Raport Lista 3

Eksploracja danych

Dominik Kowalczyk i Matylda Mordal

2025-05-28

Spis treści

1 Klasyfikacja na bazie modelu regresji liniowej

1.1 Analizowane dane

Zbiór danych `iris`, dostępny w pakiecie `datasets` języka R, stanowi klasyczny przykład wykorzystywany w analizie statystycznej i uczeniu maszynowym. Dane te pochodzą z pracy Ronalda Fishera z 1936 roku i opisują trzy gatunki irysów: `setosa`, `versicolor` oraz `virginica`. Każdy z gatunków reprezentowany jest przez 50 obserwacji, co daje łącznie 150 przypadków. Oznacza to, że mamy do czynienia z problemem klasyfikacyjnym z $K = 3$ klasami.

Zmiennymi objaśniającymi są cztery pomiary ilościowe, wyrażone w centymetrach: długość i szerokość działki kielicha (`Sepal.Length`, `Sepal.Width`) oraz długość i szerokość płatka (`Petal.Length`, `Petal.Width`). Oznaczmy je skrótowo jako: SL (`Sepal.Length`), SW (`Sepal.Width`), PL (`Petal.Length`) oraz PW (`Petal.Width`). Łącznie mamy więc $p = 4$

zmienne objaśniające, które tworzą wektor cech: $\mathbf{X} = (PL, PW, SL, SW)$. Piątą zmienną w zbiorze jest `Species`, która pełni rolę zmiennej objaśnianej i wskazuje gatunek irysa, przypisując jedną z trzech klas.

Zbiór danych jest kompletny i nie zawiera żadnych brakujących wartości, co czyni go gotowym do bezpośredniego wykorzystania w zadaniach klasyfikacyjnych. Szczegółowe informacje o liczbie obserwacji dla poszczególnych gatunków oraz o typach zmiennych znajdują się odpowiednio w Tabelach 1 i 2.

Tabela 1: Liczba obserwacji dla poszczególnych gatunków w zbiorze iris

Gatunek	Liczba obserwacji
setosa	50
versicolor	50
virginica	50

Tabela 2: Typy zmiennych w zbiorze iris

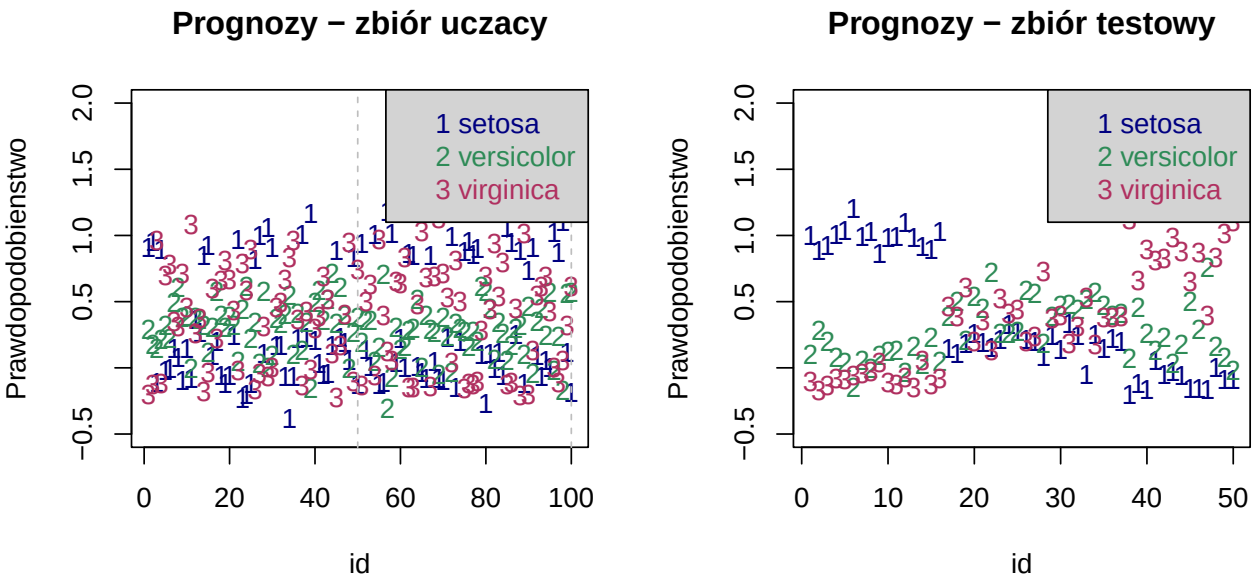
Indeks	Nazwa zmiennej	Typ zmiennej	Opis zmiennej
1	Sepal.Length	numeric	Długość działki kielicha w cm
2	Sepal.Width	numeric	Szerokość działki kielicha w cm
3	Petal.Length	numeric	Długość płatka w cm
4	Petal.Width	numeric	Szerokość płatka w cm
5	Species	factor	Gatunek rośliny (setosa, versicolor, virginica)

1.2 Podział danych na zbiór uczący i testowy

Podzieliliśmy losowo dane z zestawu `iris` na zbiór uczący (2/3, 100 obserwacji) i testowy (1/3, 50 obserwacji), ustawiając ziarno `set.seed(123)` dla powtarzalności. Dzięki temu uzyskaliśmy dwa niezależne podzbiory, które posłużą do analizy danych.

1.3 Konstrukcja klasyfikatora i wyznaczenie prognoz

Stworzyliśmy trzy oddzielne modele regresji liniowej (po jednym dla każdego gatunku irysa, przewidujące przynależność 0-1). Używając tych modeli, wyznaczyliśmy prawdopodobieństwa przynależności do każdej z klas dla zbioru uczącego i testowego, przy przypisaliśmy klasę z najwyższym prognozowanym prawdopodobieństwem. Wyniki tych predykcji wizualizują Wykresy prognozowanych prawdopodobieństw (Rysunek 1).



Rysunek 1: Wykresy prognozowanych prawdopodobieństw

Dokładność klasyfikacji wyniosła 81% na zbiorze uczącym i 84% na testowym, co pokazuje skuteczność naszego klasyfikatora opartego na regresji liniowej.

1.4 Ocena jakości modelu

1.4.1 Macierz pomyłek i z błąd klasyfikacji

Wyznaczyliśmy macierze pomyłek oraz błędy klasyfikacji dla zbioru uczącego i testowego. Macierz pomyłek dla zbioru uczącego (Tabela 3) pokazuje, że model poprawnie sklasyfikował wszystkie 34 obserwacje klasy `setosa`, ale pomylił się w przypadku `versicolor` (15 z 29 sklasyfikowano jako `virginica`) oraz `virginica` (4 z 37 sklasyfikowano jako `versicolor`). Dla zbioru testowego (Tabela 4) model również idealnie rozpoznał `setosa` (16/16), ale pomylił się przy `versicolor` (7 z 21 sklasyfikowano jako `virginica`) i `virginica` (1 z 13 sklasyfikowano jako `versicolor`).

Tabela 3: Macierz pomyłek dla zbioru uczącego

	setosa	versicolor	virginica
setosa	34	0	0
versicolor	0	14	15
virginica	0	4	33

Tabela 4: Macierz pomyłek dla zbioru testowego

	setosa	versicolor	virginica
setosa	16	0	0
versicolor	0	14	7
virginica	0	1	12

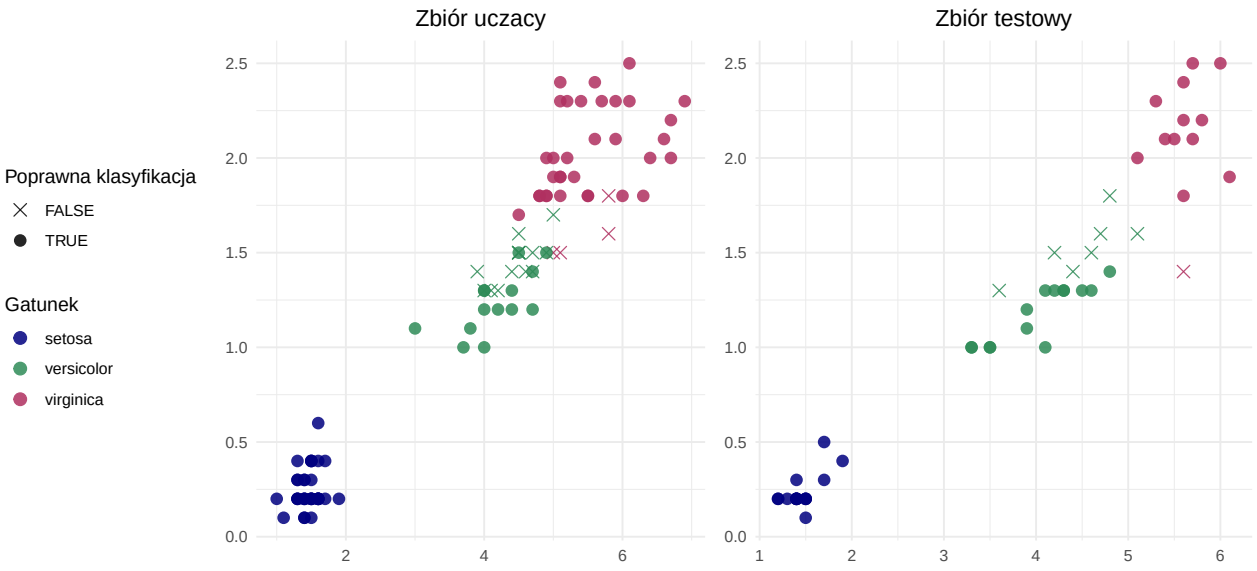
Błąd klasyfikacji wynosi 19% dla zbioru uczącego i 16% dla zbioru testowego, co wskazuje na pewien poziom błędnych predykcji w obu zestawach danych.

1.4.2 Zjawisko maskowania klas

W celu zbadania potencjalnego zjawiska maskowania klas, stworzyliśmy wizualizację wyników klasyfikacji na zbiorze uczącym i testowym, przedstawioną na Wykresach maskowania klas (Rysunek 2). Punkty są kolorowane zgodnie z rzeczywistym gatunkiem irysa, a ich kształt wskazuje, czy klasyfikacja była poprawna (kółko) czy błędna (krzyżyk).

Wykresy (Rysunek 2) pokazują, że `setosa` jest dobrze sklasyfikowana w obu zbiorach. Większość błędów występuje między nakładającymi się obszarami `versicolor` i `virginica`, co sugeruje trudności regresji liniowej w ich rozdzieleniu.

