Raport Lista 2

Eksploracja danych

Dominik Kowalczyk i Matylda Mordal

2025-05-27

Spis treści

1 Zadanie 2 - Porównywanie metod klasyfikacji

1.1 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 1: 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.

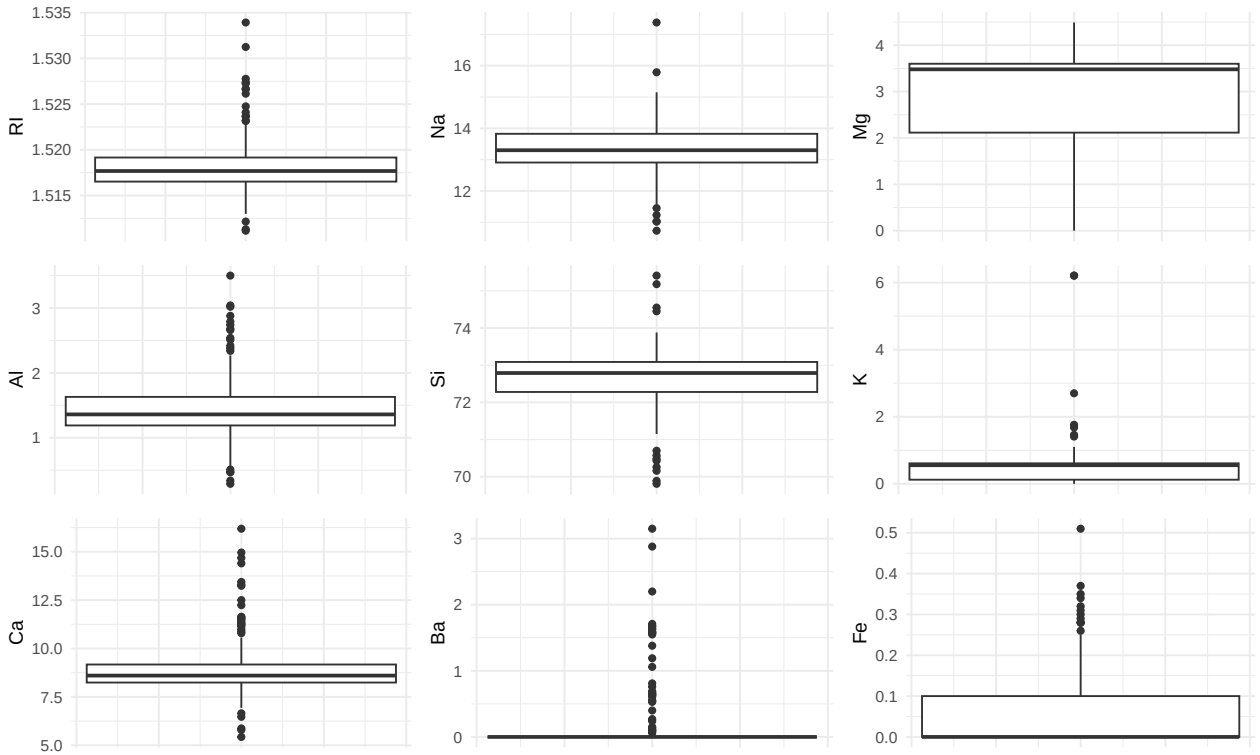
Zbiór danych `glass_data` zawiera 10 zmiennych, z czego ostatnia z nich, `Type` przechowuje informacje o przynależności obiektu do konkretnej klasy (tzw. etykieta klas). Jest ona typu: factor. Dodatkowo, pozostałe zmienne zawierają informację dotyczące występowania danego piewiastka chemicznego w szkle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

Tabela 2: Liczba Obserwacji dla Każdego Typu Szkła

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

```
#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.

