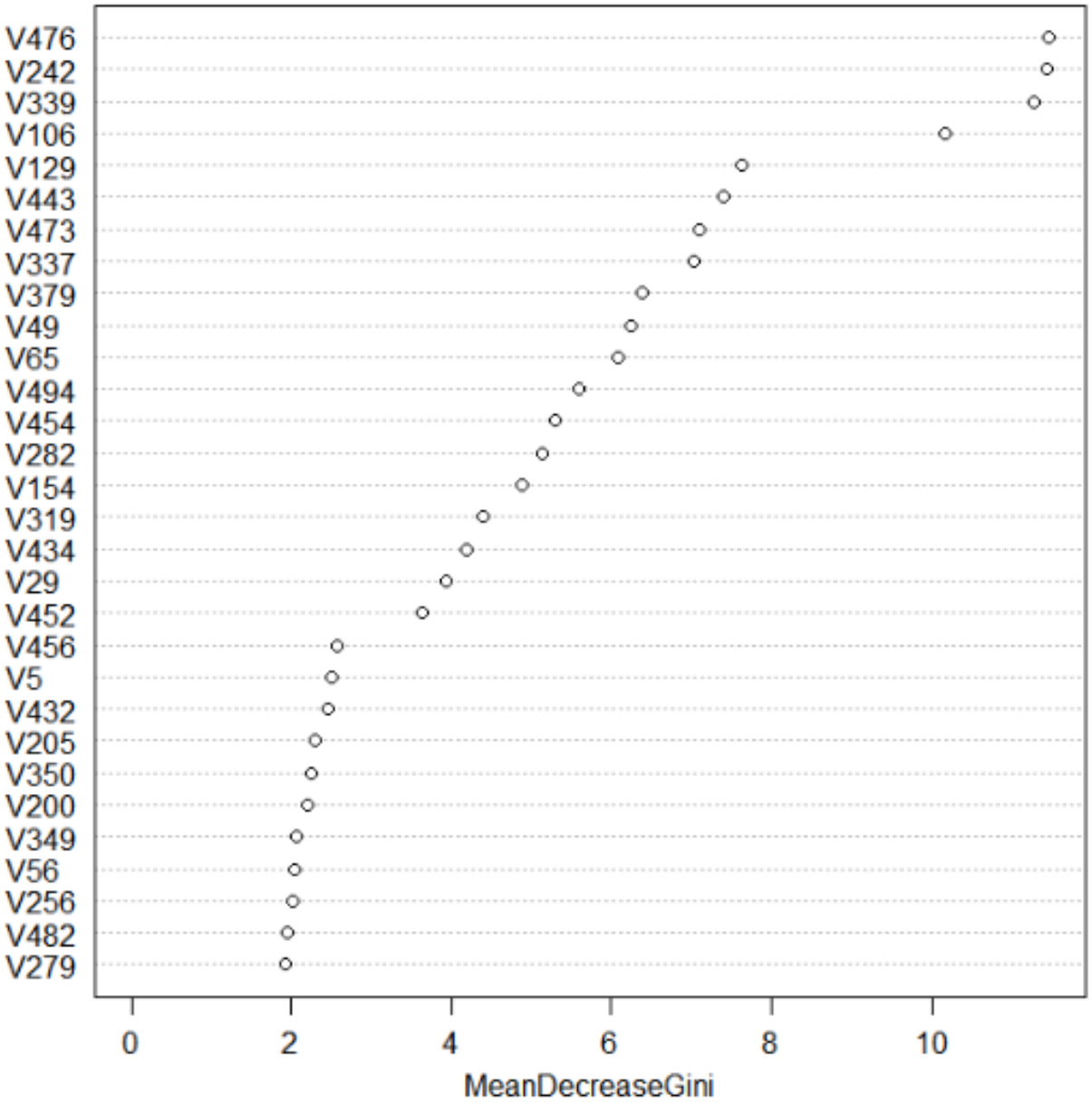
- brak braków danych
- brak par obserwacji o takich samych wartościach oraz zmiennej objaśniającej przyjmującej stałą wartość
- 12 zmiennych objaśniających z dużym wskaźnikiem vif (>10)
- za pomocą odległości Cooka, można zweryfikować w zbiorze 30 obserwacji odstających
- klasy równoliczne

Zbiór danych został podzielony losowo na część treningową(80%) i testową (20%). Część treningowa służyła do nauczenia klasyfikatora. Część testowa była używana wyłącznie do celów oceny jakości klasyfikacji.