Na podstawie wizualizacji, nie obserwujemy sytuacji, w której jedna z klas byłaby całkowicie maskowana i nigdy nie przewidywana. Jednakże, nakładanie się obszarów `versicolor` i `virginica` w przestrzeni cech oraz występowanie błędnych klasyfikacji między tymi dwoma gatunkami sugeruje, że regresja liniowa ma trudności z wyznaczeniem idealnej granicy decyzyjnej między nimi. Można to interpretować jako częściowe maskowanie lub po prostu jako problem z liniową separowalnością tych dwóch klas.



Rysunek 2: Wykresy maskowania klas: poprawne i błędne klasyfikacje

1.5 Budowa modelu liniowego dla rozszerzonej przestrzeni cech

1.5.1 Rozszerzenie przestrzeni cech

W celu poprawy jakości klasyfikacji, dokonano rozszerzenia oryginalnej przestrzeni cech, dodając do danych dodatkowe zmienne będące składnikami wielomianowymi stopnia drugiego. Oprócz podstawowych czterech cech: długości i szerokości działki kielicha (SL, SW) oraz długości i szerokości płatka (PL, PW), utworzono dziesięć nowych zmiennych: kwadraty każdej z cech (PL2, PW2, SL2, SW2) oraz wszystkie możliwe iloczyny par różnych cech. Dzięki temu model może uchwycić nieliniowe zależności między zmiennymi i skuteczniej rozróżniać klasy, ponieważ uwzględnia bardziej złożone powiązania między cechami.

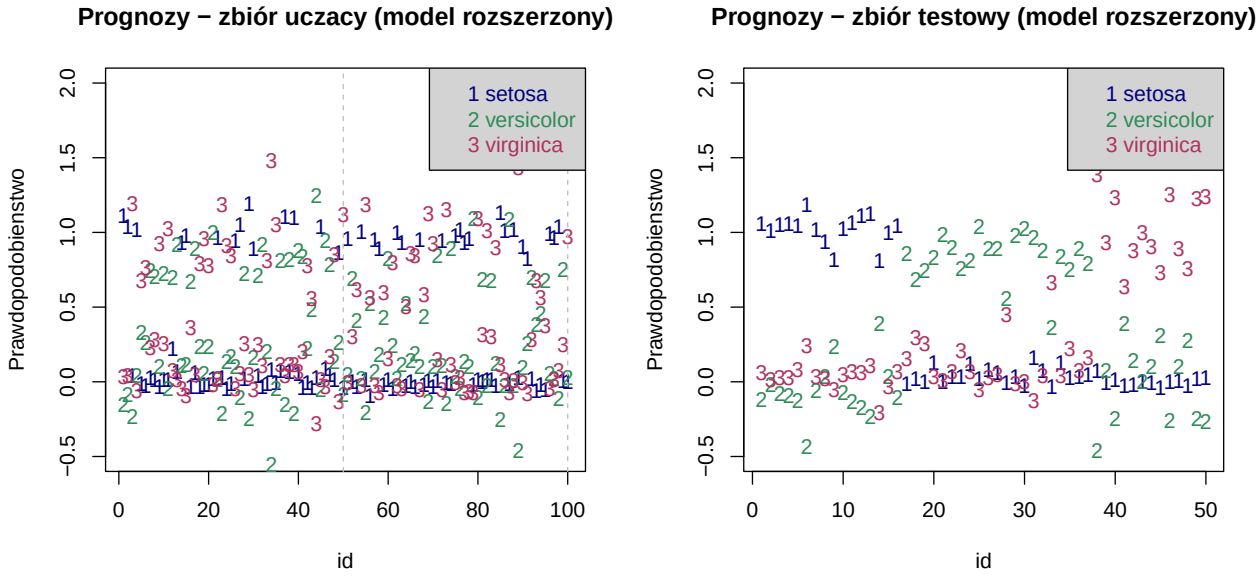
1.5.2 Konstrukcja klasyfikatora i wyznaczenie prognoz dla rozszerzonych cech

Po rozszerzeniu przestrzeni cech o składniki wielomianowe stopnia drugiego, przystąpiono do konstrukcji klasyfikatora. Podobnie jak w przypadku modelu podstawowego, dla każdej z trzech klas Iris) zbudowano oddzielny model regresji liniowej.

Wyznaczyliśmy prognozy klas oraz odpowiadające im prawdopodobieństwa przynależności do danej klasy zarówno dla zbioru uczącego, jak i testowego. Wyniki przedstawiono graficznie na Rysunku 3, który pokazuje rozkład prawdopodobieństw klasyfikacji obserwacji według modelu rozszerzonego.

Tabela 5: Porównanie dokładności modeli

Model	Zbiór Uczący	Zbiór Testowy
Podstawowy	81%	84%
Rozszerzony	99%	98%



Rysunek 3: Wykresy prognozowanych prawdopodobieństw dla rozszerzonego modelu

Zestawienie dokładności obu modeli przedstawia Tabela 5. Model podstawowy osiągnął dokładność na poziomie 81% dla zbioru uczącego i 84% dla zbioru testowego, natomiast model rozszerzony znacząco przewyższył go, uzyskując 99% i 98% odpowiednio. Tak znaczna poprawa potwierdza skuteczność rozszerzenia przestrzeni cech w kontekście klasyfikacji danych o złożonej strukturze.

1.5.3 Ocena jakości modeli

Ocena jakości klasyfikatorów została przeprowadzona na podstawie macierzy pomyłek oraz wizualizacji poprawnych i błędnych klasyfikacji, przedstawionych na Rysunku 4. Szczegółowe wyniki zawarto w Tabelach 6 i 7 dla modelu rozszerzonego oraz w Tabelach 3 i 4 dla modelu podstawowego.

Z porównania macierzy pomyłek wynika, że model rozszerzony niemal idealnie klasyfikuje dane, zarówno w zbiorze uczącym, jak i testowym liczba błędnych przypisań została zredukowana do minimum. W Tabeli 6 błędnie zaklasyfikowano tylko jeden przypadek gatunku `virginica`, a w Tabeli 7 pomyłka pojawiła się jedynie raz w przypadku `versicolor`. Dla porównania, model podstawowy (Tabela 3 i 4) miał znacznie więcej błędnych klasyfikacji między klasami `versicolor` i `virginica`.

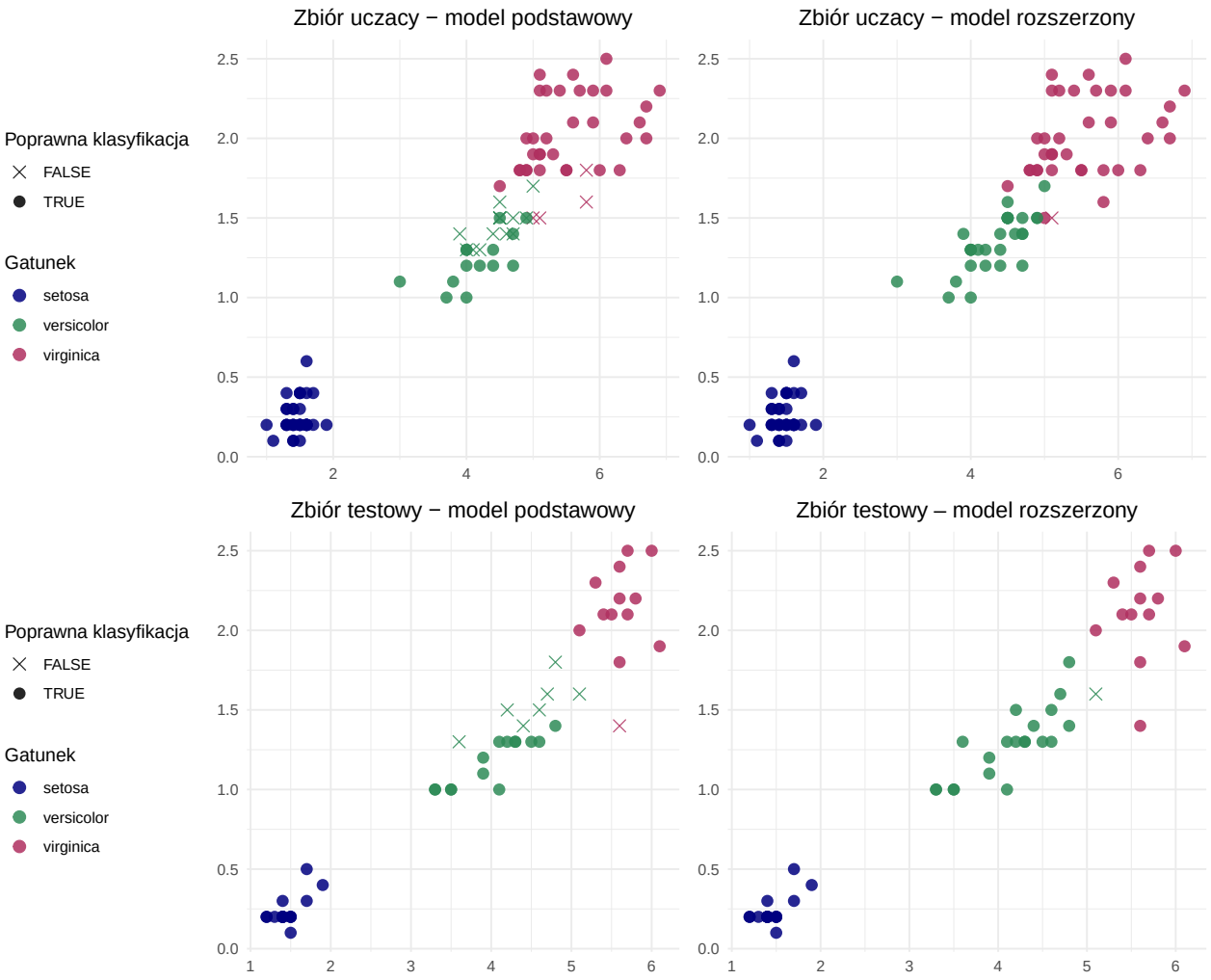
Rysunek 4 wizualnie ilustruje, że po zastosowaniu rozszerzenia przestrzeni cech, błędy klasyfikacji w modelu rozszerzonym niemal znikają, co wizualnie potwierdza wyraźniejszy podział klas.

