Sprawozdanie 3

Jakub Markowiak
album 255705

28 maja 2021

Spis treści

1 Krótki opis zagadnienia

2 Opis eksperymentów/analiz

Przeprowadzimy następujące analizy i eksperymenty:

1. klasyfikacja na bazie modelu regresji liniowej,
2. porównanie metod klasyfikacji dla danych `Glass` z pakietu `mlbench`.

3 Wyniki

3.1 Klasyfikacja na bazie modelu regresji liniowej

Rozpoczynamy od wczytania danych `iris` z pakietu `datasets`.

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.10	3.50	1.40	0.20	setosa
2	4.90	3.00	1.40	0.20	setosa
3	4.70	3.20	1.30	0.20	setosa
4	4.60	3.10	1.50	0.20	setosa
5	5.00	3.60	1.40	0.20	setosa
6	5.40	3.90	1.70	0.40	setosa

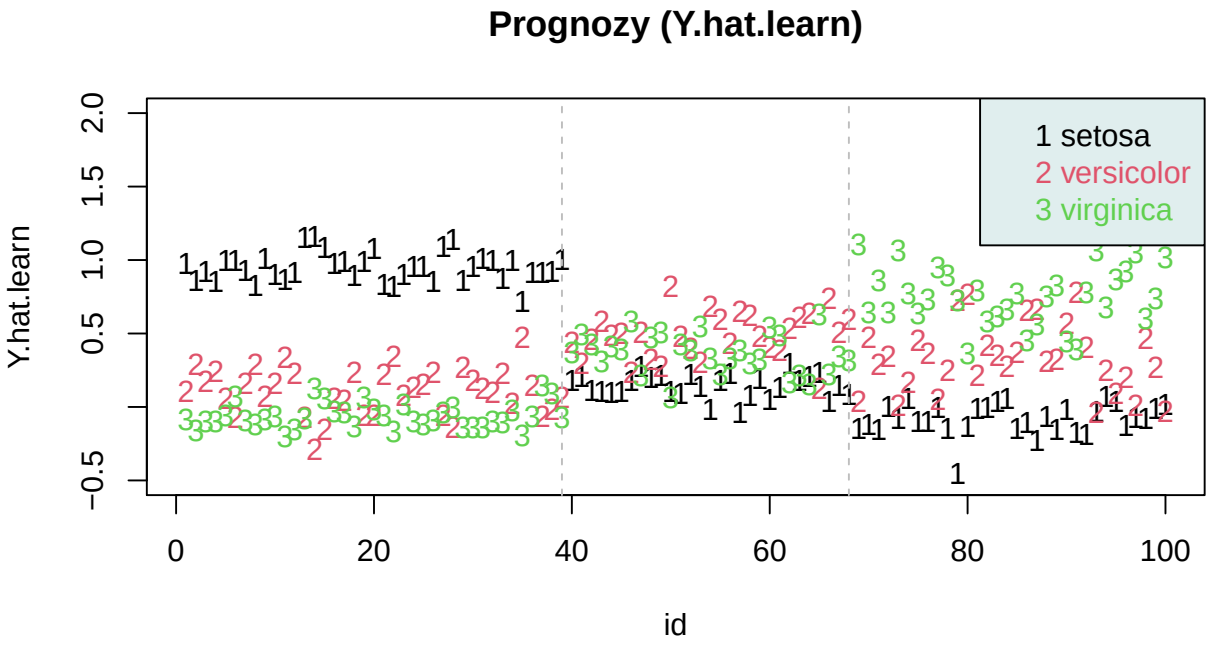
Tabela 1: Wczytane dane

Dzielimy teraz losowo dane na dwa zbiory – zbiór uczący i zbiór testowy (w proporcji 2 : 1).

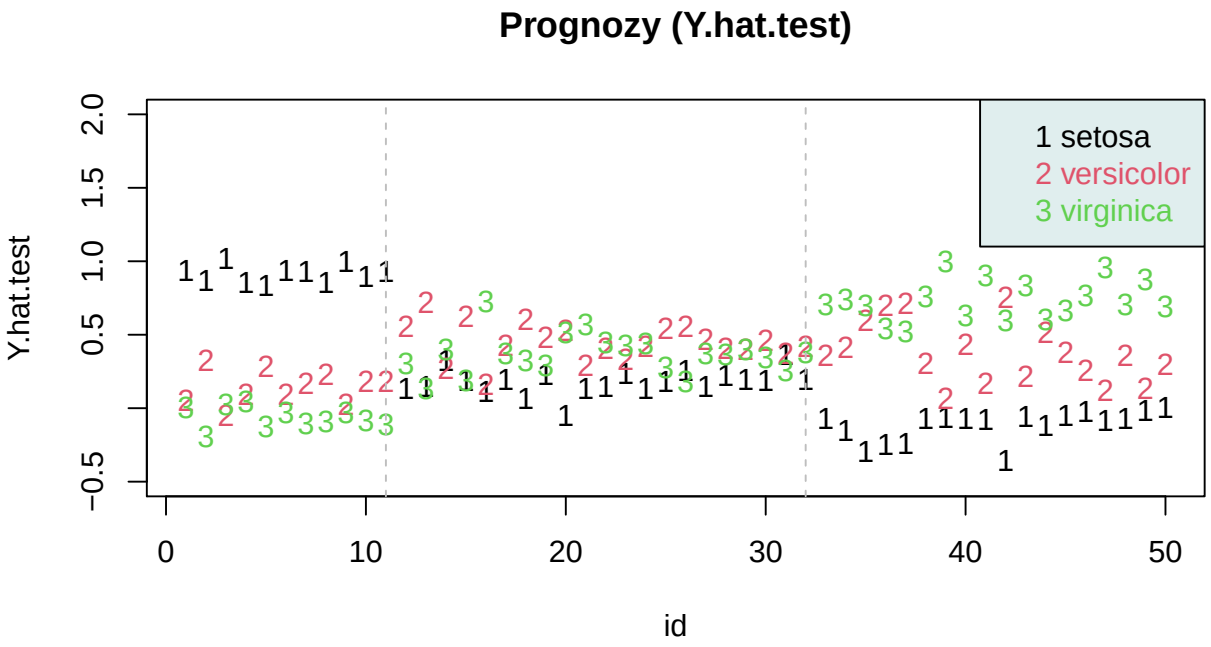
	l. obserwacji	versicolor	virginica	setosa
learn set	100	29	32	39
test set	50	21	18	11

Tabela 2: Podział na zbiór uczący i testowy

Przygotujemy teraz model regresji liniowej dla danych ze zbioru uczącego.



Rysunek 1: Prognoza dla zbioru uczącego (model 1)



Rysunek 2: Prognoza dla zbioru testowego (model 1)

Spróbujemy teraz ocenić jakość naszego modelu. W tym celu wyznaczamy macierz pomyłek dla zbioru uczącego i testowego.

Tabela 3: Macierz pomyłek dla zbioru uczącego (model 1)

Rzeczywiste etykietki	Prognozowane etkietki		
	setosa	versicolor	virginica
setosa	39	0	0
versicolor	0	21	8
virginica	0	6	26

Dokładność klasyfikacji dla zbioru uczącego wynosi 0.86.

Tabela 4: Macierz pomyłek dla zbioru testowego (model 1)

Rzeczywiste etykietki	Prognozowane etkietki		
	setosa	versicolor	virginica
setosa	11	0	0
versicolor	0	14	7
virginica	0	3	15

Dokładność klasyfikacji dla zbioru testowego natomiast jest równa 0.8.

Możemy zauważyć, że w zbiorze testowym odsetek pomyłek, gdy w rzeczywistości mamy `versicolor`, to około 0.3333333. Również spoglądając na wykres widzimy, że obserwacje w środkowej grupie są słabo odseparowane. Stąd wyciągamy wniosek, że w naszym modelu występuje efekt maskowania klasy `versicolor`.