Selekcja zmiennych

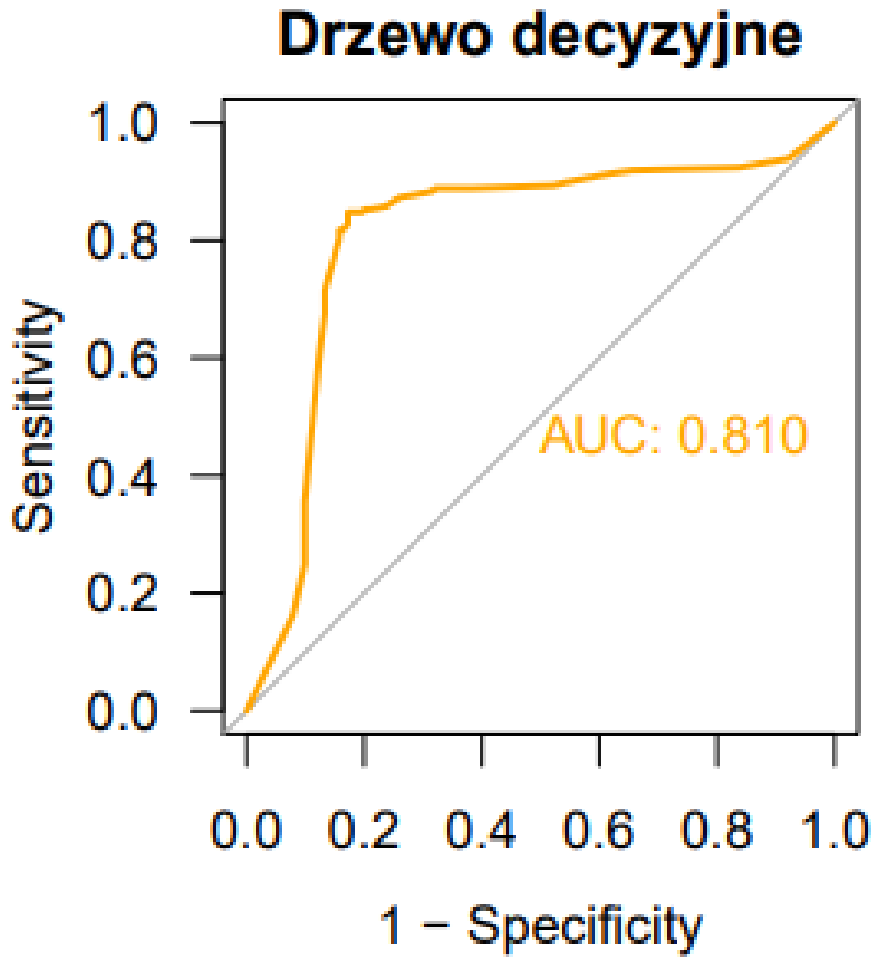
1. Boruta z pakietu 'Boruta'
   V29 V49 V65 V106 V129 V154 V242 V282 V319 V337 V339 V379 V434 V443 V452 V454 V473 V476 V494 V456
2. Metoda selekcji CMIM z pakietu "praznik"
   V106 V494 V339 V337 V65 V154 V443 V454 V476 V456 V242 V434 V379 V67 V79 V398 V170 V340 V444 V473
3. importance() z lasu losowego zbudowanego na wszystkich danych