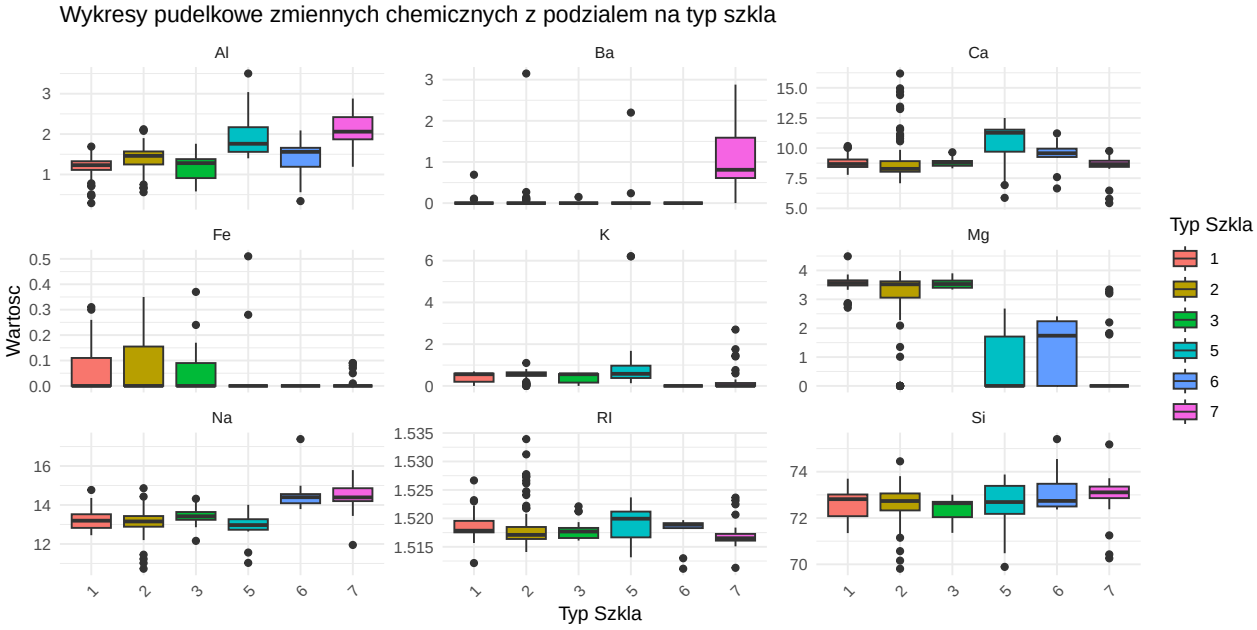


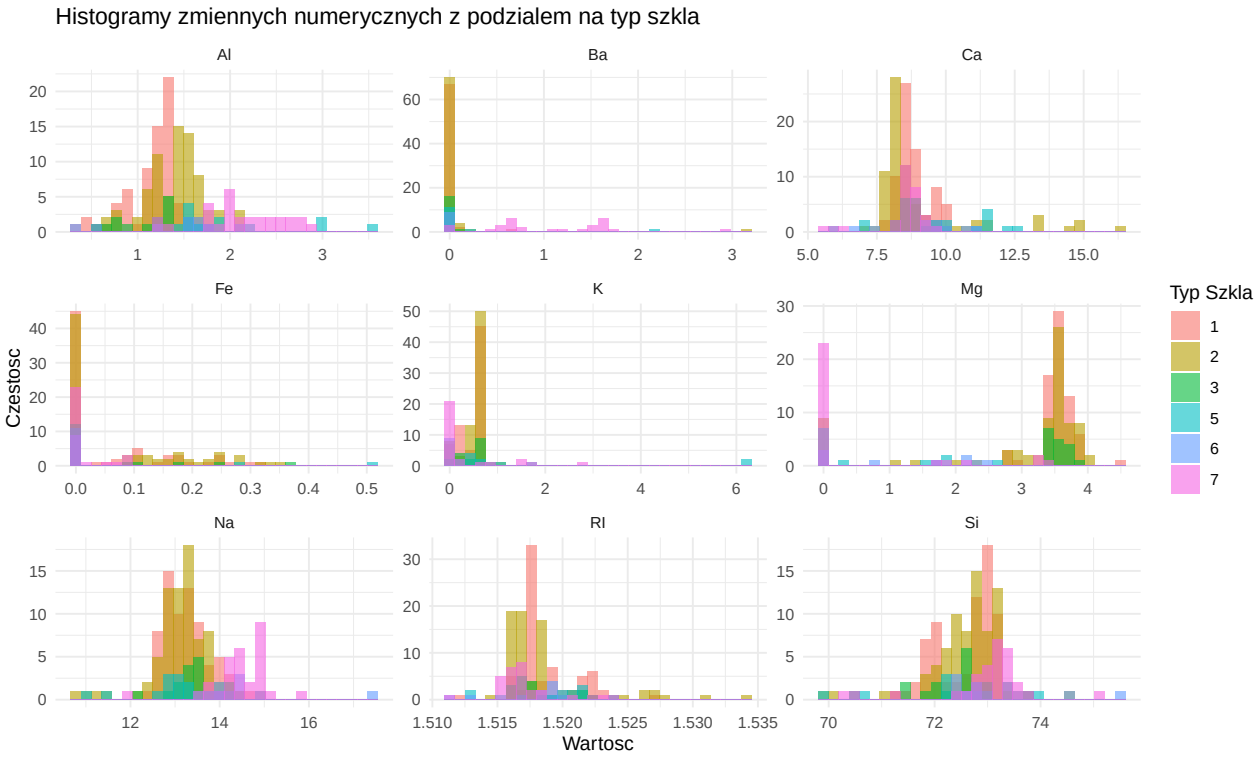
W analizowanym zbiorze nie mamy doczynienia z "nieścisłościami" rozkładów w sensie błędów w danych, ale raczej charakterystycznymi cechami chemicznymi różnych typów szkła. Dziwne rozkłady (silnie skośne, z licznymi odstającymi) dla takich pierwiastków jak Mg, K i Ba są prawdopodobnie wynikiem specyficznych 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).

1.2 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.