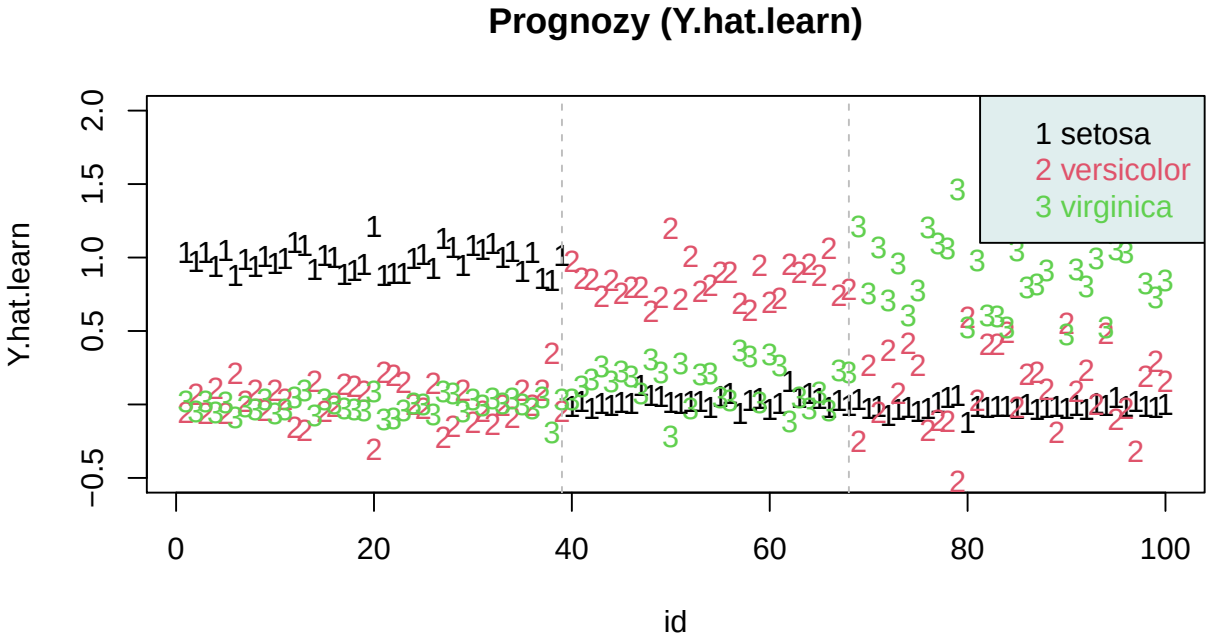
Spróbujemy teraz zbudować model regresji po uzupełnieniu wyjściowych cech o składniki wielomianowe stopnia 2. Konstruujemy w tym celu nową ramkę danych.

	SL	SW	PL	PW	Species	PL2	PW2	SL2	SW2
1	5.10	3.50	1.40	0.20	setosa	1.96	0.04	26.01	12.25
2	4.90	3.00	1.40	0.20	setosa	1.96	0.04	24.01	9.00
3	4.70	3.20	1.30	0.20	setosa	1.69	0.04	22.09	10.24
4	4.60	3.10	1.50	0.20	setosa	2.25	0.04	21.16	9.61
5	5.00	3.60	1.40	0.20	setosa	1.96	0.04	25.00	12.96
6	5.40	3.90	1.70	0.40	setosa	2.89	0.16	29.16	15.21

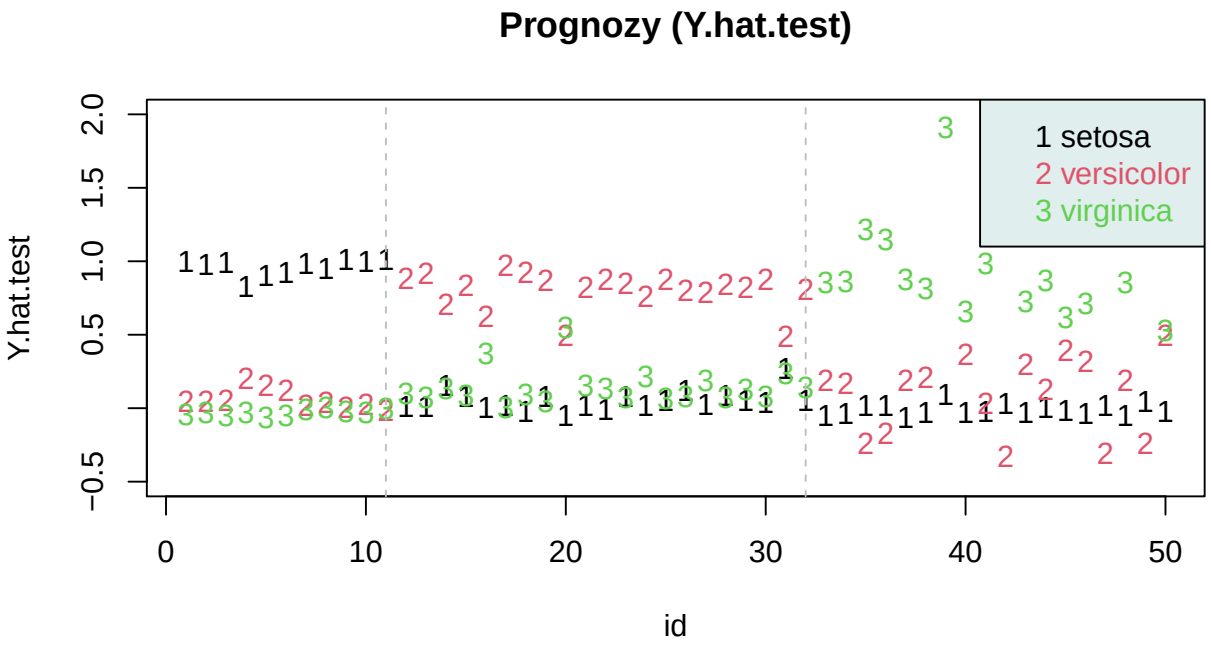
	PL*PW	PL*SW	PL*SL	PW*SL	PW*SW	SL*SW
1	0.28	4.90	7.14	1.02	0.70	17.85
2	0.28	4.20	6.86	0.98	0.60	14.70
3	0.26	4.16	6.11	0.94	0.64	15.04
4	0.30	4.65	6.90	0.92	0.62	14.26
5	0.28	5.04	7.00	1.00	0.72	18.00
6	0.68	6.63	9.18	2.16	1.56	21.06

Tabela 5: Nowa ramka danych

Teraz powtarzając poprzednie kroki (wybieramy taki sam zbiór uczący i testowy) rysujemy wykresy dla modelu 2.



Rysunek 3: Prognozy dla zbioru uczącego (model 2)



Rysunek 4: Prognozy dla zbioru testowego (model 2)

Analogicznie jak dla modelu 1, wyznaczymy również macierz pomyłek.

Tabela 6: Macierz pomyłek dla zbioru uczącego (model 2)

Rzeczywiste etykietki	Prognozowane etkietki		
	setosa	versicolor	virginica
setosa	39	0	0
versicolor	0	29	0
virginica	0	2	30

W modelu 2 dokładność klasyfikacji dla zbioru uczącego wynosi 0.98.

Tabela 7: Macierz pomyłek dla zbioru testowego (model 2)

Rzeczywiste etykietki	Prognozowane etkietki		
	setosa	versicolor	virginica
setosa	11	0	0
versicolor	0	20	1
virginica	0	0	18

Dokładność klasyfikacji dla zbioru testowego natomiast jest równa 0.98. Porównajmy teraz dokładność klasyfikacji dla obu modeli.

	Zbiór uczący	Zbiór testowy
Model 1	0.86	0.80
Model 2	0.98	0.98

Tabela 8: Porównanie dokładności klasyfikacji

Widzimy, że zdecydowanie lepszym modelem jest model 2. Również spoglądając na wykresy możemy zauważyć, że pozbyliśmy się zjawiska maskowania cechy `versicolor`. Możemy stąd wywnioskować, że uwzględnienie składników wielomianowych stopnia 2 miało kluczowy wpływ na jakość naszego modelu klasyfikacyjnego.

3.2 Porównanie metod klasyfikacji dla danych Glass z pakietu mlbench

Rozpoczynamy od wczytania danych `Glass`, które zawierają informacje o współczynniku załamania światła oraz zawartości poszczególnych pierwiastków chemicznych dla badanych szkieł.

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe	Type
1	1.52	13.64	4.49	1.10	71.78	0.06	8.75	0.00	0.00	1
2	1.52	13.89	3.60	1.36	72.73	0.48	7.83	0.00	0.00	1
3	1.52	13.53	3.55	1.54	72.99	0.39	7.78	0.00	0.00	1
4	1.52	13.21	3.69	1.29	72.61	0.57	8.22	0.00	0.00	1
5	1.52	13.27	3.62	1.24	73.08	0.55	8.07	0.00	0.00	1
6	1.52	12.79	3.61	1.62	72.97	0.64	8.07	0.00	0.26	1