Tabela 6: Macierz pomyłek dla rozszerzonego zbioru uczącego

	setosa	versicolor	virginica
setosa	34	0	0
versicolor	0	29	0
virginica	0	1	36

Tabela 7: Macierz pomyłek dla rozszerzonego zbioru testowego

	setosa	versicolor	virginica
setosa	16	0	0
versicolor	0	20	1
virginica	0	0	13



Rysunek 4: Wizualizacja poprawnych i błędnych klasyfikacji: model podstawowy i model rozszerzony

Błąd klasyfikacji dla rozszerzonego zbioru uczącego wynosi 1%, natomiast dla rozszerzonego zbioru testowego 2%. Wyniki te są znacznie niższe niż w modelu podstawowym, co potwierdza znaczną poprawę jakości klasyfikacji.

1.5.4 Analiza wyników

Rozszerzenie przestrzeni cech znacząco poprawiło jakość klasyfikacji. Model liniowy, wzbogacony o składniki wielomianowe drugiego stopnia, uchwycił nieliniowe zależności między zmiennymi, co przełożyło się na wyraźny wzrost dokładności i niemal całkowitą eliminację błędów klasyfikacji. W szczególności zredukowano problemy z rozróżnianiem gatunków versicolor i virginica, które były widoczne w modelu podstawowym. Wizualizacje potwierdzają, że granice między klasami stały się wyraźniejsze, a przypisania znacznie bardziej trafne.

Zatem, zastosowanie składników wielomianowych drugiego stopnia w przestrzeni cech okazało się skutecznym podejściem, pozwalającym modelowi liniowemu na lepsze modelowanie

złożonych zależności danych `Iris` i osiągnięcie znacznie wyższej dokładności klasyfikacji.

2 Porównywanie metod klasyfikacji

2.1 Przygotowanie danych

Do wykonania zadania wykorzystamy zbiór danych Glass (mlbench). Opisuje on dane identyfikacyjne szkła używane oraz potrzebne przy śledztwach kryminalistycznych. Przyjrzyjmy się zatem, co znajduje się w analizowanym zbiorze.

Tabela 8: Opis danych Glass

Indeks	Nazwa zmiennej	Typ zmiennej	Opis zmiennej
1	RI	numeric	Współczynnik załamania światła
2	Na	numeric	Zawartość sodu
3	Mg	numeric	Zawartość magnezu
4	Al	numeric	Zawartość glinu
5	Si	numeric	Zawartość krzemu
6	K	numeric	Zawartość potasu
7	Ca	numeric	Zawartość wapnia
8	Ba	numeric	Zawartość baru
9	Fe	numeric	Zawartość żelaza
10	Type	factor	Typ szkła

Liczba przypadków w zbiorze danych `glass_data` wynosi 214.

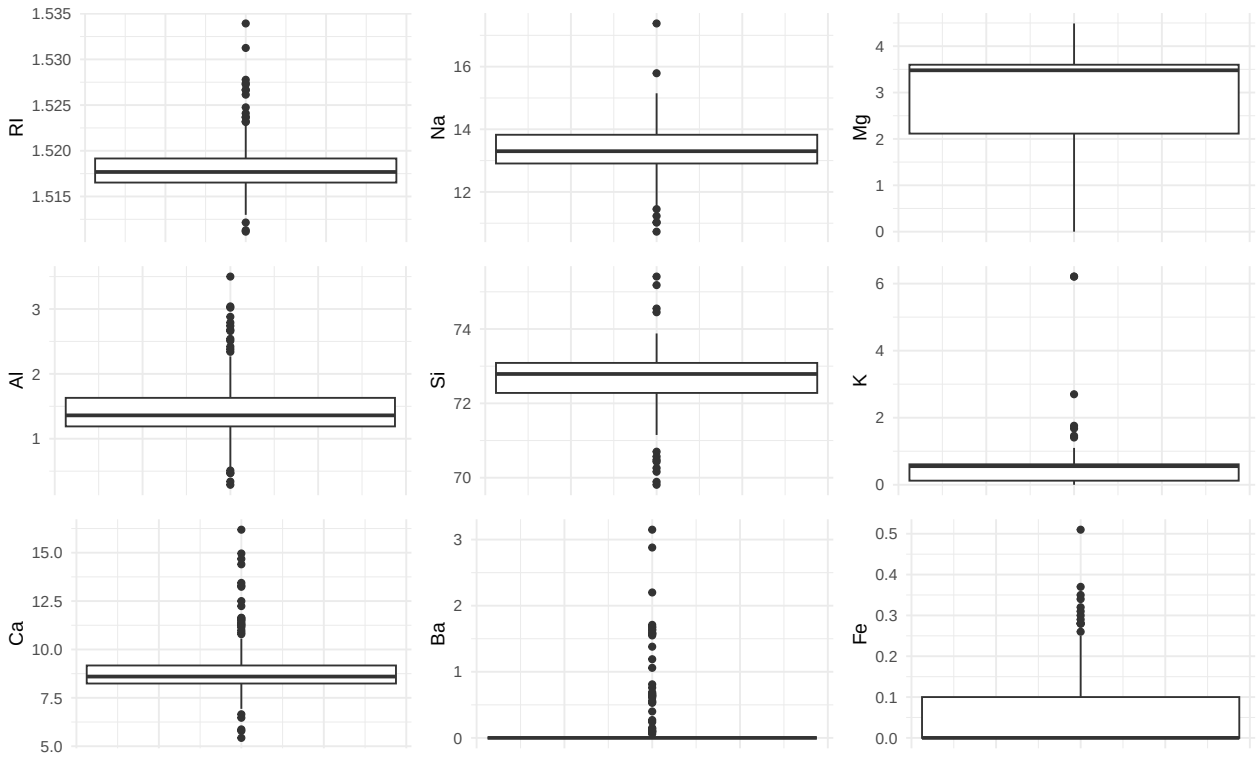
Zbiór danych `glass_data` zawiera 10 zmiennych, z czego ostatnia z nich, `Type` przechowuje informacje o przynależności obiektu do konkretnej klasy (tzw. etykieta klas). Jest ona typu: factor. Dodatkowo, pozostałe zmienne zawierają informację dotyczące występowania danego piewiastka chemicznego w szkle i są one typu numeric, co pozwala nam stwierdzić, że wszystkie zmienne w naszym analizowanym zbiorze mają prawidłowo przypisane typy.

Tabela 9: Liczba Obserwacji dla Każdego Typu Szkła

Typ.Szkła	Liczba.Obserwacji
1	70
2	76
3	17
5	13
6	9
7	29

```r
#Czy istnieją jakieś braki w danych?
any(is.na(glass_data)) ||
any(sapply(glass_data, function(col) is.character(col)
           & (col == "" | grepl(" ", col))))
```

[1] FALSE - Zatem nasz zbiór danych jest kompletny i nie wsytępują tam żadne braki danych. Sprawdźmy rozkład danych w celu zrozumienia z czym mamy doczynienia, jak również poszukując nieścisłości lub różnego rodzaju nietypowych wartości.



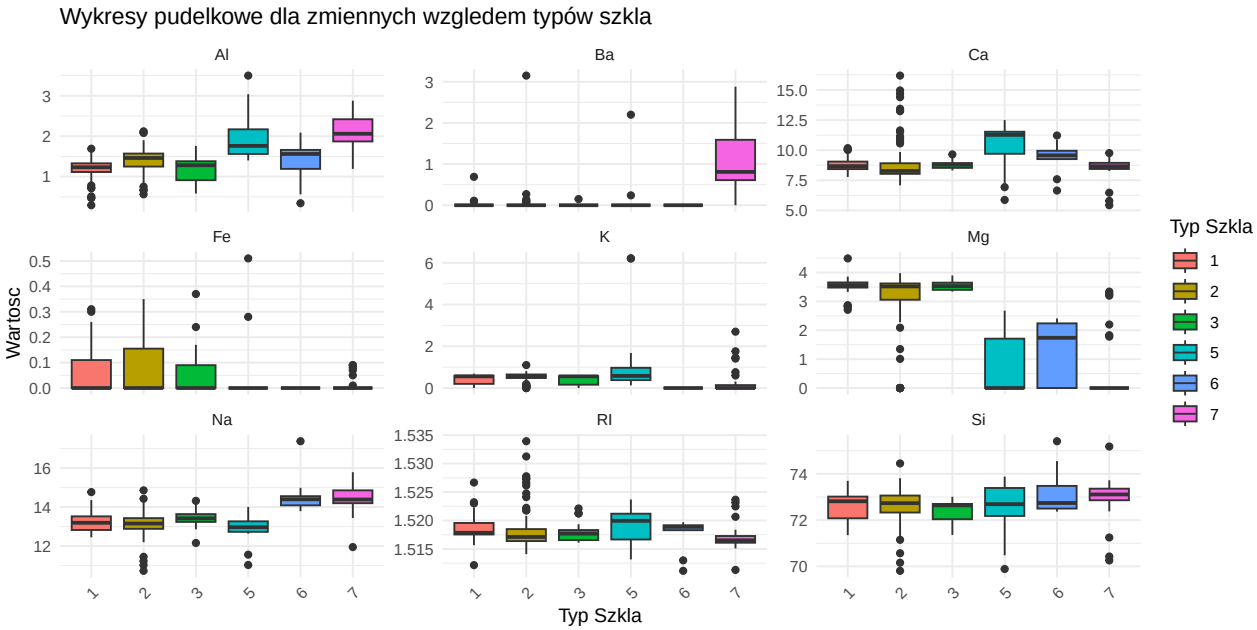
Rysunek 5: Wykresy pudełkowe względem zmiennych

W analizowanym zbiorze nie mamy doczynienia z "nieścisłościami" rozkładów w sensie błędów w danych, ale raczej charakterystycznymi cechami chemicznymi różnych typów szkła. Dziwne rozkłady (silnie skośne, z licznymi odstającymi) dla takich pierwiastków jak Mg, K i Ba są prawdopodobnie wynikiem specyficznych receptur chemicznych stosowanych do wytwarzania różnych rodzajów szkła o odmiennych właściwościach (np. szkło budowlane vs. szkło optyczne vs. szkło kryształowe), szczególnie że ilość obserwacji dla każdego typu szkła znacznie się różni (dla klasy 1 jest 70 a dla 6 tylko 9).

