Projekt 2 - raport

Anna Kozioł

7 czerwca 2019

Celem analizy jest sprawdzenie metod klasyfikacji na podstawie danych zawartych w pliku "train_projekt2.txt". W pliku zawartych jest 2000 obserwacji, należących do jednej z dwóch klas, oznaczonych odpowiednio 0 i 1. W raporcie przedstawiona zostanie analiza zbioru danych, identyfikacja istotnych zmiennych,opis testowanych metod klasyfikacji oraz wybranie ostatecznej.

Wstępna analiza danych

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

Zbiór danych dzielę na część treningową(80%) i testową (20%). Część treningowa posłuży do nauczenia klasyfikatora. Część testową używać będę wyłącznie do do celów oceny jakości klasyfikacji.

Selekcja zmiennych

- Boruta z pakietu "Boruta"

V29 V49 V65 V106 V129 V154 V242 V282 V319 V337 V339 V379 V434 V443 V452 V454 V473 V476 V494 V456 Algorytm wybrał, 20 istotnych zmiennych, 480 pozostałych zostało odrzucone.

- 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

Przy deklaracji wyboru 20 zmiennych wyniki w znacznej części pokrywają.

- Dodatkowo buduję las losowy i z niego wybieram ważne zmienne(importance())

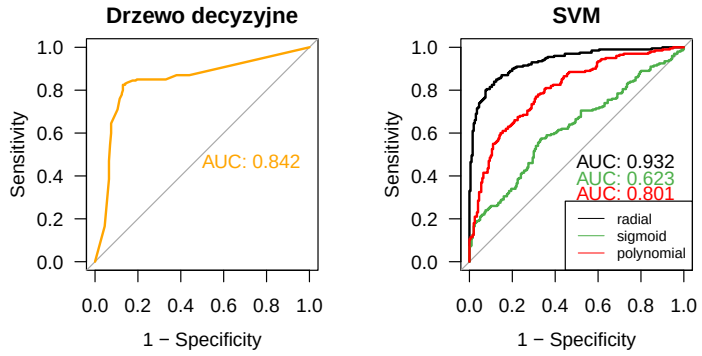
W dalszej części raportu wyniki są oparte na zmiennych wybranych przez Borutę, gdyż dały najlepszą jakość klasyfikacji i wystąpiły w większości wyników innych algorytmów.

Drzewo decyzyjne

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

Maszyny wektorów nośnych

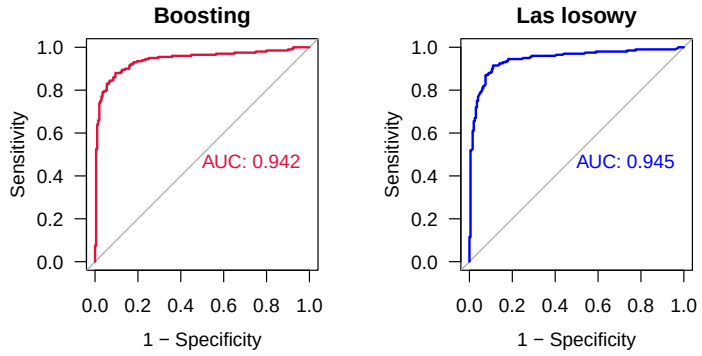
- Następnym sposobem klasyfikacji jest SVM. W tym przypadku widać wyraźne różnice klasyfikacji w zależności od przyjętego jądra.



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

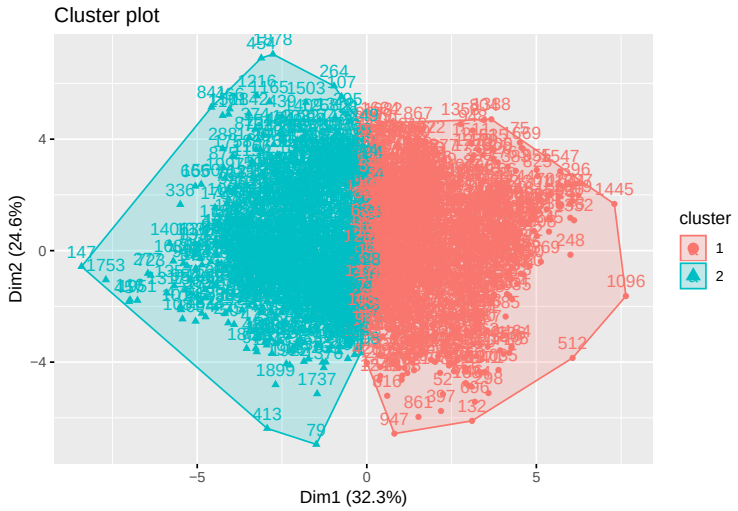
ADA BOOST

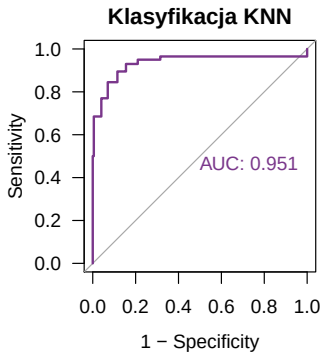
AdaBoost został, użyty z parametrami mfinal=1000 oraz boos=T.



KNN i klasyfikacja KNN

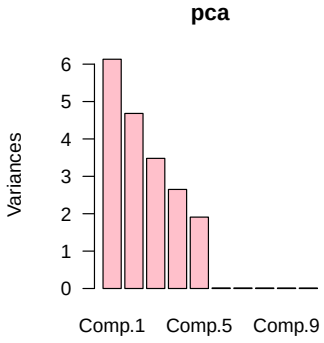
- Lewy wykres jest narysowany dla skalowanych danych bez zmiennej wyjaśnianej. Prawy to ROC dla klasyfikacji na testowym.

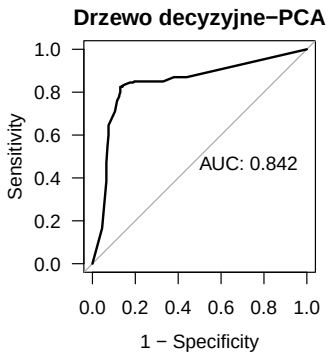




PCA i drzewo oparte na składowych głównych

- Analiza składowych głównych, wyodrębniła pięć głównych składowych, wyjaśniających ponad 99% wariancji. Klasyfikacja na tych składowych nie poprawiła wyniku.





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.