Na podstawie analizy histogramów oraz wcześniejszych wykresów pudełkowych 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.

Podsumowując, zmienne takie jak *Magnez*, *Bar*, *Potas* i *Glin* wydają się najbardziej obiecujące w rozróżnianiu typów szkła na podstawie ich rozkładów i centralnych tendencji. Zmienne *Wapń* i *Sód* mogą dostarczać dodatkowych informacji, podczas gdy *Żelazo* i *Krzem* prawdopodobnie mają mniejsze znaczenie w ogólnej dyskryminacji. *Współczynnik załamania światła* wykazuje

subtelne różnice, które mogą być istotne w bardziej złożonych modelach klasyfikacyjnych. Ta wstępna ocena sugeruje, na których cechach warto skupić się podczas budowania i analizowania modeli klasyfikacyjnych dla zbioru danych

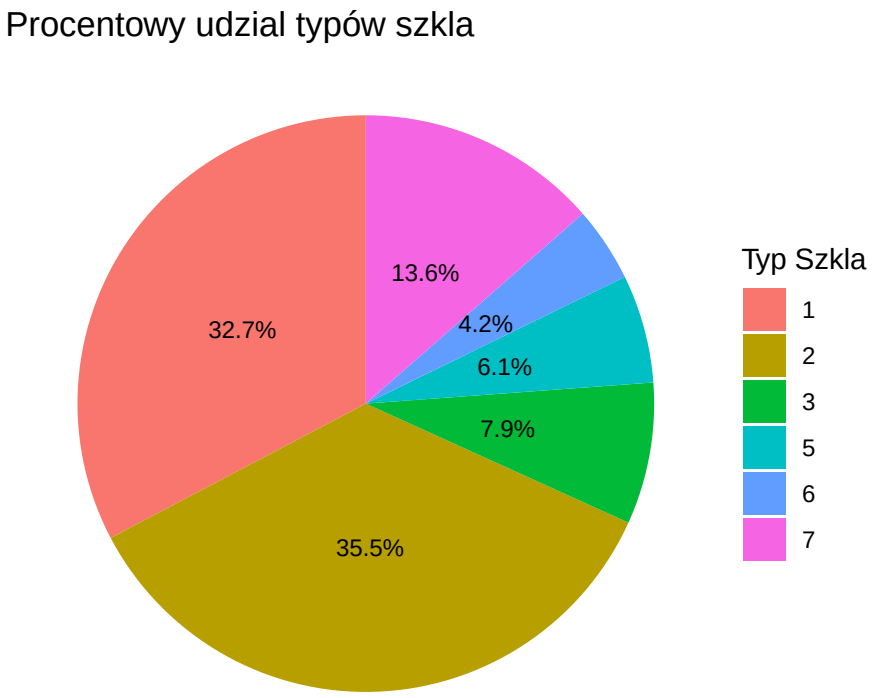


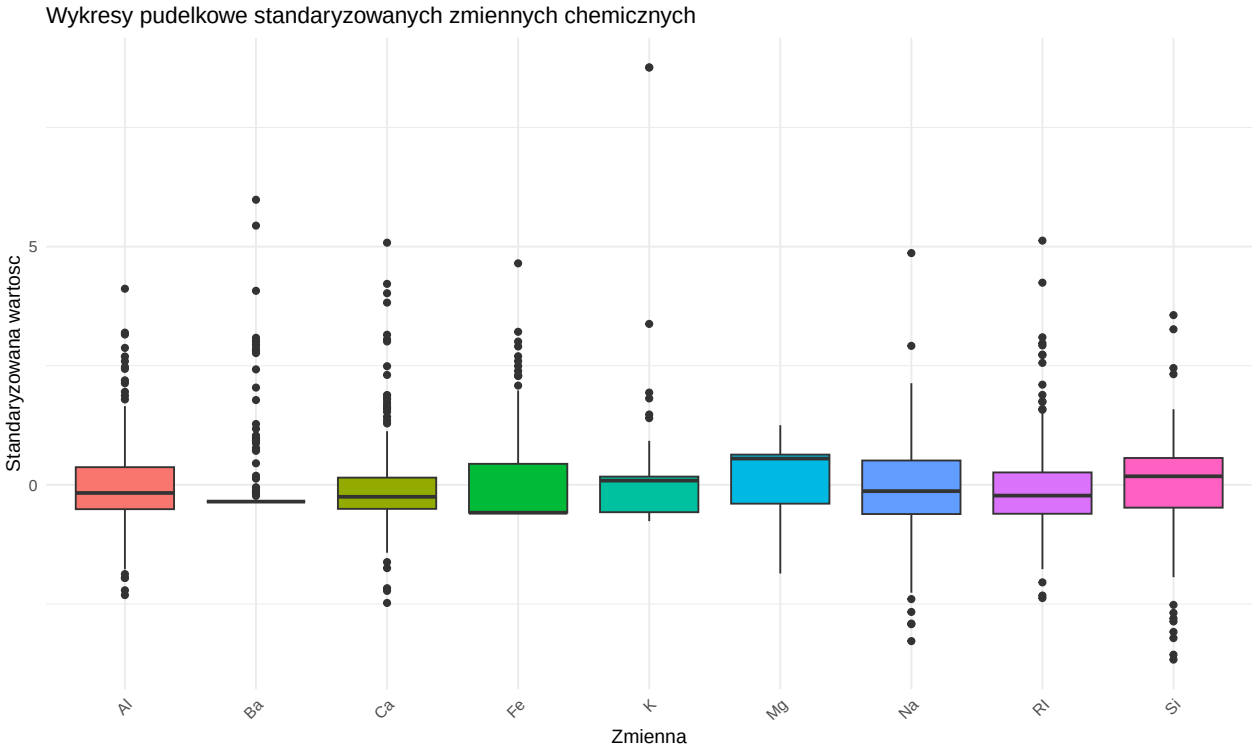
Tabela 3: Błąd Klasyfikacji przy Przypisaniu Wszystkich do Najczęstszej Klasy

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

Tabela 4: 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



1.3 Ocena dokładności klasyfikacji i porównanie metod

1.3.1 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.