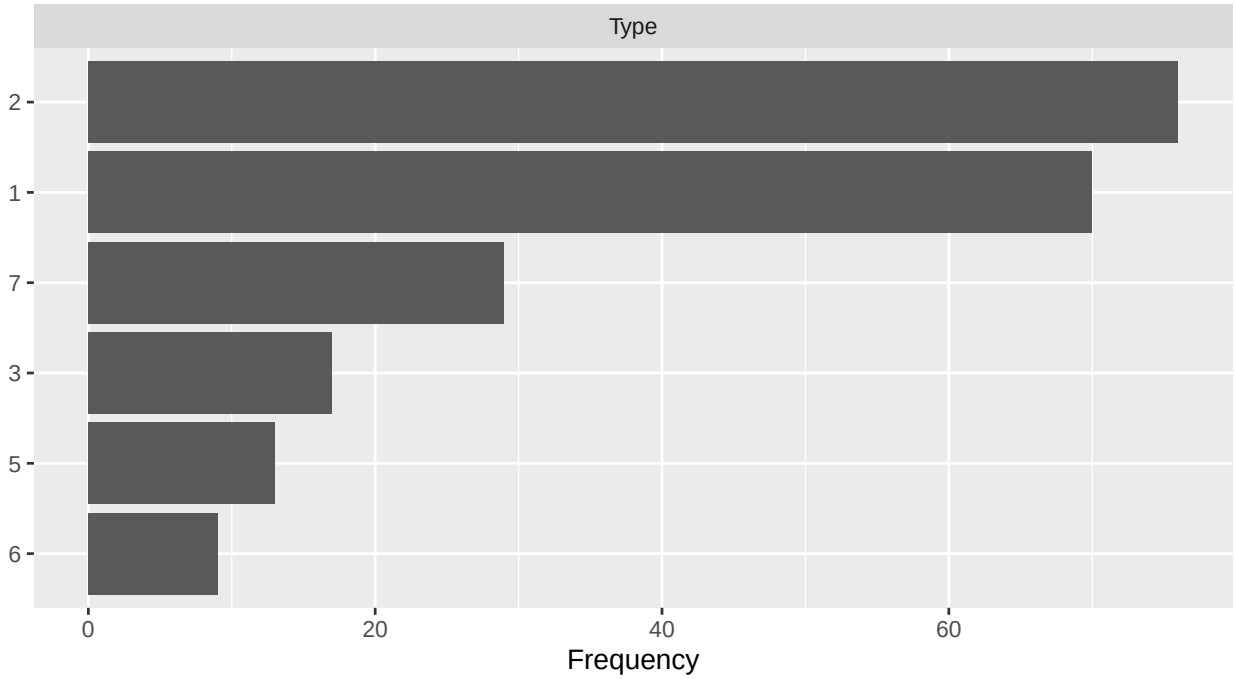
Tabela 9: Dane Glass - kilka pierwszych wierszy

Etykietą jest cecha `Type` – mamy 6 klas. Nietypowy jest fakt, że nie pojawia się szkło oznaczone jako „4”, mamy natomiast 1, 2, 3, 5, 6, 7.

Liczba cech	Liczba obserwacji	Cechy ilościowe	Cechy jakościowe	Brakujące wartości
10	214	9	1	0

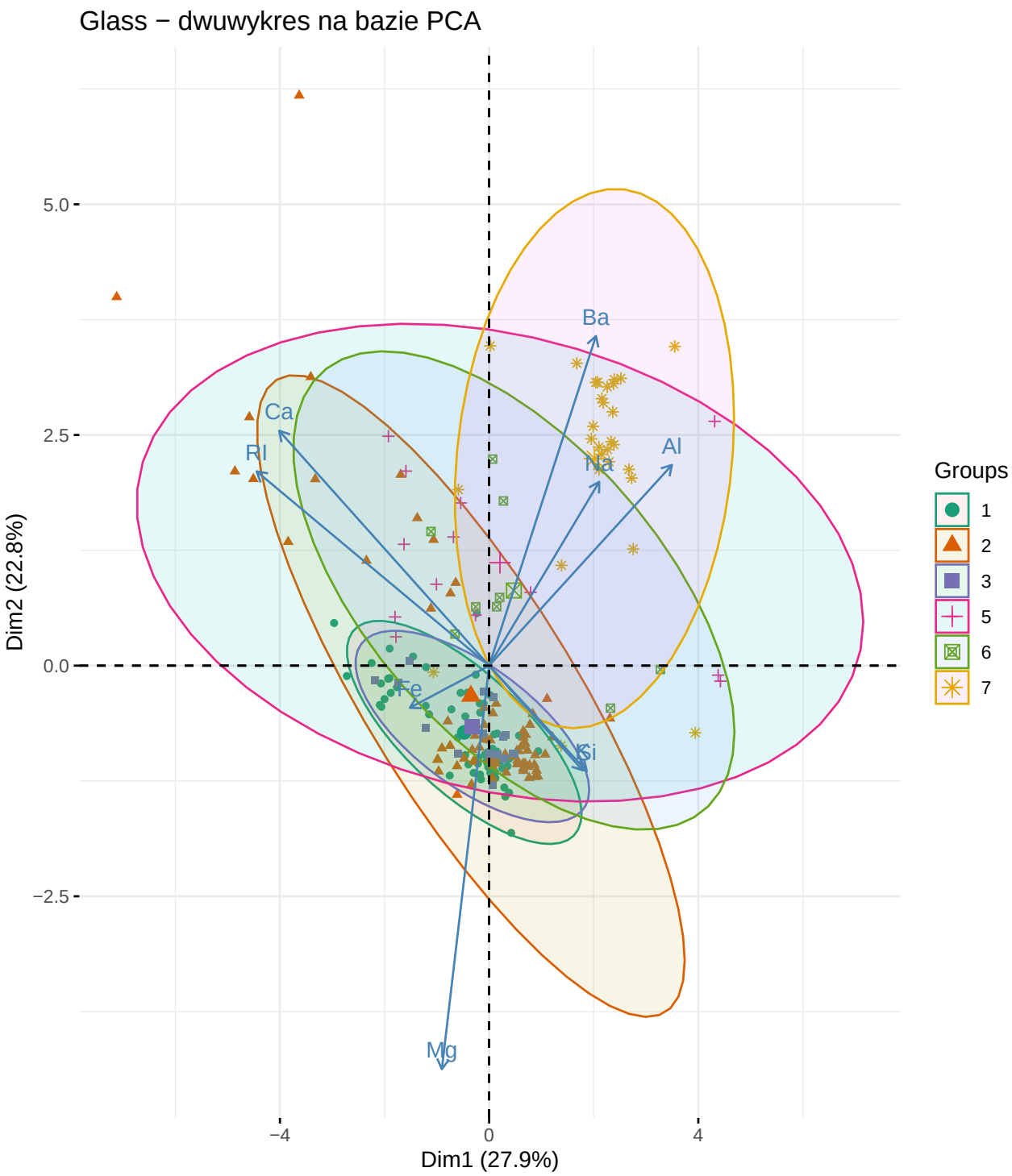
Tabela 10: Wstępne spojrzenie na dane

Sprawdźmy, jak często pojawiają się szkła danego typu w tej ramce dnaych.