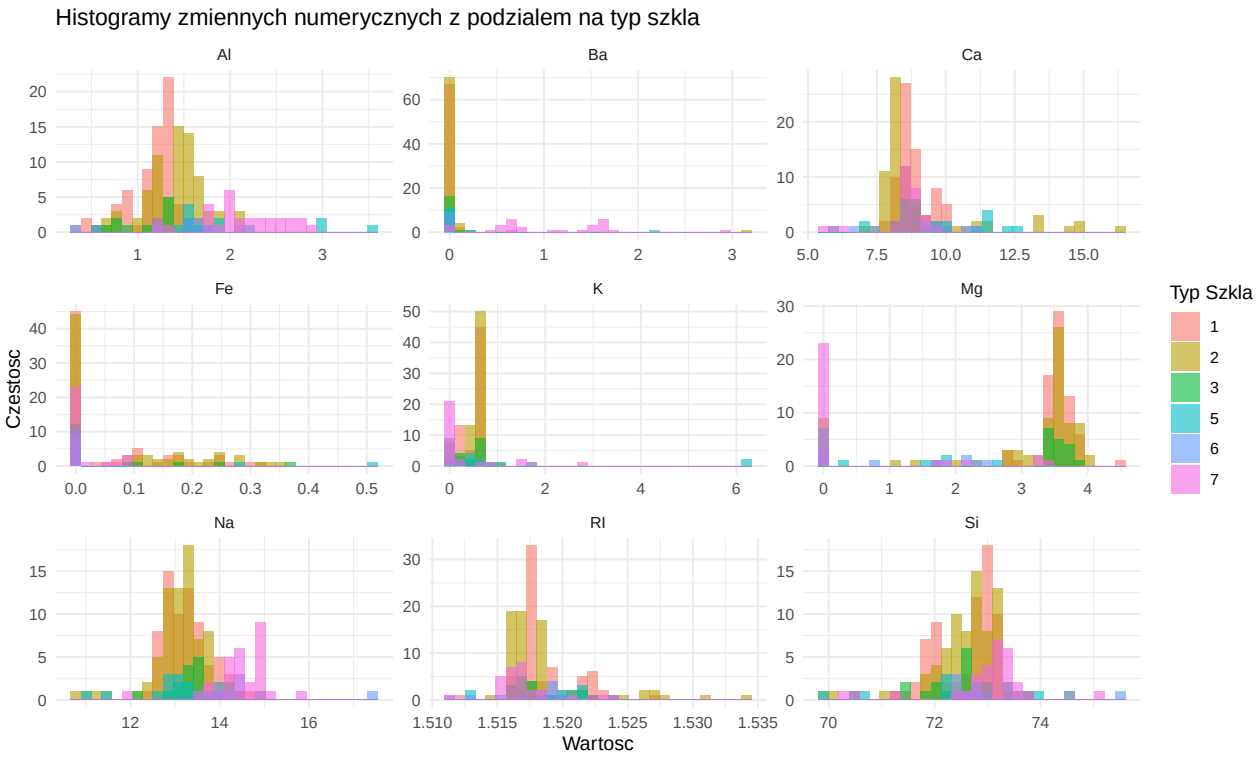
2.2 Wstępna analiza danych

Przed budową modeli klasyfikacyjnych przyjrzyjmy się analizowanym danym, zwracając uwagę m.in. na ich charakterystyczne własności oraz spróbujmy (wstępnie) ocenić zdolności dyskryminacyjne (predykcyjne) poszczególnych zmiennych/cech.

W powyższym podpunkcie przeanalizowaliśmy dane w oparciu o wykresy pudełkowe. Spójrzmy teraz, co możemy wywnioskować z histogramów oraz wykresów pudełkowych.



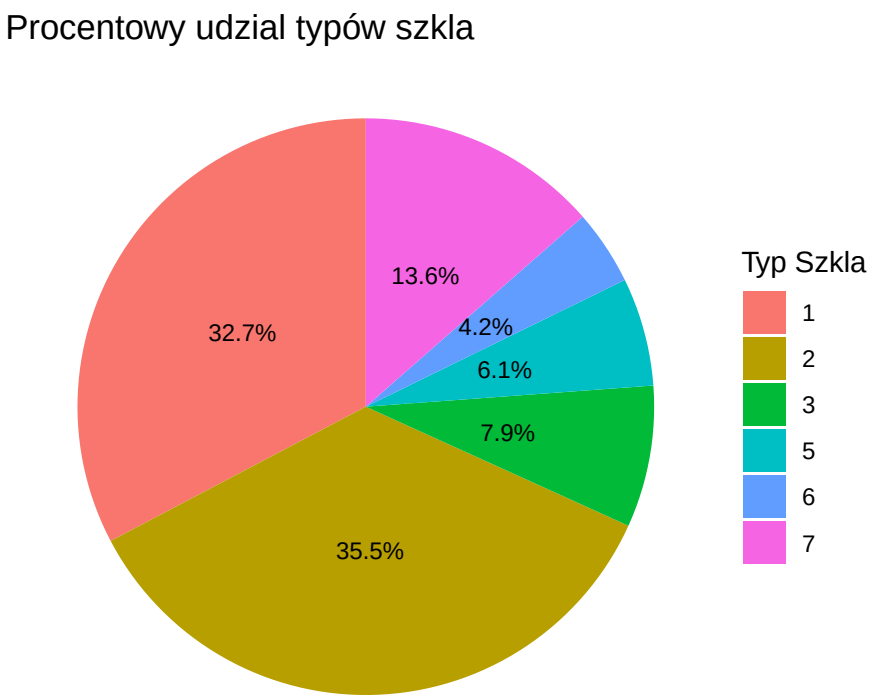
Rysunek 6: Wykresy pudełkowe dla zmiennych względem typów szkła



Rysunek 7: Histogramy dla danych zbioru Glass

Na podstawie analizy histogramów (Rysunek 7) oraz wykresów pudełkowych (Tabela 6) zbioru danych `Glass` możemy wstępnie ocenić zdolności dyskryminacyjne poszczególnych zmiennych chemicznych w kontekście rozróżniania typów szkła. Magnez (`Mg`) wydaje się być silnym predyktorem, ponieważ typ 1 charakteryzuje się znacznie wyższą zawartością `Mg` w porównaniu do typów 2 i 3, które mają niskie jego wartości. Bar (`Ba`) również wykazuje duży potencjał dyskryminacyjny, szczególnie w identyfikacji typów 2 i 3 (oraz potencjalnie 7), które mają tendencję do posiadania wyższych wartości `Ba`, w przeciwieństwie do pozostałych typów z niską zawartością. Podobnie, Potas (`K`) silnie wyróżnia typ 2, który zawiera próbki o znacznie wyższych stężeniach `K`. Glin (`Al`) także przyczynia się do rozróżnienia, z typami 2 i 6 generalnie wykazującymi wyższe wartości `Al` niż typy 1 i 3.

Wapń (`Ca`) i Sód (`Na`) zdają się mieć umiarkowaną moc dyskryminacyjną. Rozkłady ich wartości dla różnych typów szkła częściowo się pokrywają, ale widoczne są pewne różnice w centralnej tendencji, na przykład typ 2 ma tendencję do niższych wartości Ca i wyższych Na. Żelazo (`Fe`), ze względu na ogólnie niskie i skupione wartości, prawdopodobnie ma ograniczoną zdolność do rozróżniania typów szkła, chociaż wyższe wartości w niektórych próbkach mogą być specyficzne dla pewnych typów. Krzem (`Si`), jako główny składnik szkła, wykazuje niewielką zmienność między typami, sugerując ograniczoną moc dyskryminacyjną, chociaż subtelne różnice mogą być istotne. Współczynnik załamania światła (`Ri`) również wydaje się mieć umiarkowany potencjał predykcyjny, z niewielkimi różnicami w rozkładach między typami.



Rysunek 8: Wykres kołowy wkładu danej klasy w zbiór danych Glass (procentowy)

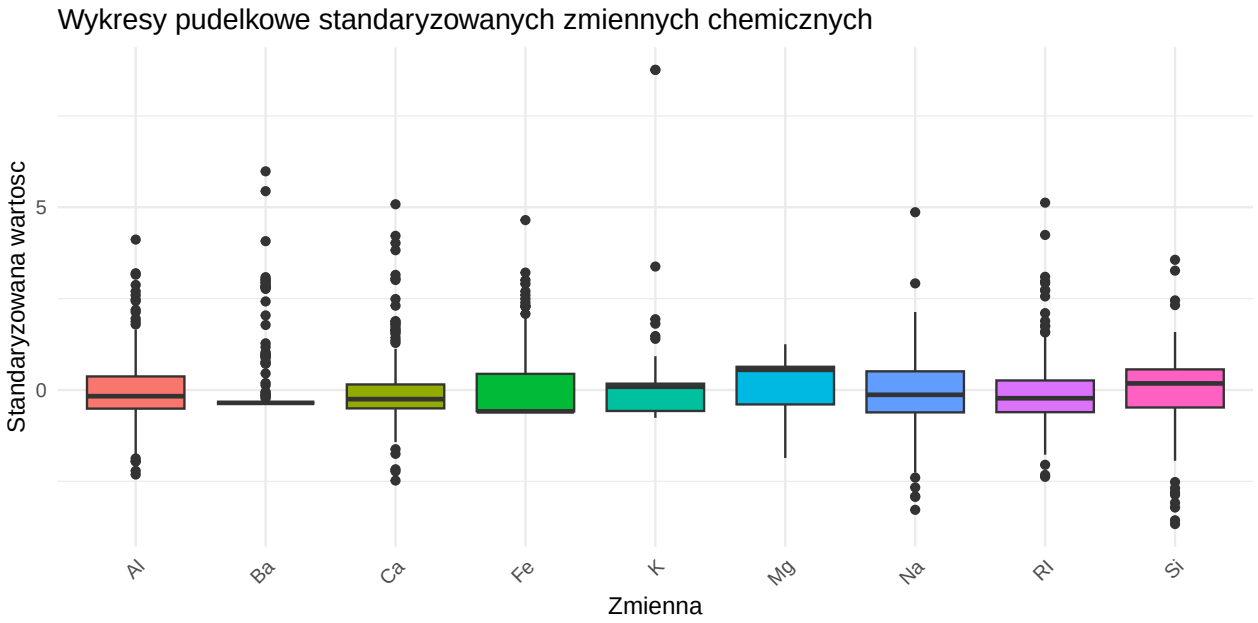
Tabela 10: Błąd Klasyfikacji przy Przypisaniu Wszystkich do Najczęstszej Klasy

Najczęściej.występująca.klasa	Procent.obserwacji	Błąd.klasyfikacji
2	35.51%	64.49%

Tabela 11: Wariancje poszczególnych zmiennych numerycznych

Zmienna	Wariancja
RI	0.0000092
Na	0.6668414
Mg	2.0805404
Al	0.2492702
Si	0.5999212
K	0.4253542
Ca	2.0253658
Ba	0.2472270
Fe	0.0094943

Tabela 11 przedstawiająca wartości wariancji analizowanych zmiennych pokazuje, że należy zastosować standaryzację naszych zmiennych jako, że wartości wahają się od 0,0000092 dla `RI` do ponad 2 dla `Mg` oraz `Ca`. Zatem zastosujmy standaryzację i zwizualizujmy ją na wykresie pudełkowym.