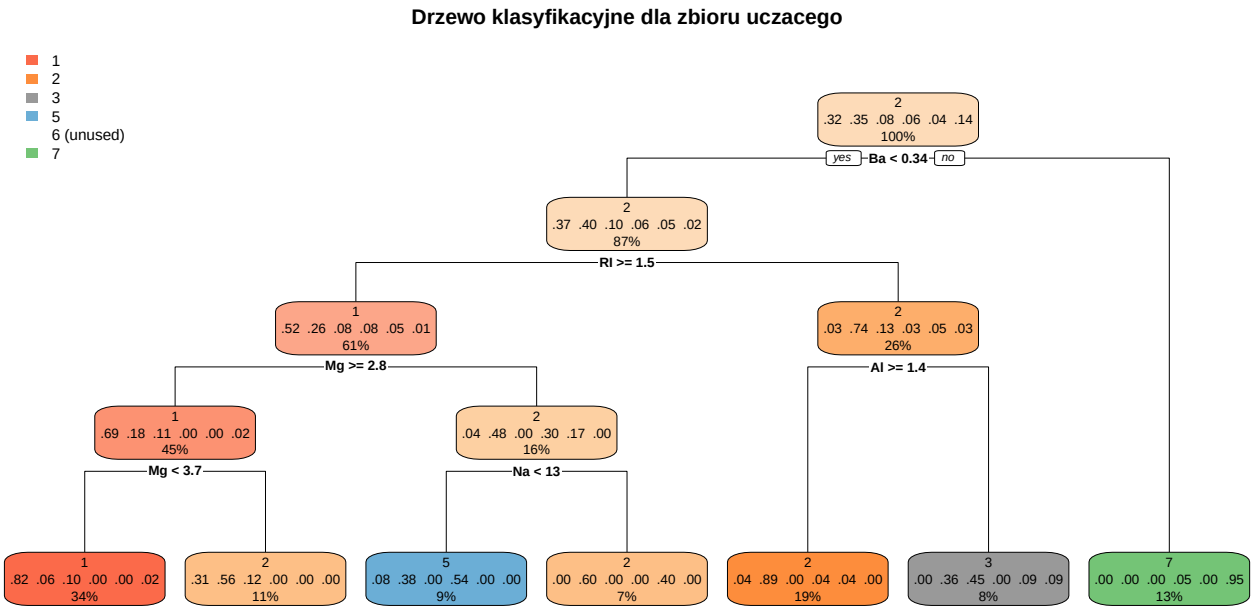
Tabela 5: Macierz pomyłek dla zbioru uczącego (K-najbliższych sąsiadów)

	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 6: 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 6) 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 5) 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 11).



Rysunek 1: Drzewo klasyfikacyjne dla zbioru danych Glass

Tabela 7: Macierz pomyłek dla zbioru uczącego (Drzewo klasyfikacyjne)

	1	2	3	5	6	7
1	40	6	0	1	0	0

	1	2	3	5	6	7
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 8: 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 1. 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 8) potwierdziła te obserwacje, szczególnie widoczny był brak trafnych klasyfikacji dla klasy 6, której wszystkie przypadki zostały błędnie przypisane. Na zbiorze uczącym (Tabela 7) 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 11). Ta różnica sugeruje, że model słabo radzi sobie z dostosowaniem do nowych danych, co sprawia, że jest mniej skuteczny niż metoda k-NN.