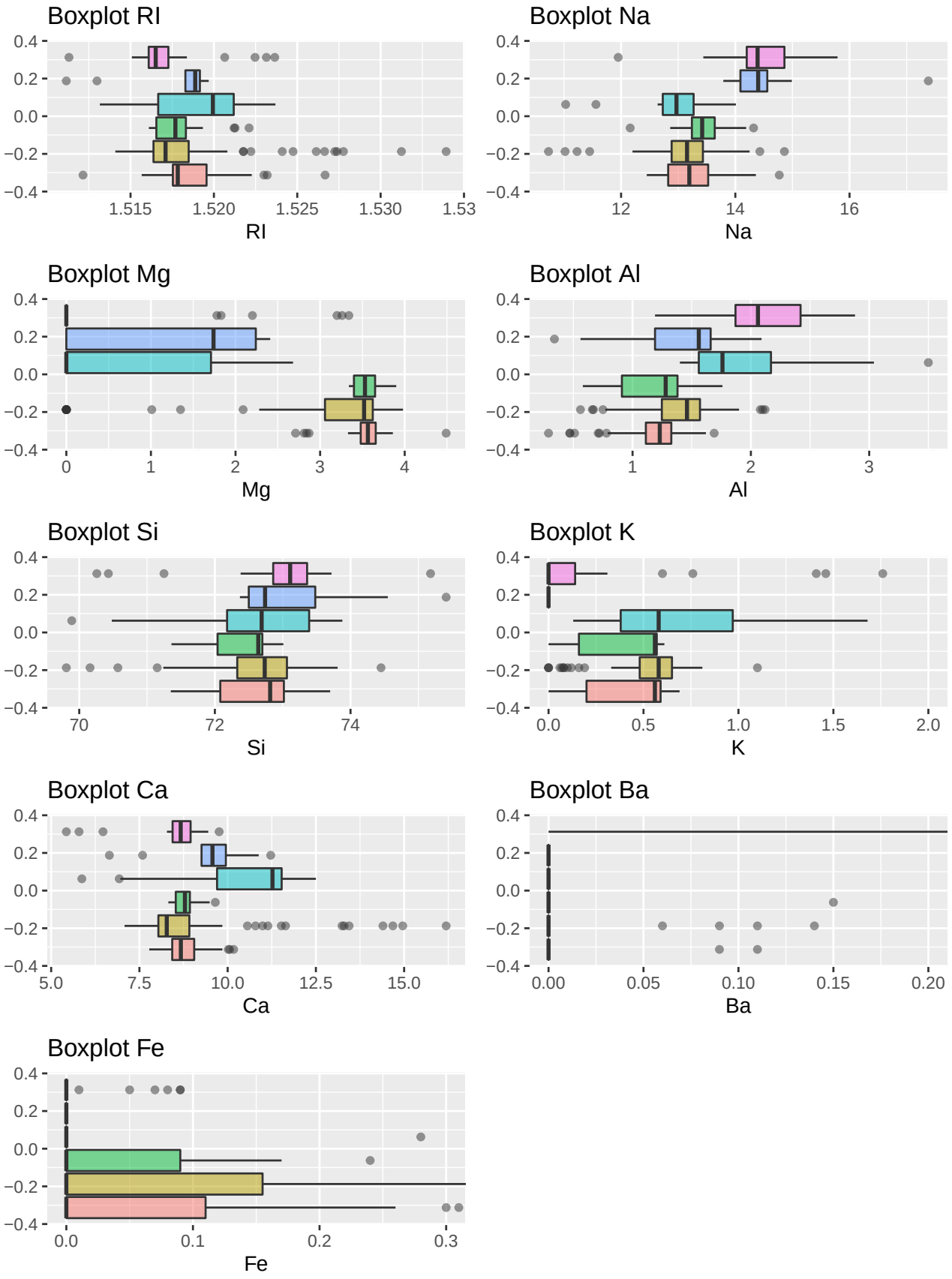
Rysunek 5: Wykres częstości dla Type

Widzimy, że najwięcej mamy szkieł typu 1 oraz 2. Zbadajmy również rozkład dla pozostałych cech. Przeprowadzimy teraz analizę PCA oraz sporządzimy kolejno wykresy pudełkowe, wykresy gęstości oraz macierze korelacji z podziałem na kolejne grupy.

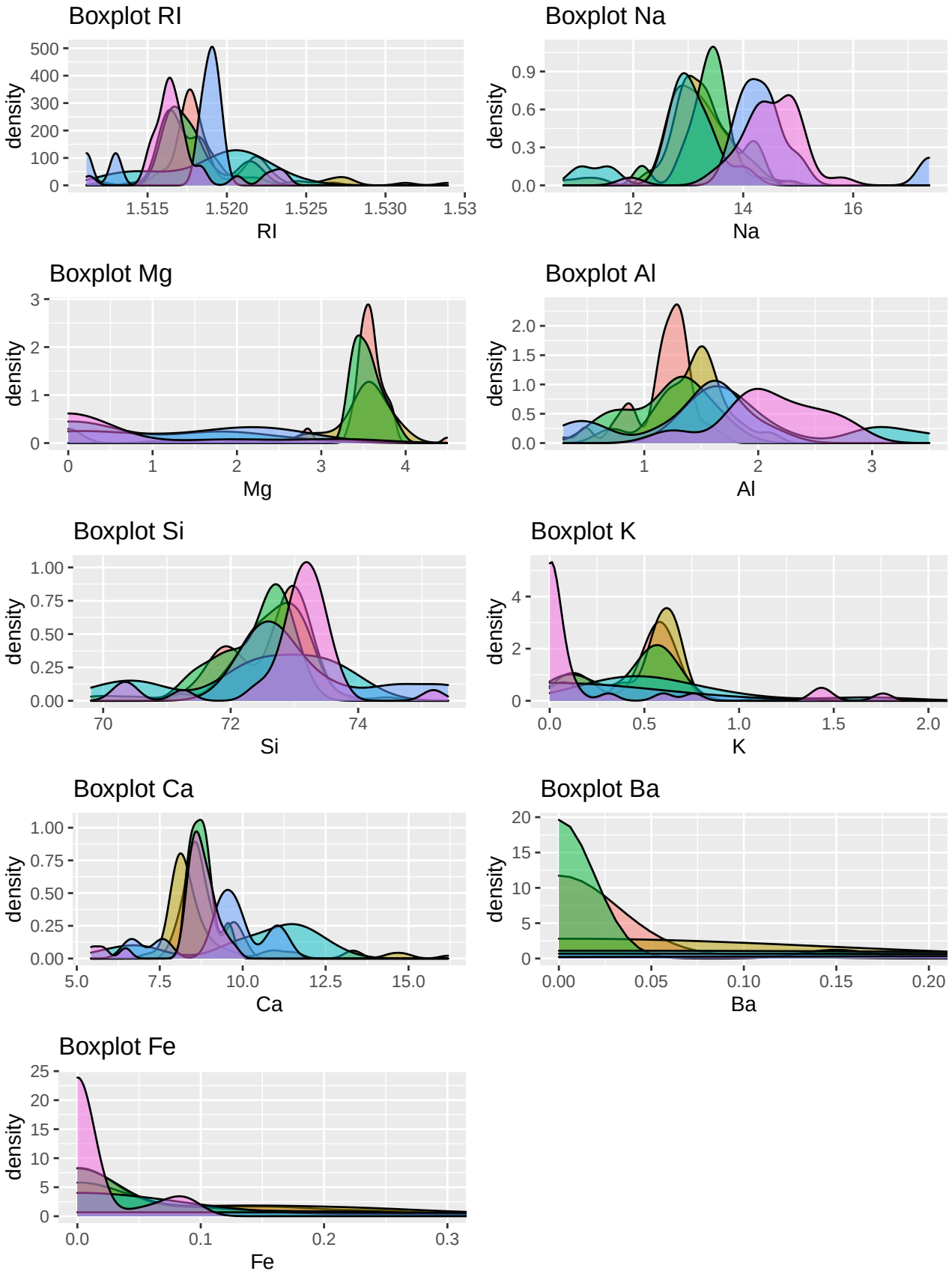


Rysunek 6: Analiza PCA

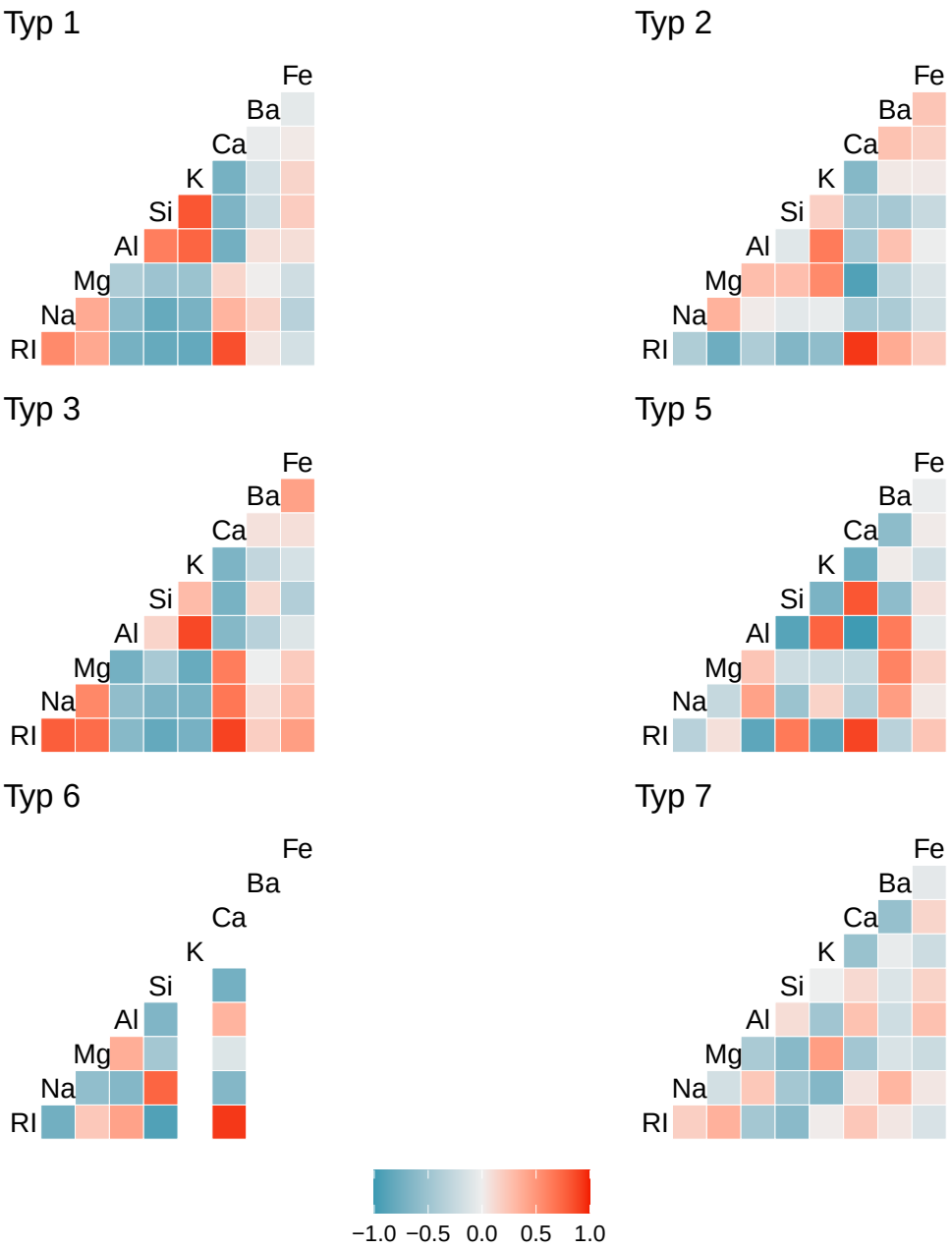
Wstępnie widzimy, że typ 7 powinien być dobrze charakteryzowany przez zawartość baru, glinu oraz sodu, a typ 2 przez zawartość wapienia oraz współczynnik załamania. Najbardziej rozrzuconymi zdają się natomiast być typ 5 i 6, a więc najmniej liczne rodzaje szkieł.



