Agata Pałdyna, 277019

ZMUM - **Projekt 2** – Raport

1. Cel projektu.

Celem projektu było zbadanie metod selekcji zmiennych. Należało zaproponować metody selekcji zmiennych oraz klasyfikacji, które umożliwiają zbudowanie modelu o dużej mocy predykcyjnej przy użyciu możliwie małej liczby zmiennych.

2. Opis przetwarzania danych.

Dane treningowe znajdowały się w pliku `artificial_train.data`, etykiety danych treningowych w pliku `artificial_train.labels`, a dane walidacyjne w pliku `artificial_valid.data`. W danych treningowych było równo po 1000 obserwacji z klasy 1 oraz z klasy -1. Pierwszym krokiem jaki zrobiłam było połączenie danych z plików `artificial_train.data` i `artificial_train.labels` w jedną ramkę danych, aby ułatwić sobie pracę z modelami. Następnie wykonywałam takie kroki:

1) Dzielenie za pomocą kroswalidacji (metoda `createDataPartition()` z pakietu `caret`) danych treningowych na dwa zbiory:
   - train – do selekcji zmiennych i trenowania modeli; tu 90% obserwacji,
   - test – do testowania; tu pozostałe 10% obserwacji.
2) Selekcja zmiennych na zbiorze train.
3) Budowanie modeli random forest na wybranych w poprzednim kroku zmiennych, począwszy od najbardziej istotnej zmiennej, następnie dodając kolejne, coraz mniej istotne zmienne.
4) Predykcja na zbiorze test.
5) Obliczanie balanced accuracy (BA).

Powyższe kroki wykonywałam iteracyjnie, by móc uśrednić otrzymane wyniki miary BA. Następnie dla najlepszych zestawów zmiennych, tj. dających największe BA, dla każdej z metod zrobiłam 50, 100, i 2000 iteracji, w których dzieliłam na train i test oraz budowałam modele i dokonywałam predykcji, by móc jak najlepiej uśrednić miarę BA, a oprócz niej także dokładność i precyzję.

Kolejnym krokiem było wybranie spośród analizowanych metod selekcji najlepszej i dokonanie za jej pomocą predykcji dla danych ze zbioru `artificial_valid.data`, przypisując każdej obserwacji oszacowane prawdopodobieństwo aposteriori dla klasy 1. Wyniki predykcji zostały zapisane w pliku `AGAPAL_artificial_prediction.txt`, natomiast w pliku `AGAPAL_artificial_features.txt` zostały zapisane indeksy wybranych zmiennych.

3. Podsumowanie eksperymentów.

Testowałam następujące metody selekcji zmiennych:

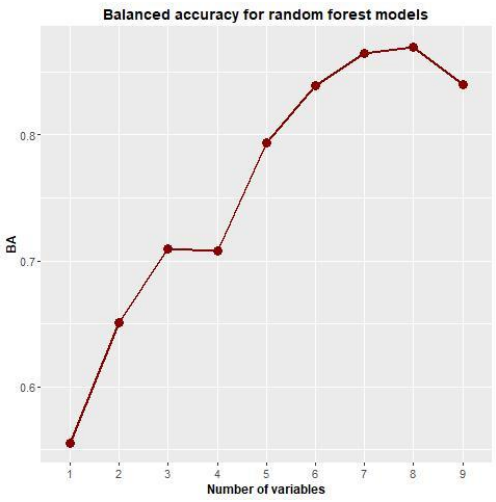
1) variable importance – metoda `varImp()` z pakietu `caret`,
2) Boruta – metoda `Boruta()` z pakietu `Boruta`.

Jeśli chodzi o metody klasyfikacji, za każdym razem używałam `randomForest()` z pakietu `randomForest` z parametrem `ntree = 100`.

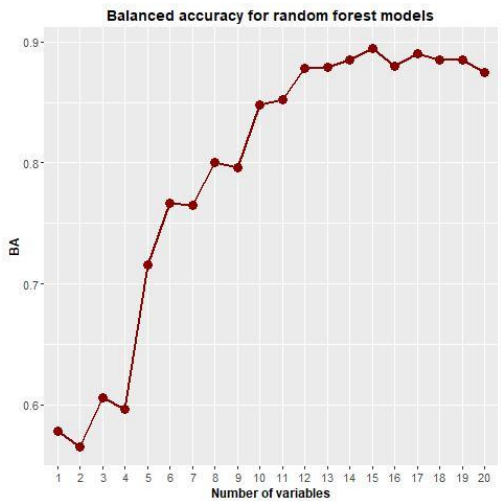
Każdą z metod najpierw przetestowałam jednorazowo, w większości przypadków z domyślnymi parametrami, w celu zorientowania się, jakie mniej więcej wyniki dają poszczególne metody, oraz ile czasu zajmuje ich wykonanie. Dzięki temu mogłam zdecydować, na jak dużo iteracji mogę sobie pozwolić.