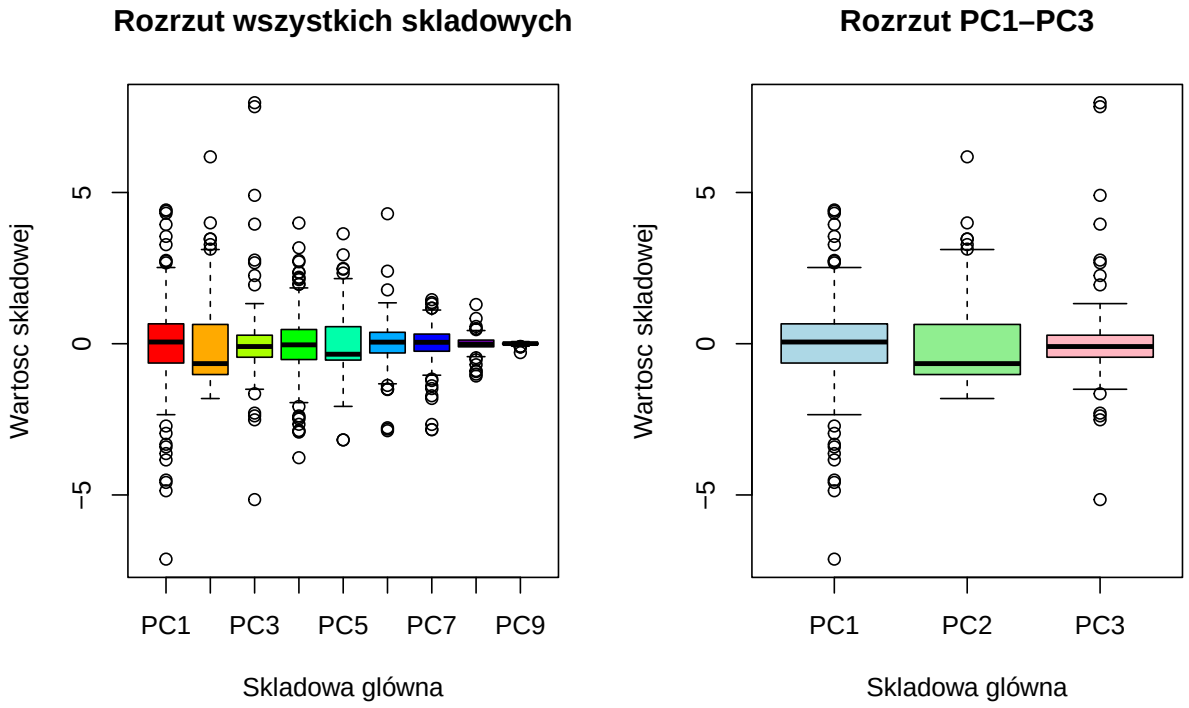


Rysunek 9: Wykres pudełkowy po standaryzacji zmiennych zbioru Glass

Na podstawie Rysunku 6, Rysunku 7, jak również 11 możemy wywnioskować, że zmiennymi reprezentującymi najlepsze zdolności dyskryminacyjne/predykcyjne są `Mg`, `Al`, `Na`, `K` oraz `Ba`, gdyż charakteryzują się znaczącymi odstępstwami wartości pomiędzy różnymi klasami.

2.3 ***PCA (dodatkowe)

Mając ustandaryzowane dane możemy przeprowadzić analizę PCA.



Rysunek 10: Rozrzut składowych głównych

Na Rysunkach 10 widzimy rozrzuty danej składowej. Możemy dostrzec, że mamy doczynienia z dosyć nietypową sytuacją, gdyż nie możemy powiedzieć, że PC1 charakteryzuje się największym rozrzutem wykresu pudełkowego. Jednakże, jest na to wytłumaczenie, gdyż jak widać na Rysunku 6 oraz 9 można wskazać obecność dużej ilości wartości odstających, które mogą "rozciągać" wąsy wykresu pudełkowego, sprawiając, że cała składowa ma większy rozrzut, podczas gdy główna masa danych jest nadal mniejsza niż dla PC1. Aby to potwierdzić sprawdźmy wariancje.

Tabela 12: Wariancje i proporcje wyjaśniane przez składowe główne

Składowa	Wariancja	Proporcja Wariancji	Skumulowana Proporcja
PC1	2.5112	0.2790	0.2790
PC2	2.0501	0.2278	0.5068

PC3	1.4048	0.1561	0.6629
PC4	1.1579	0.1287	0.7915
PC5	0.9140	0.1016	0.8931
PC6	0.5276	0.0586	0.9517
PC7	0.3690	0.0410	0.9927
PC8	0.0639	0.0071	0.9998
PC9	0.0016	0.0002	1.0000

Zatem mamy potwierdzone na Tabeli 12, że kolejność składowych jest prawidłowa.

Przyjrzyjmy się głębiej interesującymi nas składowymi głównymi: PC1, PC2, PC3:

Tabela 13: Największe obciążenia zmiennych na PC1

Zmienna	Loading
RI	-0.545
Ca	-0.492
Al	0.429
Na	0.258
Ba	0.250
Si	0.229

Tabela 14: Największe obciążenia zmiennych na PC2

Zmienna	Loading
Mg	-0.594
Ba	0.485
Ca	0.345
Al	0.295
RI	0.286
Na	0.270

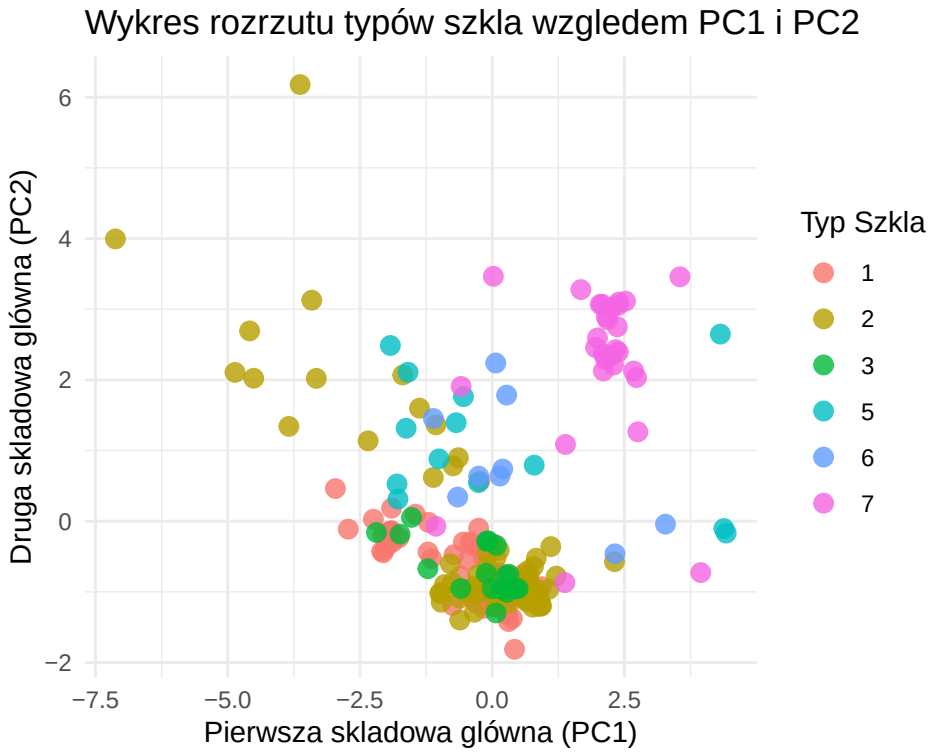
Tabela 15: Największe obciążenia zmiennych na PC3

Zmienna	Loading
K	0.663
Si	-0.459
Na	-0.385
Al	0.329
Fe	0.284
RI	0.087

PC1 Ta składowa może reprezentować oś kontrastu między szkłami o wysokim `RI` i zawartości wapnia a szkłami o wyższej zawartości aluminium, sodu, baru i krzemu. Może to odzwierciedlać ogólną "typologię" szkła w zależności od jego podstawowego składu masowego i wynikających z niego właściwości optycznych (`RI`)

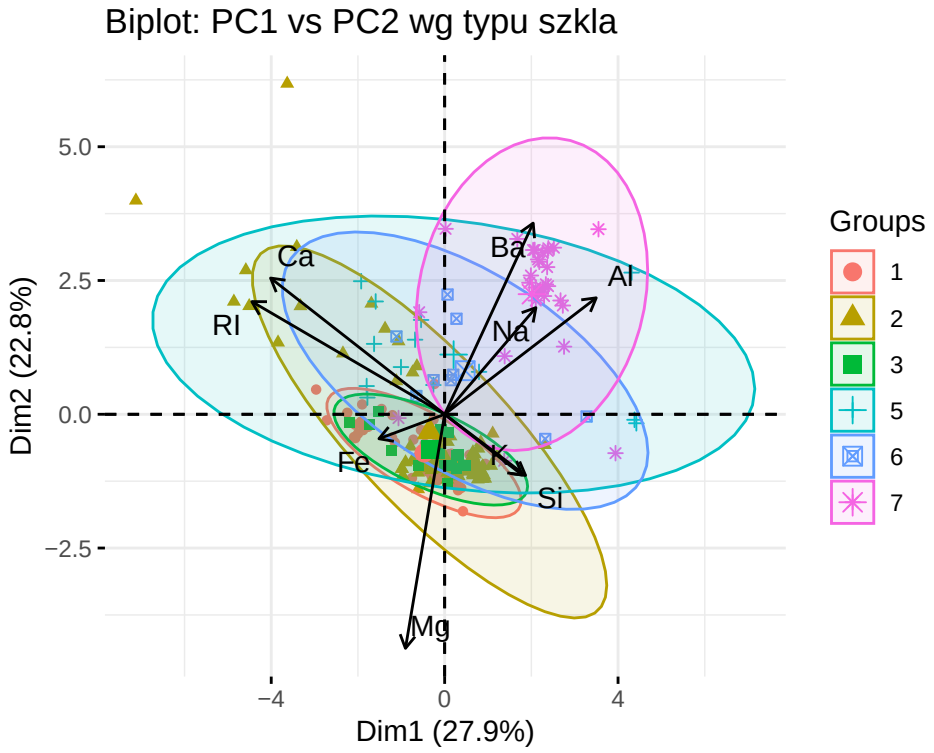
PC2 Ta składowa wydaje się wychwytywać zmienność w składzie szkła, która jest niezależna od PC1 i która głównie różnicuje szkła pod względem zawartości magnezu versus baru i wapnia. Może to być kluczowe dla rozróżnienia rodzajów szkła, gdzie Mg jest używane jako stabilizator, a `Ba` i `Ca` w innych specyficznych zastosowaniach lub do uzyskania innych właściwości.