Rysunek 7: Wykresy pudełkowe według Type



Rysunek 8: Wykresy gęstości



Rysunek 9: Macierze korelacji dla kolejnych typów

Kilka wstępnych obserwacji, które możemy odczytać z powyższych wykreśów:

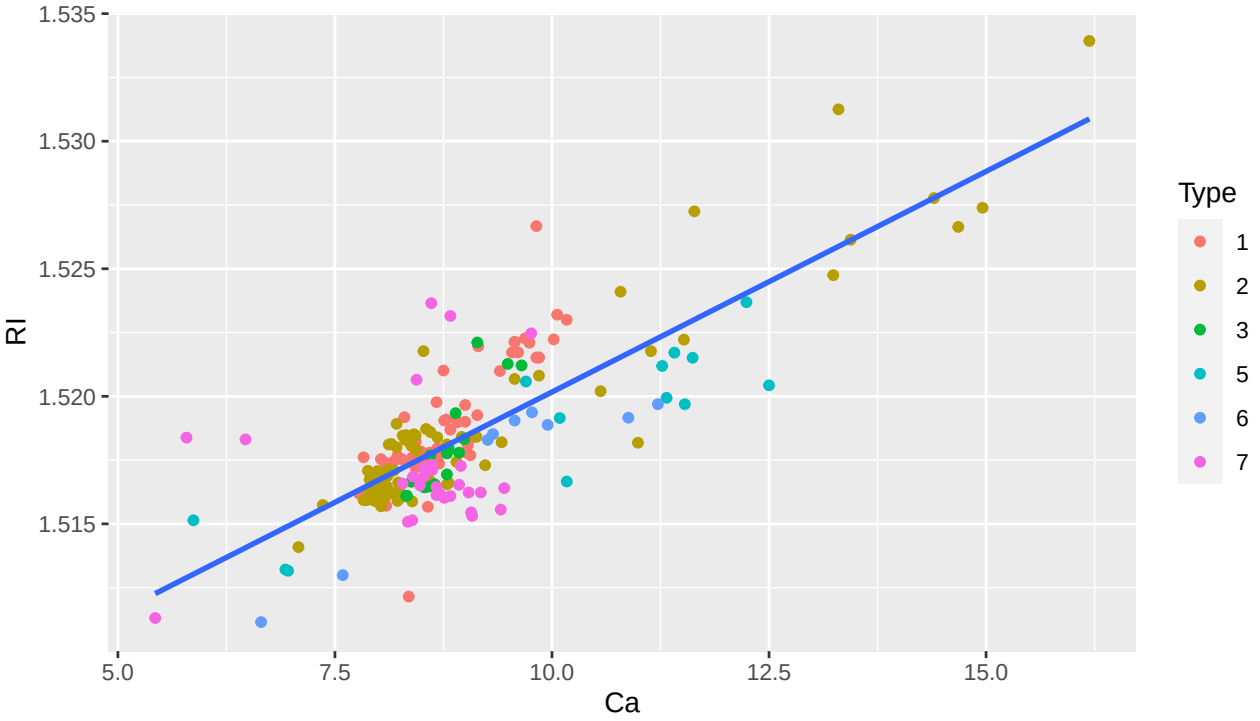
1. Typy 1, 2, 3 wyróżniają się na tle pozostałych podobną zawartością magnezu oraz żelaza,
2. typ 5 wyróżnia większa zawartość wapienia oraz niewielka zawartość magnezu,
3. typ 6 wyróżnia się niewielką zawartością magnezu oraz wysoką zawartością sodu, ma także mocno skupiony wykres gęstości dla RI,
4. brak zawartości baru powinna dobrze charakteryzować typ 1 i 3,
5. istnieje niemal liniowa korelacja pomiędzy zawartością wapienia, a współczynnikiem załamania (z wyjątkiem typu 7),

6. typ 6 wyróżnia się wysoką korelacją pomiędzy krzemem i sodą,

7. typy 2 i 7 wyróżniają się wyższą korelacją pomiędzy magnezem i potasem.

8. najgorsze zdolności dyskryminacyjne zdają się mieć żelazo oraz potas.

Narysujemy teraz wykres rozrzutu dla współczynnika załamnia i krzemu z podziałem na grupy.



Rysunek 10: Wykres rozrzutu RI vs Ca

Faktycznie widzimy, że mamy do czynienia z niemal liniową zależnością.

Przygotujemy teraz zbiór uczący i zbiór testowy (w proporcji 2 : 1), na których będziemy przeprowadzali porównanie kolejnych metod.

	l. obserwacji	1	2	3	5	6	7
learn set	143	49	43	12	10	7	22
test set	71	21	33	5	3	2	7

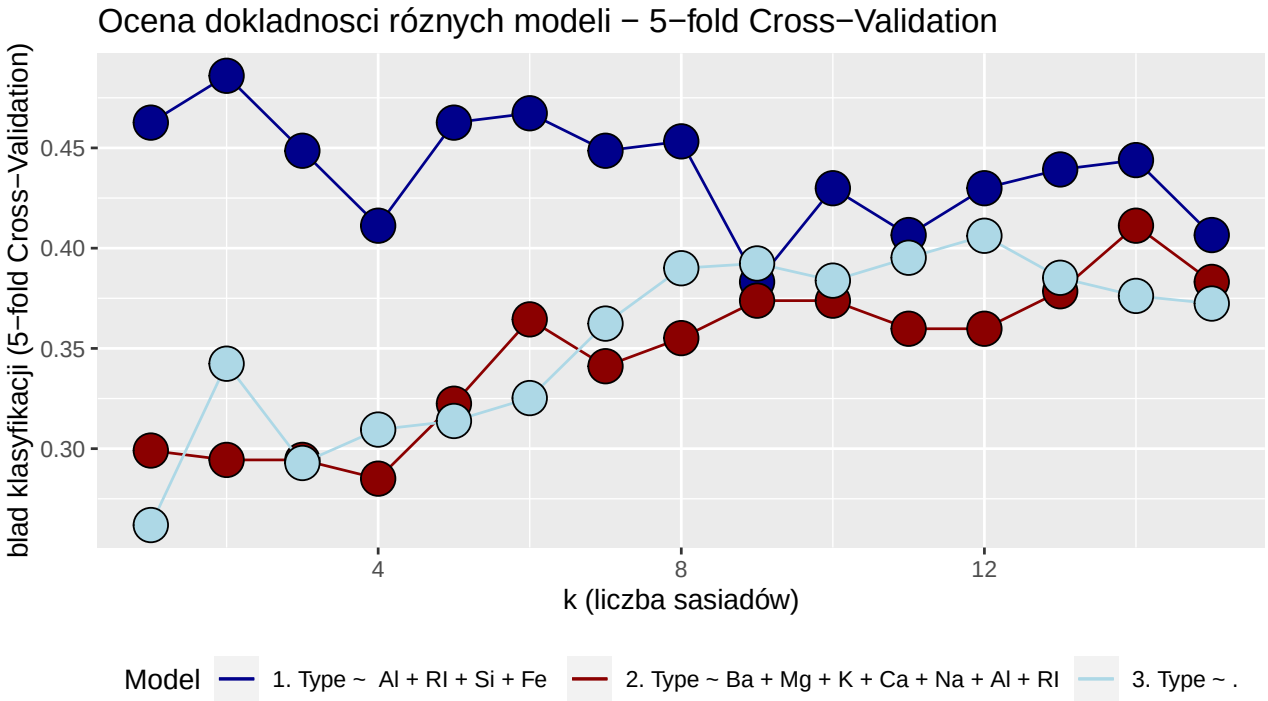
Tabela 11: Podział na zbiór uczący i testowy

Analizę przeprowadzimy dla następujących metod:

1. k-najbliższych sąsiadów,
2. drzewa klasyfikacyjne,
3. naiwny klasyfikator bayesowski.