Selekcja zmiennych



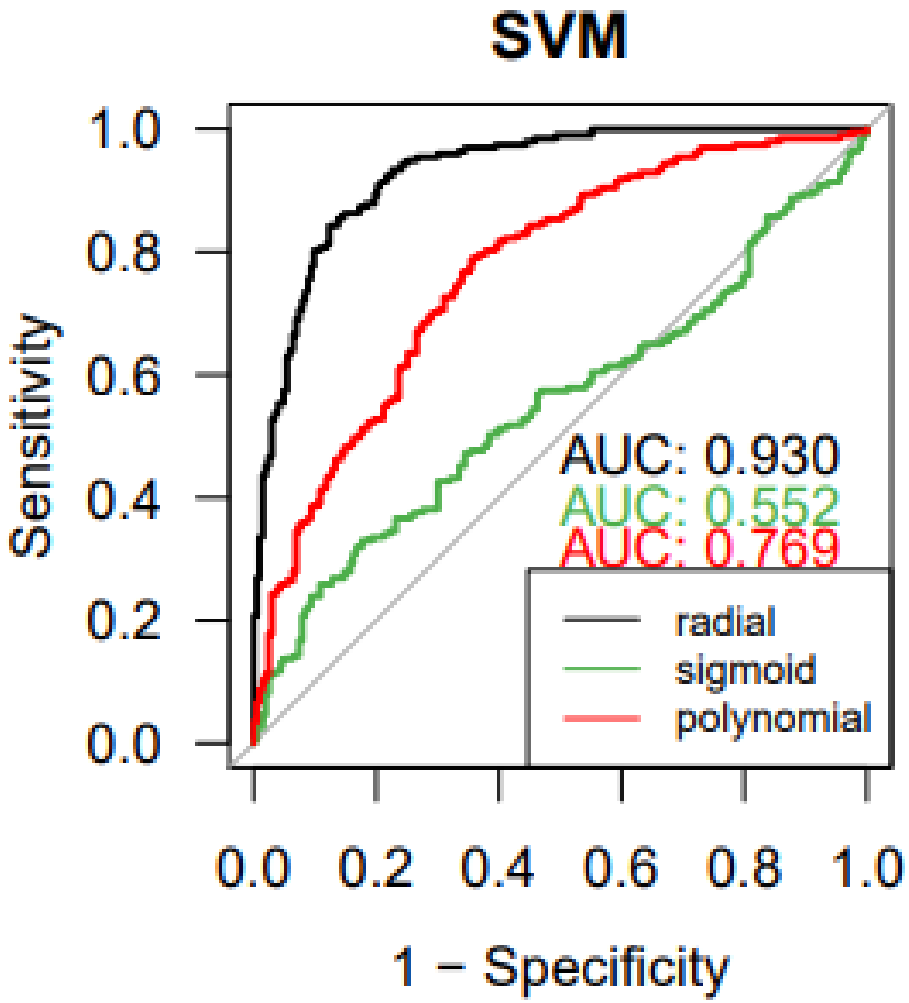
Drzewo decyzyjne

- Drzewo wygenerowane jest z parametrami minsplit=5 i cp= 0.001, zmiana tych parametrów nie wpływa znacząco na wskaźnik AUC jakości klasyfikacji.