PC3 Ta składowa wyraźnie oddziela szkła o wysokiej zawartości potasu od tych o wysokiej zawartości krzemu i sodu. Jest to szczególnie interesujące, ponieważ potas i sód są często używane zamiennie w produkcji szkła, wpływając na jego właściwości takie jak temperatura topnienia i lepkość. PC3 może zatem odzwierciedlać zmienność związaną z substytucją alkaliów (potasu za sód) oraz rolą krzemu jako głównego składnika strukturalnego. Obecność żelaza (`Fe`) może wskazywać na rolę zanieczyszczeń lub specyficznych barwników.



Rysunek 11: Wykres rozrzutu dla głównych składowych zbioru danych Glass

Rysunek 11, obrazuje pozycje poszczególnych próbek szkła w dwuwymiarowej przestrzeni zdefiniowanej przez dwie pierwsze składowe główne. Analizując ten wykres, można dostrzec zróżnicowanie w separacji poszczególnych typów szkła. Typy 1 i 2, będące dominującymi w zbiorze danych, wykazują znaczne nakładanie się w centralnej części wykresu, co wskazuje, że same PC1 i PC2 nie są wystarczające do ich pełnego rozróżnienia. Z kolei typy 7 i 5 prezentują się jako względnie dobrze oddzielone klastry, odpowiednio w lewej górnej i prawej górnej części wykresu. Typy 3 i 6, z mniejszą liczbą obserwacji, są bardziej rozproszone i częściowo nakładają się na inne grupy. Na wykresie widoczne są również pojedyncze punkty, często dla typów 2 i 7, które znacząco odbiegają od głównych skupisk swoich typów, co potwierdza obecność wartości odstających.



Rysunek 12: Biplot dla głównych składowych zbioru danych Glass

Rysunek 12 wzbogaca tę analizę, łącząc rozrzut obserwacji ze strzałkami reprezentującymi oryginalne zmienne chemiczne. Strzałki wskazują kierunek wzrostu danej zmiennej w przestrzeni PC1-PC2, a ich długość odzwierciedla siłę wkładu zmiennej w wyjaśnianą wariancję. Na osi PC1, która odpowiada za 27.9% wyjaśnionej wariancji, zmienne `RI` (współczynnik załamania światła) i `Ca` (wapń) mają silne obciążenia negatywne (strzałki w lewo), co sugeruje, że próbki z wysokimi wartościami tych pierwiastków lokują się po lewej stronie osi PC1. Z drugiej strony, zmienne takie jak `Ba` (bar), `Al` (aluminium), `Na` (sód) i `Si` (krzem) mają obciążenia pozytywne (strzałki w prawo), co oznacza, że ich wysokie stężenia przesuwają próbki w prawo na osi PC1. W odniesieniu do osi PC2, która wyjaśnia 22.8% wariancji, `Mg` (magnez) jest silnie związane z ujemnymi wartościami (strzałka w dół), podczas gdy `Ba` i `Al` mają również komponenty w górę.

W sumie, PC1 i PC2 skutecznie wyróżniają niektóre typy szkła na podstawie ich unikalnego składu chemicznego (jak np. typ 7, czy 5), choć nie są wystarczające do pełnego rozdzielenia i dokładnego, jasnego scharakteryzowania wszystkich typów.

2.4 Pojedyńczy podział na zbiór uczący i testowy

W ramach naszej analizy oceniliśmy dokładność klasyfikacji trzech algorytmów: metody k-najbliższych sąsiadów (k-NN), drzew klasyfikacyjnych oraz naiwnego klasyfikatora bayesowskiego. Do tego celu wykorzystaliśmy zbiór danych `Glass`, dzieląc go na zbiór uczący (2/3 danych) i testowy (1/3 danych). Naszą ocenę oparliśmy na macierzach pomyłek i błędach

klasyfikacji. Naszym celem było nie tylko ocena skuteczności klasyfikatorów na danych uczących i testowych w oparciu o macierze pomyłek i błędy klasyfikacji, ale także zastosowanie bardziej zaawansowanych schematów oceny dokładności.

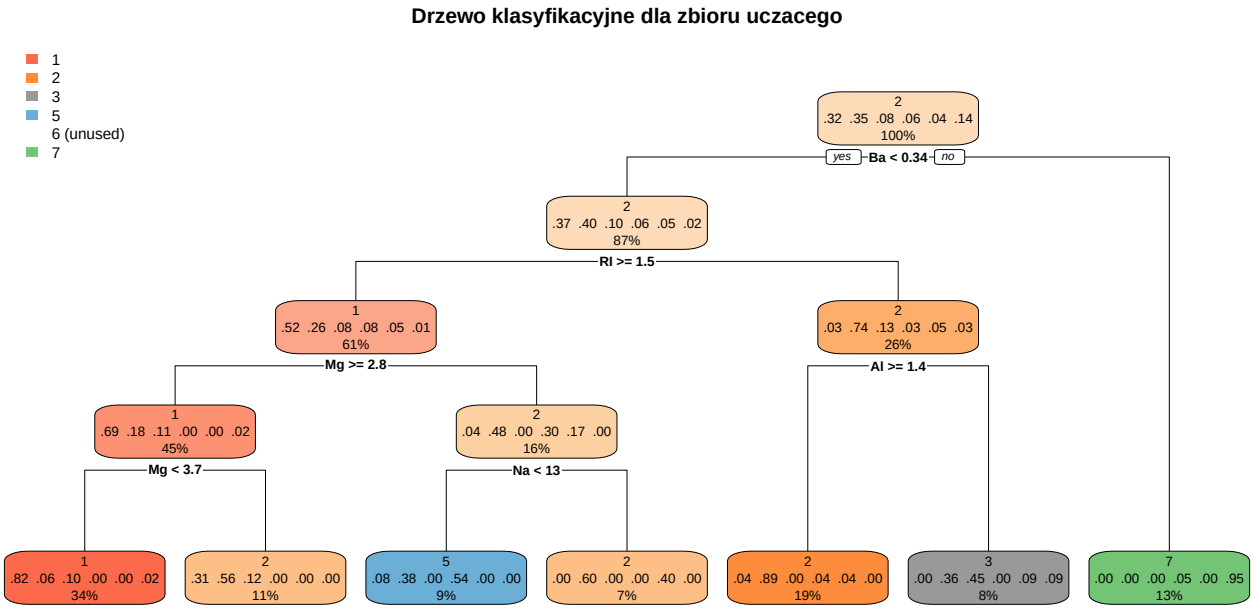
Tabela 16: Macierz pomyłek dla zbioru uczącego (K-najbliższych sąsiadów)

	1	2	3	5	6	7
1	41	6	0	0	0	0
2	8	37	1	5	0	0
3	4	3	5	0	0	0
5	0	1	0	6	0	2
6	0	1	0	0	4	1
7	1	2	0	0	0	17

Tabela 17: Macierz pomyłek dla zbioru testowego (K-najbliższych sąsiadów)

	1	2	3	5	6	7
1	15	7	1	0	0	0
2	5	19	0	1	0	0
3	4	1	0	0	0	0
5	0	1	0	2	0	1
6	0	1	0	0	2	0
7	1	0	0	0	0	8

Dla metody k-NN, z parametrem `k=5`, zaobserwowaliśmy, że macierz pomyłek dla zbioru testowego (Tabela 17) ujawniła poprawne klasyfikacje dla większości przypadków, ale także wskazała na istotne pomyłki, na przykład 7 przypadków klasy 1 zostało błędnie sklasyfikowanych jako klasa 2, a 5 przypadków klasy 2 jako klasa 1. Na zbiorze uczącym (Tabela 16) model k-NN wykazał lepsze dopasowanie. Finalnie, błąd klasyfikacji dla k-NN wyniósł 24.14% na zbiorze uczącym i 33.33% na zbiorze testowym (Tabela 22).



Rysunek 13: Drzewo klasyfikacyjne dla zbioru danych Glass

Tabela 18: Macierz pomyłek dla zbioru uczącego (Drzewo klasyfikacyjne)

	1	2	3	5	6	7
1	40	6	0	1	0	0
2	3	39	4	5	0	0
3	5	2	5	0	0	0
5	0	1	0	7	0	1
6	0	5	1	0	0	0
7	1	0	1	0	0	18

Tabela 19: Macierz pomyłek dla zbioru testowego (Drzewo klasyfikacyjne)

	1	2	3	5	6	7
1	11	8	3	0	0	1
2	7	15	0	2	0	1
3	1	2	2	0	0	0
5	0	2	0	2	0	0
6	0	3	0	0	0	0
7	0	1	0	0	0	8

Przechodząc do analizy drzew klasyfikacyjnych, zbudowaliśmy model, którego struktura została przedstawiona na Rysunku 13. Drzewo rozpoczęło podział od zmiennej dotyczącej

zawartości baru (Ba < 0.34), a kolejne rozgałęzienia opierały się na wartościach współczynnika załamania światła (RI), zawartości glinu (Al), magnezu (Mg) oraz sodu (Na). Choć klasy 1 i 2 dominowały w przewidywaniach, obecność różnych klas w poszczególnych liściach drzewa wskazuje na trudności modelu w jednoznacznym rozróżnianiu niektórych przypadków, co może prowadzić do błędów klasyfikacyjnych i sugeruje potrzebę dalszej optymalizacji.

Macierz pomyłek dla zbioru testowego (Tabela 19) potwierdziła te obserwacje, szczególnie widoczny był brak trafnych klasyfikacji dla klasy 6, której wszystkie przypadki zostały błędnie przypisane. Na zbiorze uczącym (Tabela 18) drzewo klasyfikacyjne osiągnęło lepsze wyniki, ale wciąż były widoczne pomyłki. Ostatecznie, błąd klasyfikacji dla drzewa wyniósł 24.83% na zbiorze uczącym oraz aż 44.93% na zbiorze testowym (Tabela 22). Ta różnica sugeruje, że model słabo radzi sobie z dostosowaniem do nowych danych, co sprawia, że jest mniej skuteczny niż metoda k-NN.