Metoda k-najbliższych sąsiadów

Rozpoczniemy od metody **k-najbliższych sąsiadów**. Najpierw, przy użyciu **5-cross validation** spróbujemy porównać różne modele i wybrać, dla jakich zmiennych objaśniających otrzymujemy najlepsze wyniki.



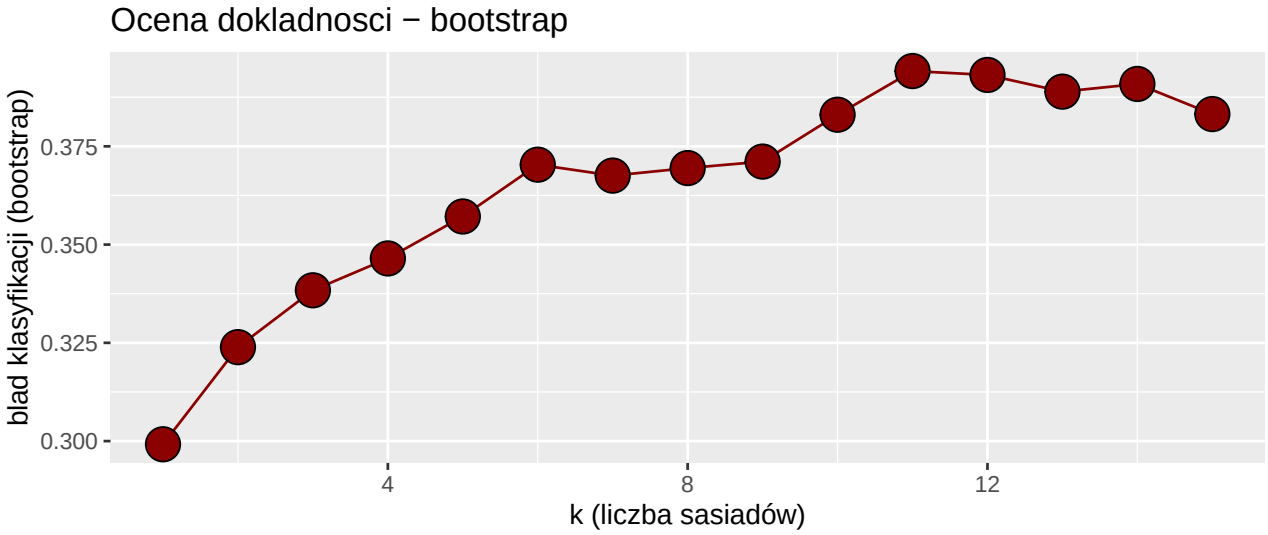
Rysunek 11: Porównanie doboru zmiennych objaśniających

Widzimy, że model 1 radzi sobie najgorzej (zmienne objaśniające to te o najgorszych zdolnościach dyskryminacyjnych), natomiast najlepiej poradził sobie model 2, a więc ten w którym uwzględniono wszystkie zmienne z wyjątkiem dwóch o najsłabszych zdolnościach dyskryminacyjnych – krzemu oraz żelaza.

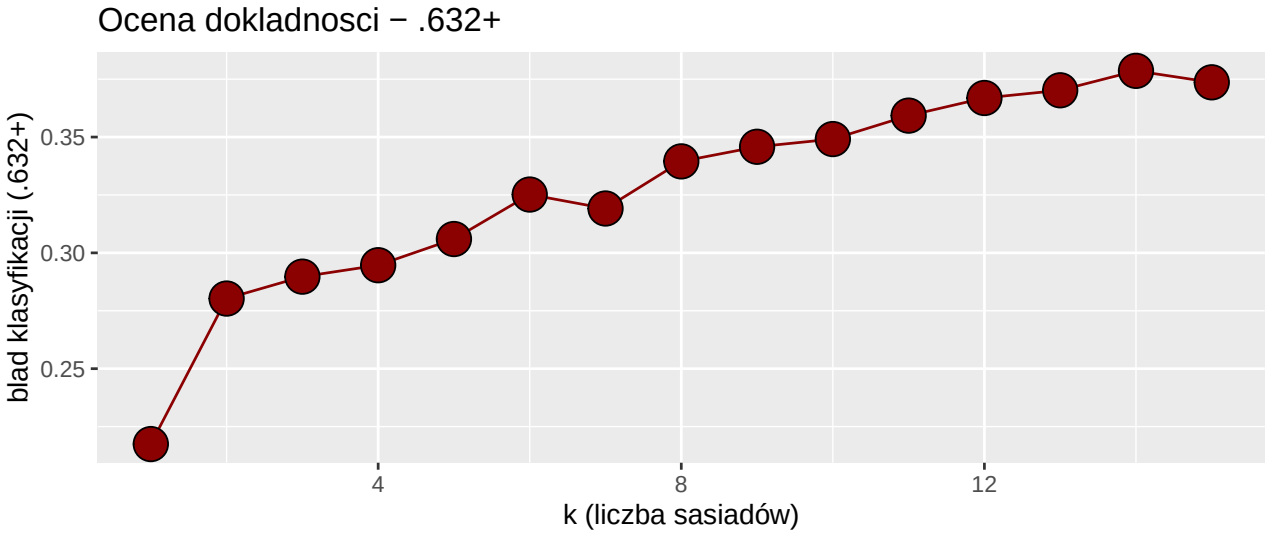
Przeprowadzimy ocenę metodami **K-cross validation**, **Bootstrap** oraz **.632+**, aby dobrać najlepszą liczbę sąsiadów.

	1	2	3	4	5	Średnia
Błędne predykcje w bloku, k=1	0.286	0.262	0.238	0.238	0.286	0.262
Błędne predykcje w bloku, k=3	0.238	0.190	0.310	0.333	0.333	0.281
Błędne predykcje w bloku, k=5	0.286	0.262	0.357	0.333	0.262	0.300
Błędne predykcje w bloku, k=10	0.357	0.333	0.429	0.452	0.405	0.395

Tabela 12: 5-Cross validation dla k-nn



Rysunek 12: Bootstrap dla różnej liczby sąsiadów



Rysunek 13: .632+ dla różnej liczby sąsiadów

Widzimy, że najlepszą liczbą sąsiadów zdaje się być 1. Sporządzimy zatem macierz pomyłek oraz sprawdzimy dokładność klasyfikacji w zbiorze uczącym i testowym.

Tabela 13: Macierz pomyłek dla zbioru uczącego (1-NN)

Rzeczywiste etykietki	Prognozowane etkietki					
	1	2	3	5	6	7
1	49	0	0	0	0	0
2	0	43	0	0	0	0
3	0	0	12	0	0	0
5	0	0	0	10	0	0
6	0	0	0	0	7	0
7	0	0	0	0	0	22

Tabela 14: Macierz pomyłek dla zbioru testowego (1-NN)

Rzeczywiste etykietki	Prognozowane etkietki					
	1	2	3	5	6	7
1	14	6	4	0	0	0
2	3	25	0	2	0	0
3	4	0	1	0	0	1
5	0	1	0	1	0	1
6	0	1	0	0	2	0
7	0	0	0	0	0	5

Dokładność klasyfikacji dla zbioru uczącego wynosi 1.

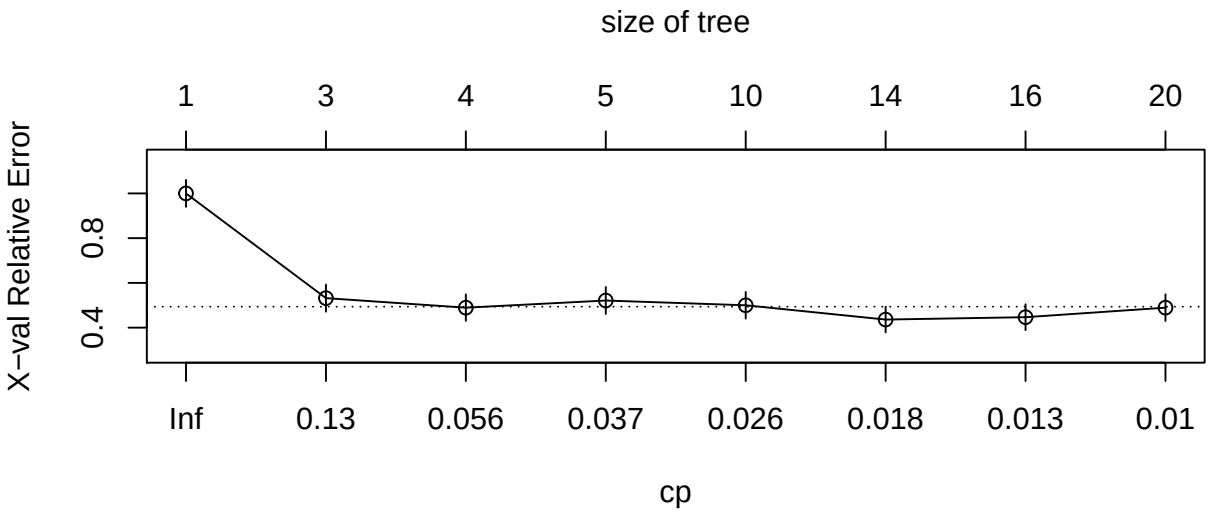
Dokładność klasyfikacji dla zbioru testowego wynosi 0.6760563. Wyniki dla najlepszego modelu podsumowuje poniższa tabela:

	5-cv	bootstrap	.632+	predykcje na zb. testowym
błąd klasyfikacji (k-nn)	0.295	0.302	0.219	0.324

Tabela 15: Błędy klasyfikacji dla k-nn

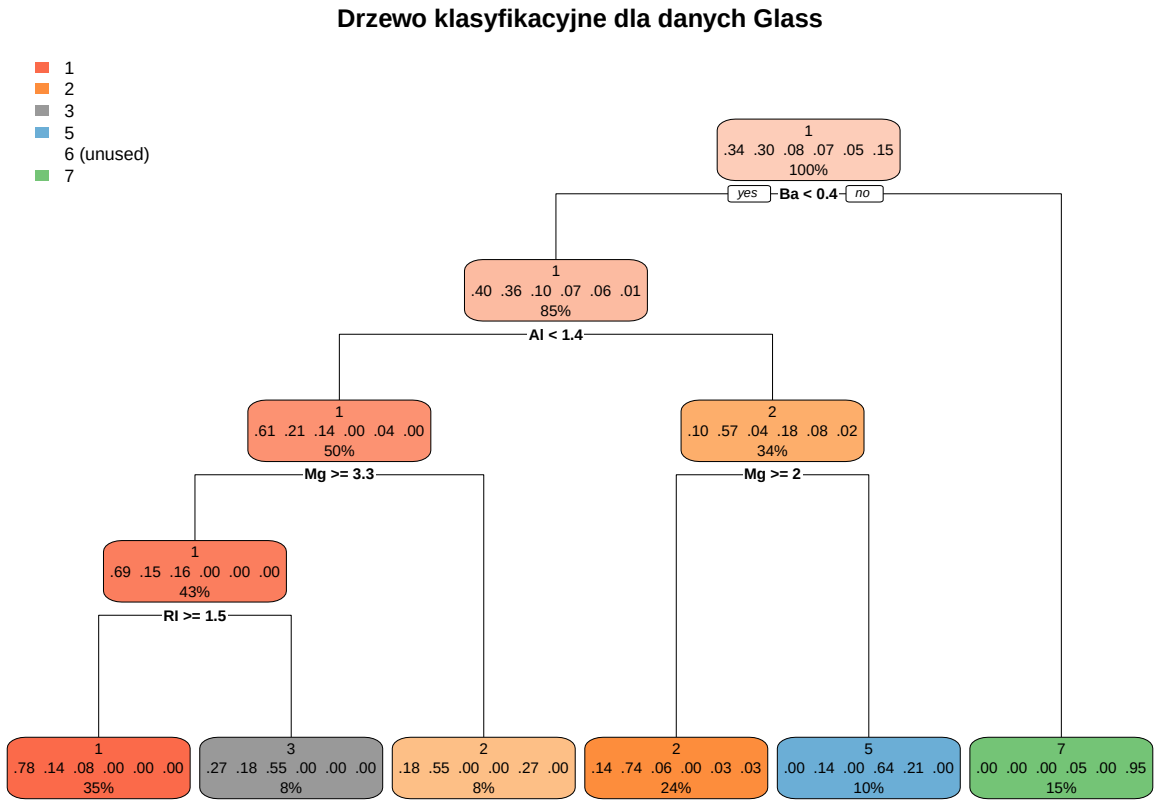
Metoda drzew klasyfikacyjnych

Podobnie skonstruujemy model przy użyciu metody **drzew klasyfikacyjnych**. Najpierw korzystam z kryterium kosztu złożoności, aby wybrać optymalny rozmiar drzewa.



Rysunek 14: Kryterium kosztu złożoności

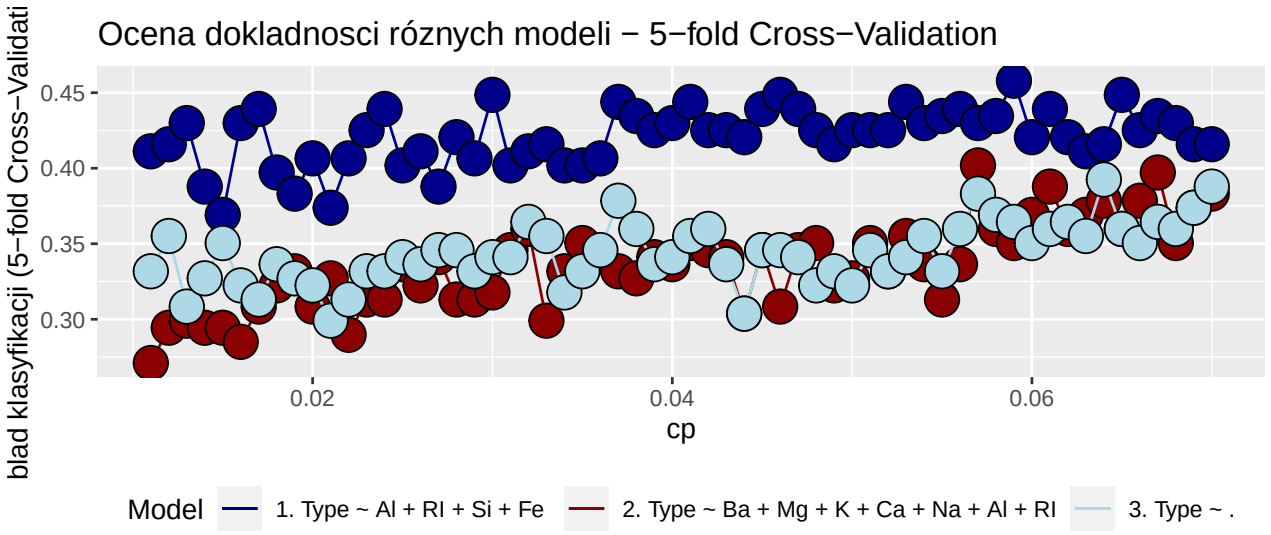
Optymalną wartością `cp` jest 0.0159574.



Rysunek 15: Drzewo klas. po zastosowaniu kryt. kosztu złożoności

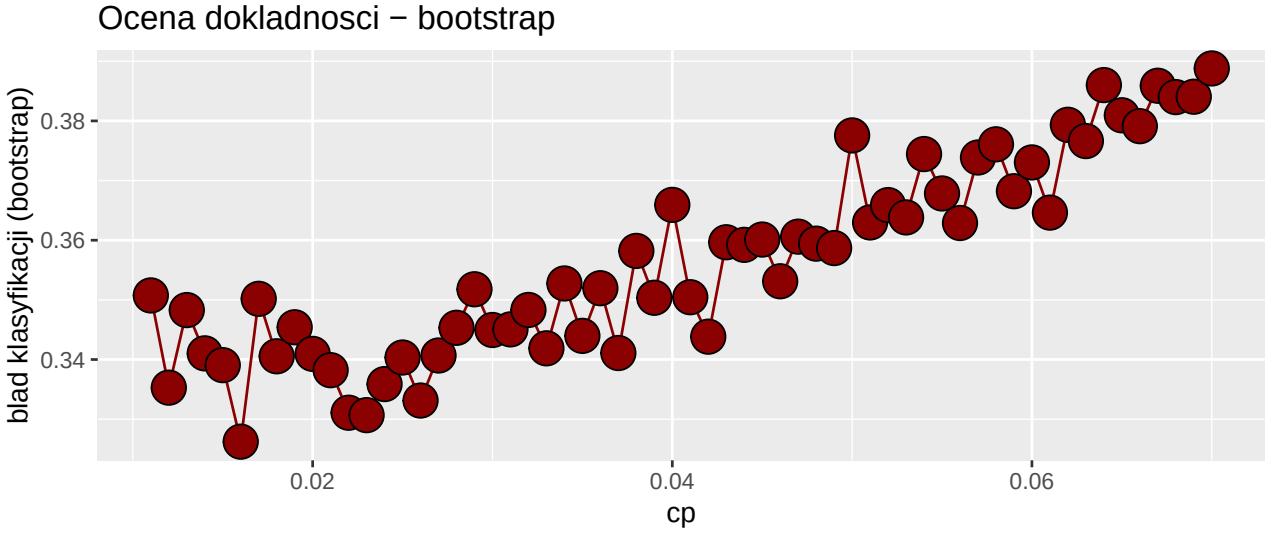
Sprawdzimy teraz, jaki wpływ na kształt drzewa klasyfikacyjnego ma dobór zmiennych objaśniających. Podobnie jak poprzednio, zbudujemy dwa nowe modele na podstawie zmiennych o

najlepszych i najgorszych zdolnościach dyskryminacyjnych. Przeprowadzimy w tym celu ocenę przy użyciu **5-cross validation**.