Tabela 9: 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

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

Tabela 10: 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 11: 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%
Naiwny Bayes	48.28%	59.42%

Najsłabsze wyniki w naszej analizie uzyskał naiwny klasyfikator bayesowski. Macierz pomyłek dla zbioru uczącego (Tabela 9) oraz zbioru testowego (Tabela 10) jasno pokazała liczne pomyłki. Model ten charakteryzował się najwyższymi błędami klasyfikacji: 48.28% na zbiorze uczącym i 59.42% na zbiorze testowym (Tabela 11).

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 12: 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.

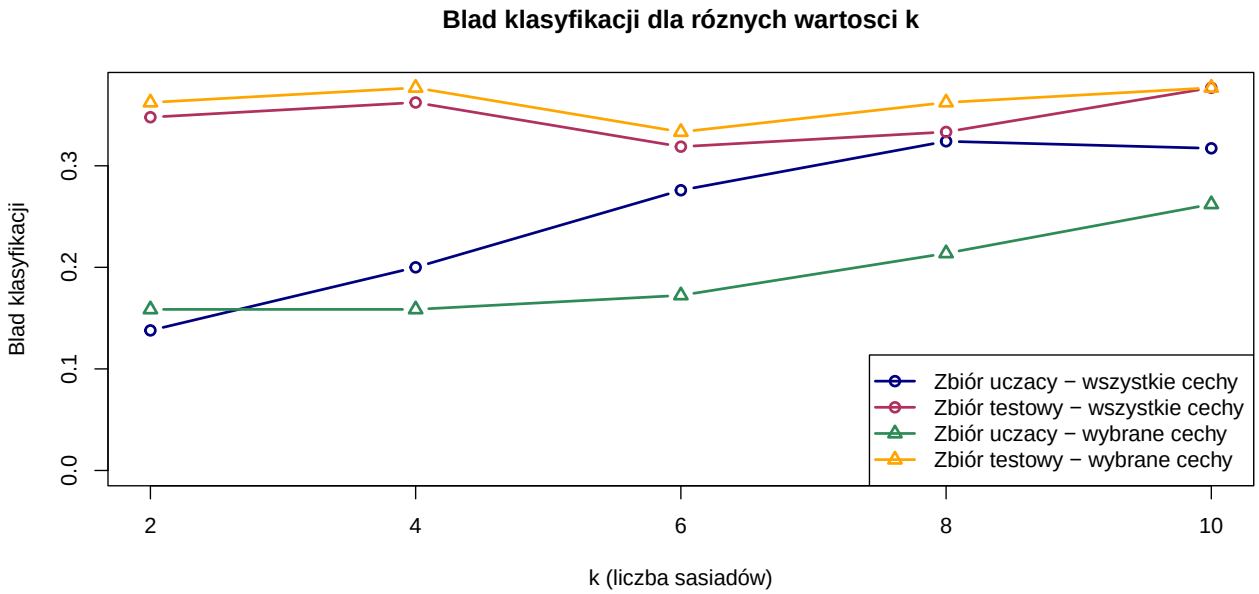
1.3.2 Różne parametry i różne podzbiory cech

Tabela 13: 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)	14.48%	36.23%
Drzewo klasyfikacyjne (wszystkie cechy)	24.83%	44.93%
Drzewo klasyfikacyjne (wybrane cechy)	24.83%	44.93%
Naive Bayes (wszystkie cechy)	48.28%	59.42%
Naive Bayes (wybrane cechy)	50.34%	52.17%

Tabela 14: 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%	15.86%	36.23%
4	20.00%	36.23%	15.86%	37.68%
6	27.59%	31.88%	17.24%	33.33%
8	32.41%	33.33%	21.38%	36.23%
10	31.72%	37.68%	26.21%	37.68%



Rysunek 2: Wpływ liczby sąsiadów (k) na błąd klasyfikacji