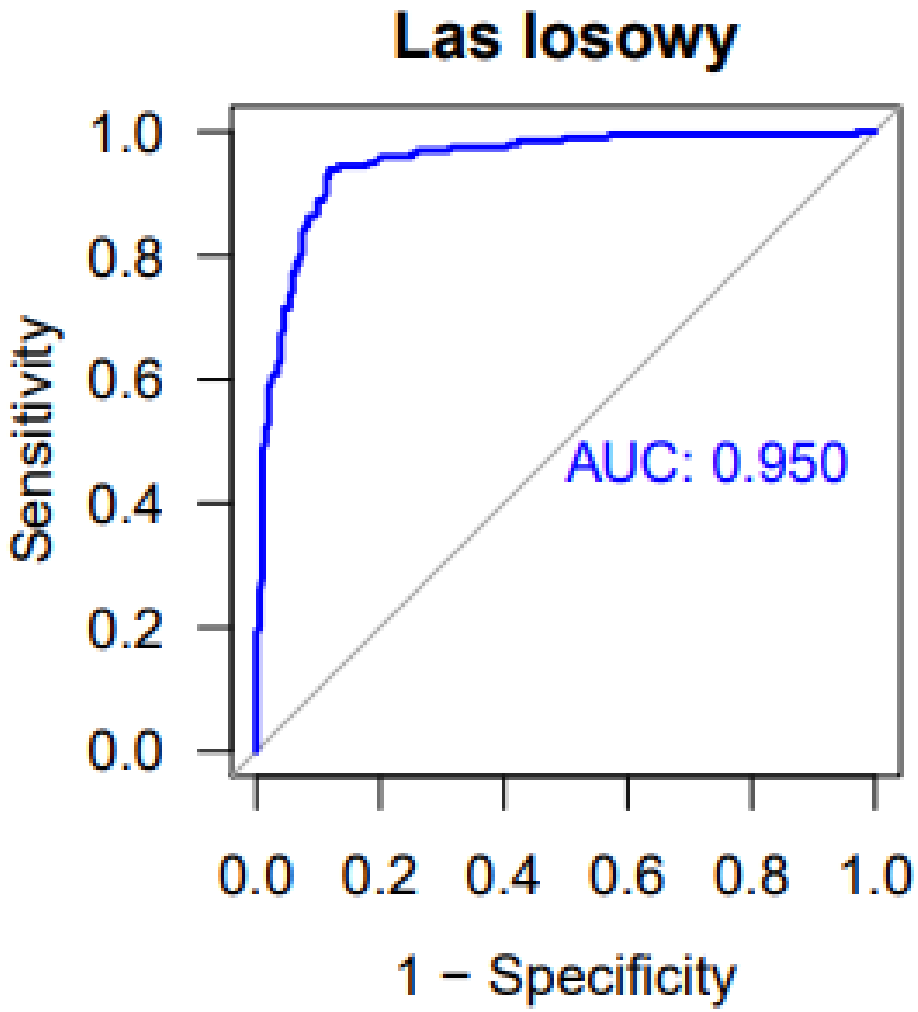
Maszyny wektorów nośnych

- W tym przypadku widać wyraźne różnice jakości klasyfikacji w zależności od przyjętego jądra. Najlepsze wyniki daje jądro radialne.



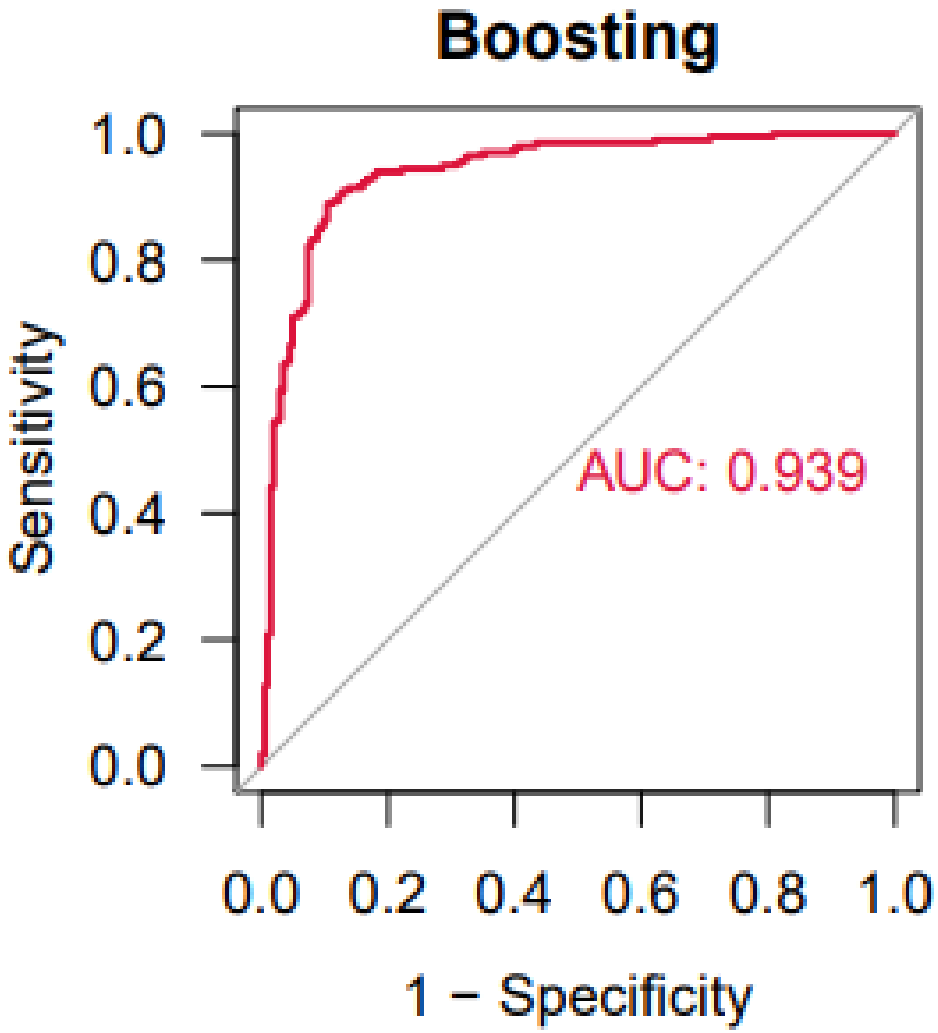
Las losowy

- Las losowy został wygenerwoany z parametrem ntree=500.Zmiana tego parametru nie powodowała poprawy klasyfikacji.