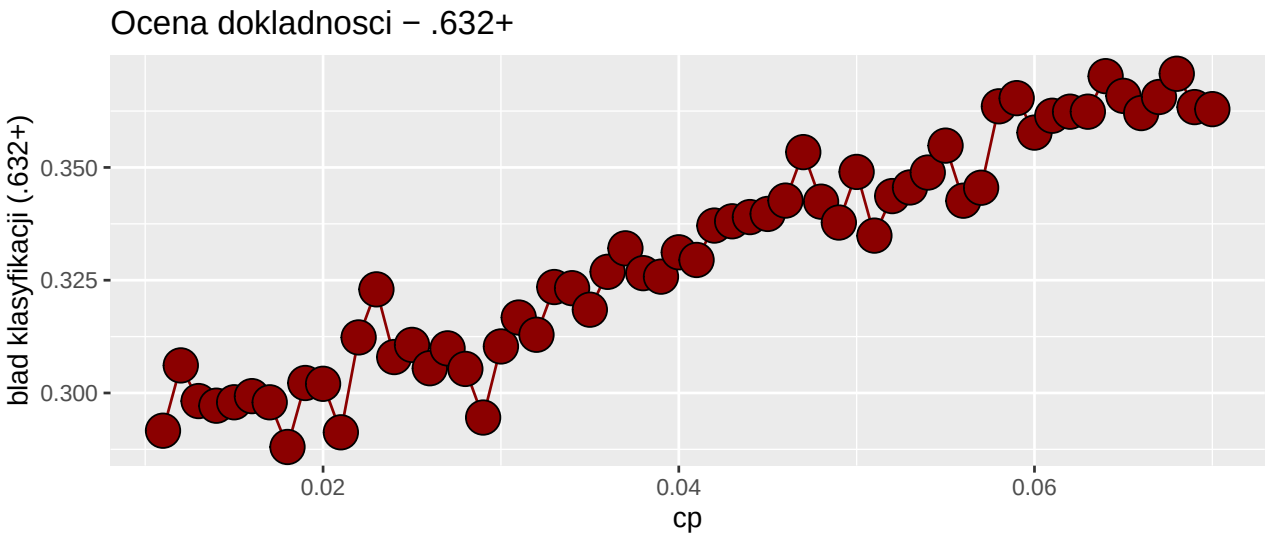


Rysunek 16: Porównanie błędów predykcji dla różnych modeli

Model 2 nie uwzględnia zmiennych o najgorszych cechach dyskryminacyjnych – krzemu oraz żelaza. Widzimy, że zdaje się on mieć mniejszy błąd klasyfikacyjny od pozostałych. Dlatego do dalszej analizy wybieramy właśnie ten model.



Rysunek 17: Ocena dokładności klasyfikacji metodą bootstrap



Rysunek 18: Ocena dokładności klasyfikacji metodą .632+

	1	2	3	4	5	Średnia
Błędne predykcje w bloku, cp= 0.016	0.381	0.357	0.286	0.286	0.357	0.333
Błędy predykcyjne w bloku, cp= 0.11	0.452	0.429	0.333	0.429	0.405	0.410
Błędy predykcyjne w bloku, cp= 0.06	0.429	0.405	0.286	0.381	0.429	0.386

Tabela 16: 5-Cross validation dla drzew klasyfikacyjnych, różne cp

Widzimy, że dla `cp` innych niż wybrane na mocy kryterium kosztu złożoności mamy większe rozbieżności pomiędzy wynikami w poszczególnych blokach.

Tabela 17: Macierz pomyłek dla zbioru uczącego (dk)

Rzeczywiste etykietki	Prognozowane etkietki					
	1	2	3	5	6	7
1	39	7	4	0	0	0
2	7	32	2	0	4	1
3	3	2	6	0	0	0
5	0	2	0	9	3	0
6	0	0	0	0	0	0
7	0	0	0	1	0	21

Dokładność klasyfikacji dla zbioru uczącego wynosi 0.7482517.

Dokładność klasyfikacji dla zbioru testowego wynosi 0.6760563. Wyniki dla najlepszego modelu podsumowuje poniższa tabela:

	5-cv	bootstrap	.632+	predykcje na zb. testowym
błąd klasyfikacji (dk)	0.345	0.330	0.301	0.324

Tabela 19: Błędy klasyfikacji dla drzewa klas.

Tabela 18: Macierz pomyłek dla zbioru testowego (dk)

Rzeczywiste etykietki	Prognozowane etkietki					
	1	2	3	5	6	7
1	17	8	2	0	0	1
2	3	23	2	1	2	1
3	0	1	1	0	0	0
5	0	0	0	2	0	0
6	0	0	0	0	0	0
7	1	1	0	0	0	5

Naiwny klasyfikator bayesowski

	Model 1	Model 2	Model 3
5-cv	0.762	0.542	0.626
bootstrap	0.708	0.567	0.600
.632+	0.734	0.538	0.578

Tabela 20: Porównanie modeli

Oznaczenia modeli jak poprzednio – model 1 jest zbudowany na podstawie zmiennych o najgorszych zdolnościach klasyfikacyjnych, model 2 na podstawie wszystkich zmiennych z wyjątkiem krzemu i żelaza, natomiast model 3 ze wszystkich zmiennych. Również widzimy, że model 2 wypadł najlepiej. Skonstruujemy macierze pomyłek dla tego klasyfikatora.

Tabela 21: Macierz pomyłek dla zbioru uczącego (NB)

Rzeczywiste etykietki	Prognozowane etkietki					
	1	2	3	5	6	7
1	44	30	8	0	0	0
2	3	9	0	7	0	1
3	2	0	3	0	0	0
5	0	3	0	3	0	1
6	0	1	1	0	7	0
7	0	0	0	0	0	20

Dokładność klasyfikacji dla zbioru uczącego wynosi 0.6013986.

Dokładność klasyfikacji dla zbioru testowego wynosi 0.4084507. Wyniki podsumowuje poniższa tabela:

	5-cv	bootstrap	.632+	predykcje na zb. testowym
błąd klasyfikacji (NB)	0.542	0.567	0.538	0.592

Tabela 23: Błędy klasyfikacji dla NB

Tabela 22: Macierz pomyłek dla zbioru testowego (NB)

Rzeczywiste etykietki	Prognozowane etkietki					
	1	2	3	5	6	7
1	16	20	5	0	0	1
2	2	5	0	2	0	0
3	1	2	0	0	0	0
5	0	3	0	1	0	1
6	1	3	0	0	2	0
7	1	0	0	0	0	5

Porównanie wyników

Umieścimy teraz w tabeli błędy klasyfikacji dla najlepszych modeli uzyskanych poprzednimi metodami.

	5-cv	bootstrap	.632+	predykcje na zb. testowym
błąd klasyfikacji (k-nn)	0.295	0.302	0.219	0.324
błąd klasyfikacji (dk)	0.345	0.330	0.301	0.324
błąd klasyfikacji (NB)	0.542	0.567	0.538	0.592

Tabela 24: Porównanie uzyskanych klasyfikatorów

Widzimy, że najlepsze rezultaty uzyskaliśmy metodą k-najbliższych sąsiadów, natomiast najgorsze z wykorzystaniem naiwnego klasyfikatora bayesowskiego.

4 Podsumowanie

Poniżej wypunktujemy najważniejsze wnioski, jakie można wyciągnąć z przeprowadzanych analiz:

- uwzględnienie składników wielomianowych 2 stopnia w modelu regresji liniowej pozwala uzyskać zdecydowanie lepszą dokładność klasyfikacji,
- nieuwzględnienie cech o najgorszych zdolnościach dyskryminacyjnych pozytywnie wpłynęło na dokładność klasyfikatorów uzyskanych metodami k-najbliższych sąsiadów, drzew klasyfikacyjnych oraz naiwnego klasyfikatora beysowskiego,
- metody k-cross validation, bootstrap, .632+ oraz macierze pomyłek pozwalają na porównywanie klasyfikatorów uzyskanych różnymi metodami,
- wybór odpowiedniej liczby sąsiadów ma kluczowe znaczenie dla otrzymanych rezultatów metodą k-nn,
- wybór optymalnego parametru `cp` na podstawie kryterium kosztu złożoności sprawia, że możemy łatwiej szacować błędy predykcji.