3.1 Variable importance.

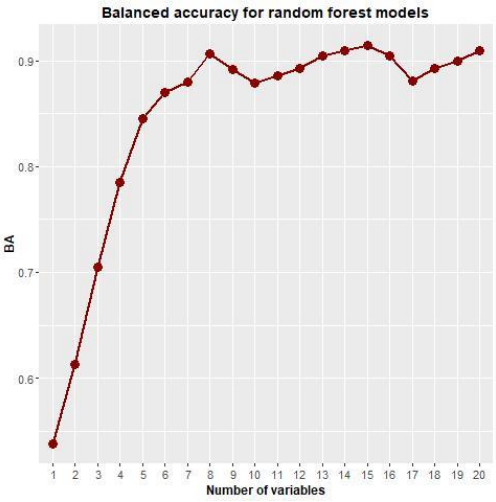
Chcąc użyć tej metody, najpierw musiałam użyć metody `train()`, również z pakietu `caret`, w celu wytrenowania modelu. W metodzie `train()` modyfikowałam parametr `method`. Funkcja `varImp()` zwracała mi dla `method` równego `rpart` zawsze 8 – 9 zmiennych, natomiast dla wartości `lvq` oraz `rf` 500 zmiennych – do dalszych testów brałam ok. 20 zmiennych o największej istotności. Poniżej są przedstawione wykresy BA od liczby zmiennych w modelu kolejno dla `method` = `rpart`, `lvq` oraz `rf` (na wykresach BA uśrednione po zestawach zmiennych w modelach, następnie max po liczbie zmiennych w modelu).



Rysunek 1 rpart



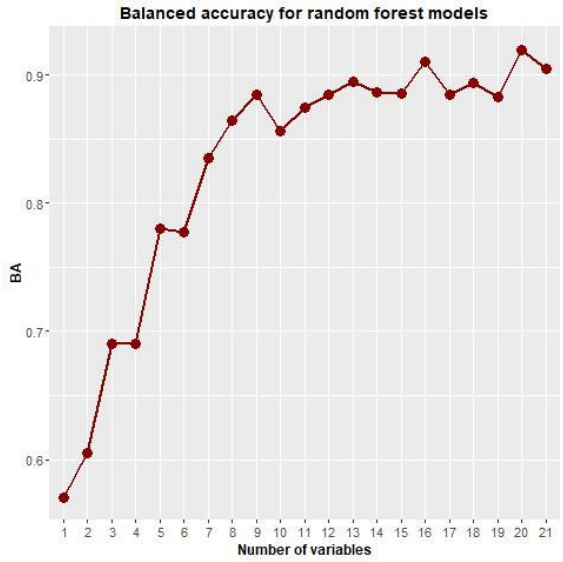
Rysunek 2 lvq



Rysunek 3 rf

3.2 Boruta.

Metoda Boruta natomiast zawsze zwracała 19 – 21 zmiennych. Tu również na wykresie BA uśrednione po zestawach zmiennych w modelach, następnie max po liczbie zmiennych w modelu.



Rysunek 4 Boruta

3.3 Podsumowanie.

Poniższa tabela przedstawia najlepsze uśrednione wyniki.

Metoda	BA	Precyzja	Liczba zmiennych	Zmienne w modelu
varImp, method = rpart	80.17 %	79.53%	8	V106 + V129 + V242 + V337 + V339 + V379 + V476 + V49
varImp, method = lvq	86.91 %	86.67%	15	V106 + V129 + V137 + V242 + V337 + V339 + V379 + V443 + V454 + V473 + V476 + V49 + V494 + V5 + V65
varImp, method = rf	87.86 %	87.73%	15	V106 + V154 + V242 + V282 + V29 + V319 + V339 + V379 + V434 + V443 + V454 + V473 + V476 + V49 + V494
Boruta	**88.29 %**	**88.48%**	**16**	**V106 + V129 + V154 + V242 + V282 + V29 + V319 + V337 + V339 + V379 + V434 + V443 + V452 + V473 + V476 + V49**
Boruta	88.94 %	88.97%	20	V106 + V129 + V154 + V242 + V282 + V29 + V319 + V337 + V339 + V379 + V434 + V443 + V452 + V454 + V456 + V473 + V476 + V49 + V494 + V65

4. Uzasadnienie wyboru końcowej metody.

Jak widać w powyższej tabelce, metoda Boruta okazała się najlepsza. Niewiele różniły się wyniki dla 16 oraz 20 zmiennych, więc ostatecznie wybrałam model z mniejszą liczbą zmiennych, tj. **metodę Boruta dla 16 zmiennych**.