Tabela 20: Macierz pomyłek dla zbioru uczącego (Naive Bayes)

	1	2	3	5	6	7
1	42	3	2	0	0	0
2	33	12	1	3	2	0
3	9	0	3	0	0	0
5	0	6	0	1	1	1
6	0	0	0	0	6	0
7	1	0	0	0	8	11

Tabela 21: Macierz pomyłek dla zbioru testowego (Naive Bayes)

	1	2	3	5	6	7
1	19	1	1	0	1	1
2	16	2	0	3	4	0
3	4	0	0	0	1	0
5	0	2	0	0	2	0
6	0	0	0	0	3	0
7	0	0	0	0	5	4

Tabela 22: Porównanie błędów klasyfikacji dla różnych metod

Metoda	Błąd.uczący	Błąd.testowy
K-najbliższych sąsiadów (k=5)	24.14%	33.33%
Drzewo klasyfikacyjne	24.83%	44.93%

Metoda	Błąd.uczący	Błąd.testowy
Naiwny Bayes	48.28%	59.42%

Najsłabsze wyniki w naszej analizie uzyskał naiwny klasyfikator bayesowski. Macierz pomyłek dla zbioru uczącego (Tabela 20) oraz zbioru testowego (Tabela 21) jasno pokazała liczne pomyłki. Model ten charakteryzował się najwyższymi błędami klasyfikacji: 48.28% na zbiorze uczącym i 59.42% na zbiorze testowym (Tabela 22).

Podsumowując wyniki z pojedynczego podziału danych, metoda k-NN okazała się najskuteczniejsza z najniższym błędem na zbiorze testowym. Drzewo klasyfikacyjne, pomimo niższego błędu uczącego, miało znacznie wyższy błąd testowy, co wskazuje na słabsze dostosowanie do nowych danych. Naiwny Bayes zdecydowanie odstawał pod względem skuteczności.

Tabela 23: Porównanie błędów klasyfikacji przy użyciu zaawansowanych metod oceny skuteczności

Metoda	Cross-validation	Bootstrap	632+
K-najbliższych sąsiadów (k=5)	31.31%	35.81%	31.21%
Drzewo klasyfikacyjne	29.44%	34.72%	30.40%
Naive Bayes	61.21%	60.33%	55.81%

Aby zwiększyć wiarygodność naszych wniosków, zastosowaliśmy również bardziej zaawansowane schematy oceny dokładności, takie jak 10-krotna cross-validation, bootstrap (z 50 próbami) oraz metodę .632+ (z 50 próbami). Wyniki tych analiz są przedstawione w Tabeli 8. Zaobserwowaliśmy, że w tych zaawansowanych schematach metoda k-NN nadal utrzymywała dobrą dokładność, z błędami 31.21-35.81%, co było zgodne z początkowymi obserwacjami. Co jednak zaskakujące, drzewo klasyfikacyjne w tych bardziej wiarygodnych schematach (wartości 29.44-34.72%) wykazało się lepszą dokładnością niż w pojedynczym podziale. Sugeruje to, że początkowy pojedynczy podział mógł być niemiarodajny dla oceny drzewa. Naiwny klasyfikator bayesowski konsekwentnie osiągał najgorsze rezultaty, z błędami przekraczającymi 55.81% we wszystkich zaawansowanych metodach, co potwierdziło jego niską skuteczność dla analizowanego zbioru danych.

Wnioskujemy zatem, że wybór schematu oceny dokładności miał istotny wpływ na ocenę skuteczności drzewa klasyfikacyjnego, co podkreśla znaczenie stosowania zaawansowanych metod w celu uzyskania bardziej stabilnych i wiarygodnych estymacji błędu. Spośród użytych metod, najlepszą stabilność i najniższe estymowane błędy dla większości modeli zapewniła metoda .632+, co czyni ją szczególnie rekomendowaną do oceny modeli w sytuacjach, gdzie dostępność danych jest ograniczona lub podział losowy może prowadzić do dużych odchyleń w wynikach.

2.5 Różne parametry i różne podzbiory cech

Analiza porównawcza skuteczności wybranych metod klasyfikacyjnych została pogłębiona o badanie wpływu różnych wartości parametrów modeli oraz wyboru podzbiorów cech. Skupiono się m.in na wpływie liczby sąsiadów w metodzie k-NN oraz współczynnika złożoności drzewa (cp), rozważając zarówno pełny zbiór cech, jak i wyselekcjonowany podzbiór cech o największej zdolności dyskryminacyjnej.

Tabela 24: Porównanie błędów klasyfikacji dla różnych podzbiorów cech

Metoda	Błąd uczący	Błąd testowy
K-najbliższych sąsiadów (wszystkie cechy)	24.14%	33.33%
K-najbliższych sąsiadów (wybrane cechy)	24.83%	33.33%
Drzewo klasyfikacyjne (wszystkie cechy)	24.83%	44.93%
Drzewo klasyfikacyjne (wybrane cechy)	26.90%	37.68%
Naive Bayes (wszystkie cechy)	48.28%	59.42%
Naive Bayes (wybrane cechy)	44.14%	49.28%

Tabela 24 przedstawia porównanie błędów klasyfikacji dla różnych podzbiorów cech. Dla metody k-NN, błąd testowy pozostał na poziomie 33.33% zarówno dla wszystkich, jak i dla wybranych cech, natomiast błąd uczący nieznacznie wzrósł przy wybranych cechach (z 24.14% do 24.83%). W przypadku drzewa klasyfikacyjnego, wykorzystanie wybranych cech przyczyniło się do zmniejszenia błędu testowego z 44.93% do 37.68%, choć błąd uczący również wzrósł (z 24.83% do 26.90%). Natomiast naiwny klasyfikator bayesowski zanotował znaczną poprawę zarówno błędu uczącego (z 48.28% do 44.14%), jak i testowego (z 59.42% do 49.28%) przy użyciu wybranych cech.

Tabela 25: Zaawansowane błędy klasyfikacji dla wybranych cech

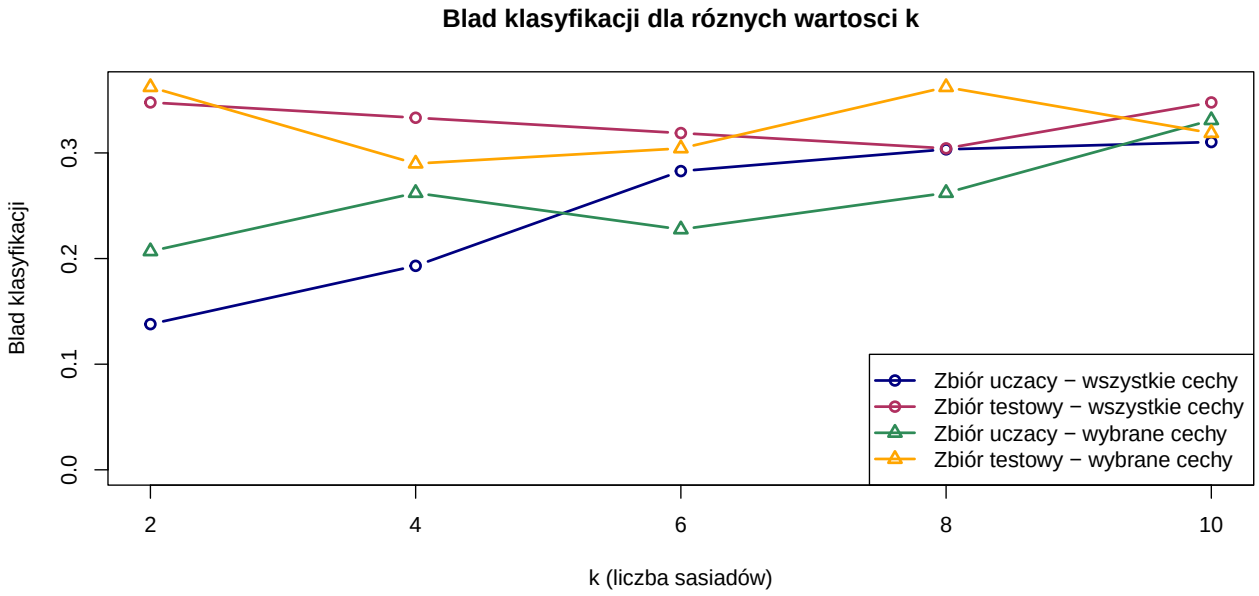
Metoda	Cross-validation	Bootstrap	632+
Drzewo klasyfikacyjne (wybrane cechy)	33.18%	35.31%	31.77%
Naive Bayes (wybrane cechy)	55.14%	59.66%	56.68%
K-najbliższych sąsiadów (wybrane cechy)	30.84%	37.31%	33.75%

Tabela 25, w porównaniu do Tabeli 23, ukazuje, że dla naiwnego klasyfikatora bayesowskiego, selekcja cech przyniosła znaczącą poprawę, obniżając błędy klasyfikacji (np. z 61.21% na 55.14% dla cross-validation). Mimo to, metoda ta nadal charakteryzowała się najwyższymi błędami spośród wszystkich analizowanych. W przypadku drzew klasyfikacyjnych, wpływ selekcji cech był bardziej złożony; dla niektórych metod oceny (np. cross-validation i bootstrap) błędy nieznacznie wzrosły po selekcji cech. Sugeruje to, że dla drzew kluczowa może być bardziej optymalizacja parametru cp niż sama redukcja liczby cech. Dla metody k-NN, selekcja

cech nie przyniosła spójnej poprawy: błąd dla cross-validation nieznacznie spadł (z 31.31% na 30.84%), ale dla bootstrapu wzrósł (z 35.81% na 37.31%). To wskazuje na bardziej subtelny i mniej przewidywalny wpływ selekcji cech na k-NN. Generalnie, metoda .632+ konsekwentnie dawała najbardziej optymistyczne (najniższe) estymacje błędu we wszystkich porównaniach.