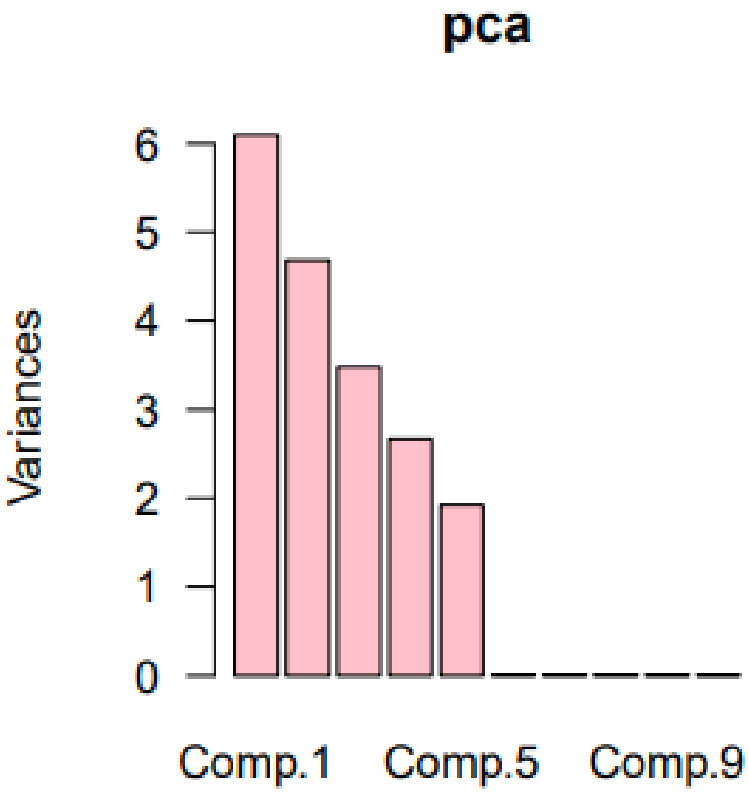


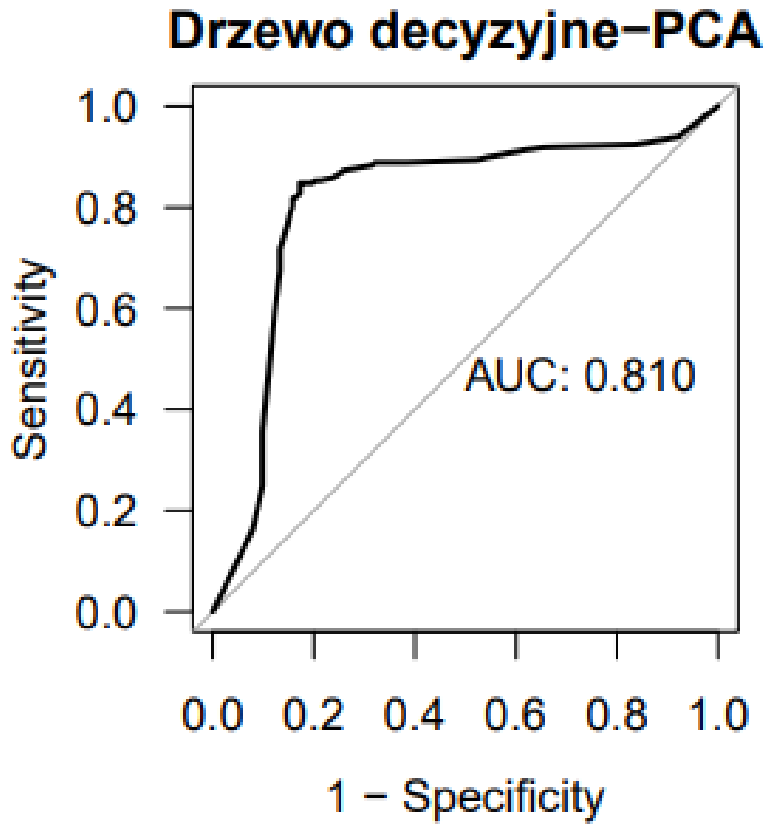
AdaBoost

- AdaBoost został użyty z parametrami mfinal=1000 oraz boos=T



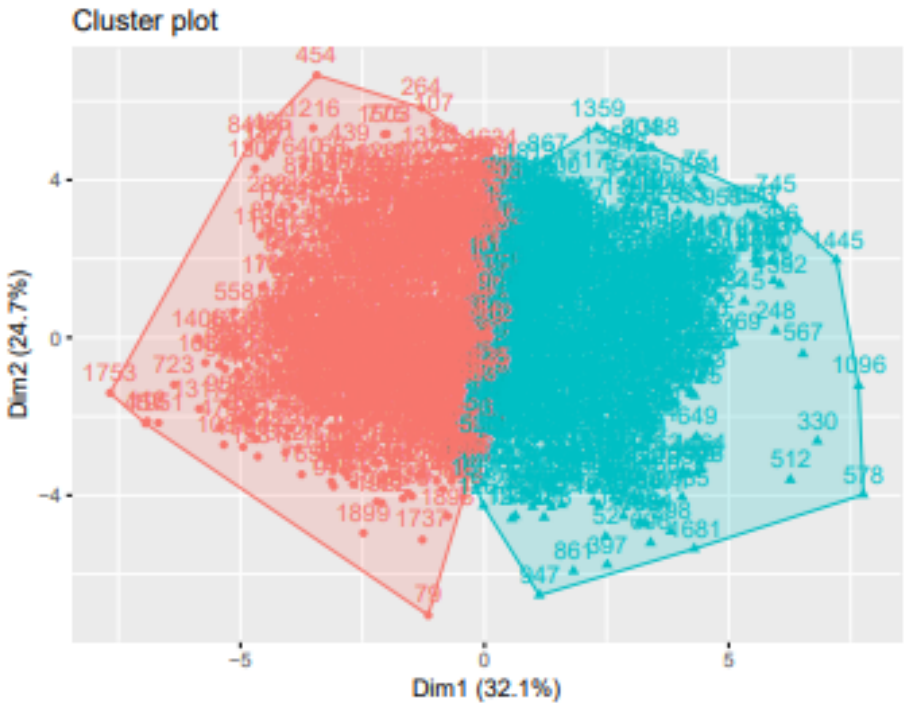
PCA i drzewo oparte na składowych głównych

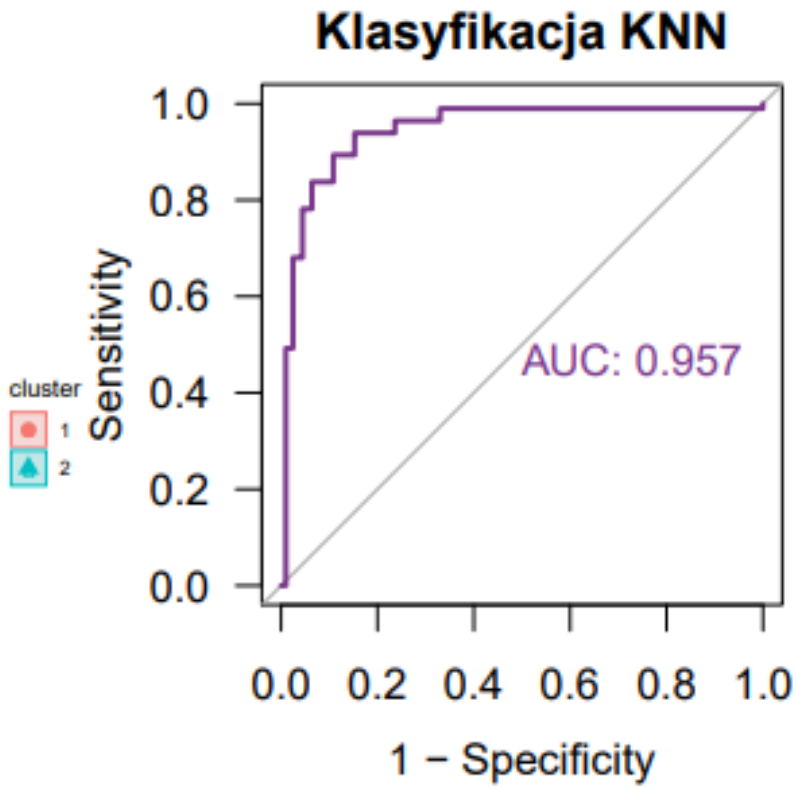




KNN i klasyfikacja KNN

- Zmienne zostały przeskalowane





Podsumowanie

Selekcja zmiennych została wykonana przy użyciu Boruty, CMIM, Lasów losowych. Przetestowane zostały metody klasyfikacyjne: Drzewo decyzyjne, Lasy losowe, AdaBoost, SVM, KNN oraz drzewo decyzyjne na zmiennych zmienionych przez PCA z różnymi parametrami. Najlepsze rezultaty klasyfikacji dała metoda KNN