Tabela 26: Porównanie błędów klasyfikacji dla różnych wartości k (k-NN)

k	Błąd uczący (wszystkie cechy)	Błąd testowy (wszystkie cechy)	Błąd uczący (wybrane cechy)	Błąd testowy (wybrane cechy)
2	13.79%	34.78%	20.69%	36.23%
4	19.31%	33.33%	26.21%	28.99%
6	28.28%	31.88%	22.76%	30.43%
8	30.34%	30.43%	26.21%	36.23%
10	31.03%	34.78%	33.10%	31.88%

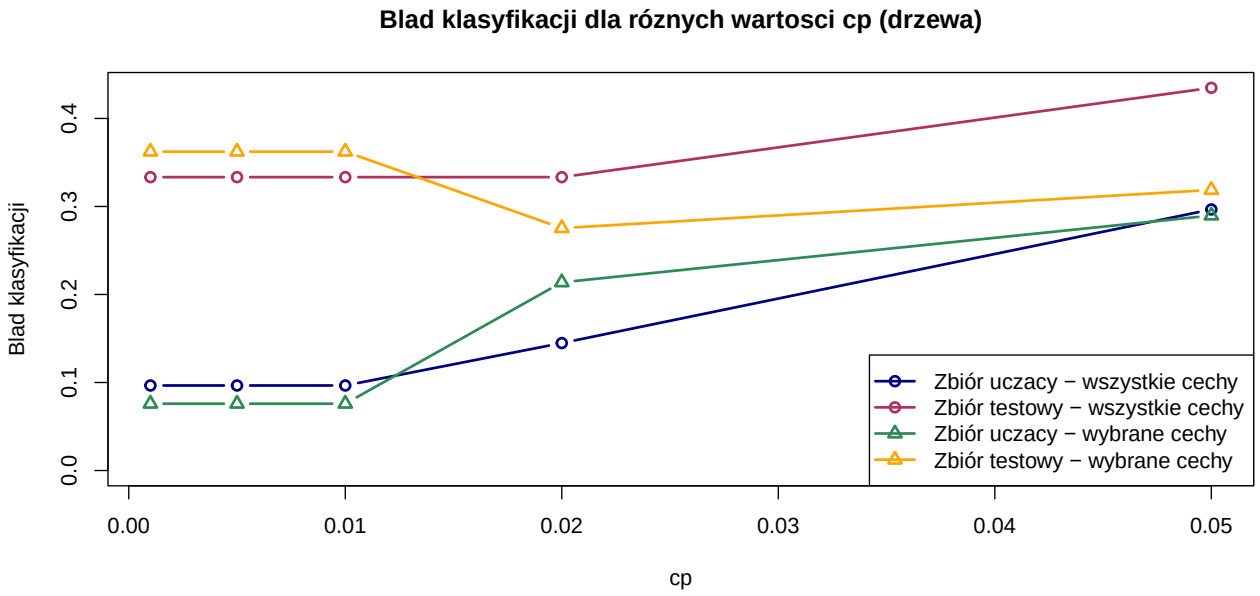


Rysunek 14: Wpływ liczby sąsiadów (k) na błąd klasyfikacji (k-NN)

Tabela 26 oraz Rysunek 14 ilustruje wpływ różnych wartości parametru k na błąd klasyfikacji metody k-NN, zarówno dla wszystkich cech, jak i dla wybranych cech. Dla wszystkich cech, błąd uczący wzrastał wraz ze wzrostem k (np. z 13.79% dla k=2 do 31.03% dla k=10), podczas gdy błąd testowy wykazywał zmienność, osiągając najniższą wartość 30.43% dla k=8. Dla wybranych cech, błąd uczący prawie w każdym przypadkuwzrastał wraz z k (np. z 20.69% dla k=2 do 33.10% dla k=10), a najniższy błąd testowy (28.99%) odnotowano dla k=4.

Tabela 27: Porównanie błędów klasyfikacji dla różnych wartości cp (drzewo klasyfikacyjne)

cp	Błąd uczący (wszystkie cechy)	Błąd testowy (wszystkie cechy)	Błąd uczący (wybrane cechy)	Błąd testowy (wybrane cechy)
0.001	9.66%	33.33%	7.59%	36.23%
0.005	9.66%	33.33%	7.59%	36.23%
0.010	9.66%	33.33%	7.59%	36.23%
0.020	14.48%	33.33%	21.38%	27.54%
0.050	29.66%	43.48%	28.97%	31.88%



Rysunek 15: Wpływ parametru cp na błąd klasyfikacji drzewa

Tabela 27 oraz Rysunek 15 pokazują, że optymalny współczynnik złożoności (cp) dla drzewa klasyfikacyjnego jest kluczowy. Małe wartości cp (np. 0.001) prowadzą do niskiego błędu na zbiorze uczącym, ale wysokiego na zbiorze testowym. Zwiększanie cp początkowo obniża błąd testowy, osiągając optimum (np. 27.54% dla wybranych cech przy cp=0.020), co wskazuje na lepszą generalizację. Jednak zbyt duże cp powoduje wzrost błędu testowego. Selekcja cech może dodatkowo poprawić wyniki na zbiorze testowym.

Tabela 28: Błędy klasyfikacji (k-NN, wszystkie cechy) dla różnych wartości k i metod oceny

k	Cross-validation	Bootstrap	632+
2	31.78%	32.54%	28.09%
4	32.24%	36.07%	31.96%

k	Cross-validation	Bootstrap	632+
6	35.98%	36.84%	32.03%
8	36.92%	37.96%	34.86%
10	40.19%	37.74%	36.13%

Tabela 29: Błędy klasyfikacji (k-NN, wybrane cechy) dla różnych wartości k i metod oceny

k	Cross-validation	Bootstrap	632+
2	41.12%	37.14%	31.00%
4	36.92%	37.91%	33.57%
6	31.78%	38.32%	34.35%
8	35.05%	38.99%	35.33%
10	34.11%	39.42%	37.04%

Tabela 30: Błędy klasyfikacji (drzewo, wszystkie cechy) dla różnych wartości cp i metod oceny

cp	Cross-validation	Bootstrap	632+
0.001	31.78%	34.56%	27.32%
0.005	30.37%	33.84%	29.24%
0.010	30.37%	34.26%	28.37%
0.020	28.97%	33.78%	29.76%
0.050	31.78%	35.75%	34.31%

Tabela 31: Błędy klasyfikacji (drzewo, wybrane cechy) dla różnych wartości cp i metod oceny

cp	Cross-validation	Bootstrap	632+
0.001	34.58%	38.10%	30.73%
0.005	33.64%	38.00%	31.55%
0.010	30.37%	35.53%	30.30%
0.020	33.18%	34.65%	30.74%
0.050	36.45%	35.84%	34.76%

Analiza Tabel 28, 29, 30 i 31 ujawnia kluczowe aspekty wydajności metod k-najbliższych sąsiadów (k-NN) i drzew klasyfikacyjnych. W przypadku k-NN (Tabele 28 i 29), zwiększanie liczby sąsiadów (k) na ogół podnosi błędy klasyfikacji we wszystkich metodach walidacji, co sugeruje, że mniejsze k jest korzystniejsze dla tego zbioru danych. Selekcja cech dla k-NN

ma zmienny wpływ, czasem nieznacznie poprawiając wyniki dla większych k, a innym razem pogarszając dla mniejszych.

Dla drzew klasyfikacyjnych (Tabele 30 i 31), optymalny współczynnik złożoności (cp) jest kluczowy: zarówno zbyt niskie, jak i zbyt wysokie wartości cp prowadzą do wyższych błędów (przeuczenie lub niedouczenie). Wartości cp w przedziale 0.010-0.020 często minimalizują błędy. Selekcja cech jest tu bardziej konsekwentnie korzystna, często prowadząc do niższych błędów testowych i lepszej generalizacji. Wszystkie zastosowane metody walidacji (cross-validation, bootstrap, .632+) wykazują spójne tendencje błędów, przy czym metoda .632+ konsekwentnie daje najbardziej optymistyczne estymacje.

2.6 Wnioski końcowe

Na podstawie przeprowadzonej analizy danych Glass i porównania trzech metod klasyfikacji – metody k-najbliższych sąsiadów (k-NN), drzew klasyfikacyjnych oraz naiwnego klasyfikatora bayesowskiego – można sformułować następujące wnioski końcowe.

W odniesieniu do najlepszych wyników, dla metody k-NN najniższy błąd testowy (28.99%) odnotowano dla parametru k=4 przy użyciu wyselekcjonowanego podzbioru cech (`Mg`, `Al`, `Na`, `K`, `Ba`). Należy jednak zaznaczyć, że błąd klasyfikacji dla k=8 przy wszystkich cechach (30.43%) był również konkurencyjny. W przypadku drzew klasyfikacyjnych, optymalny współczynnik złożoności (cp), który minimalizuje błąd testowy, znajduje się w przedziale 0.010-0.020. Dla wybranych cech, najniższy błąd testowy (27.54%) uzyskano przy cp=0.020. Selekcja cech (Mg, Al, Na, K, Ba) okazała się korzystna, często prowadząc do niższych błędów testowych i lepszej generalizacji. Naiwny klasyfikator bayesowski zanotował znaczną poprawę zarówno błędu uczącego (z 48.28% do 44.14%), jak i testowego (z 59.42% do 49.28%) przy użyciu wybranych cech. Mimo to, jego wyniki były najsłabsze spośród wszystkich analizowanych metod.

Odnośnie do skuteczności poszczególnych metod, metoda k-NN oraz drzewa klasyfikacyjne (szczególnie po optymalizacji parametru cp i selekcji cech) wykazały się znacznie lepszą skutecznością niż naiwny klasyfikator bayesowski. W przypadku zaawansowanych schematów oceny dokładności, takich jak 10-krotna cross-validation, bootstrap i metoda .632+, drzewo klasyfikacyjne okazało się nawet nieco lepsze (błędy 29.44-34.72%) niż k-NN (błędy 31.21-35.81%), co było zaskoczeniem w porównaniu do początkowego pojedynczego podziału. Naiwny klasyfikator bayesowski konsekwentnie osiągał najgorsze rezultaty, z błędami przekraczającymi 55.81% we wszystkich zaawansowanych metodach, co potwierdziło jego niską skuteczność dla analizowanego zbioru danych.

Podsumowując, wybór schematu oceny dokładności miał istotny wpływ na wnioski dotyczące skuteczności metod, zwłaszcza dla drzewa klasyfikacyjnego. Początkowy pojedynczy podział danych dla drzewa klasyfikacyjnego mógł być niemiarodajny, gdyż w zaawansowanych schematach (cross-validation, bootstrap, .632+) drzewo wykazało się lepszą dokładnością niż w pojedynczym podziale. Wszystkie zastosowane metody walidacji (cross-validation, bootstrap, .632+) wykazują spójne tendencje błędów, przy czym metoda .632+ konsekwentnie daje najbardziej optymistyczne estymacje. Podkreśla to znaczenie stosowania zaawansowanych metod w celu uzyskania bardziej stabilnych i wiarygodnych estymacji błędu. Spośród użytych

metod, najlepszą stabilność i najniższe estymowane błędy dla większości modeli zapewniła metoda .632+, co czyni ją szczególnie rekomendowaną do oceny modeli w sytuacjach, gdzie dostępność danych jest ograniczona lub podział losowy może prowadzić do dużych odchyleń w wynikach.