Sprawozdanie z listy nr 3

Eksploracja Danych

Dawid Skowroński 282241, Mateusz Cieślak 272633

2025-06-04

Spis treści

Spis wykresów

Spis tabel

1 Zadanie 1

1.1 Wstępna analiza danych

Tabela 1: Struktura zbioru danych iris

Zmienna	Opis	Typ	Przykładowe wartości
Sepal.Length	Długość działki kielicha	numeric	5.1, 4.9, 4.7, 4.6
Sepal.Width	Szerokośc działki kielicha	numeric	3.5, 3, 3.2, 3.1
Petal.Length	Długość płatka	numeric	1.4, 1.4, 1.3, 1.5
Petal.Width	Szerokość płatka	numeric	0.2, 0.2, 0.2, 0.2
Species	Gatunek irysa	factor	setosa, setosa, setosa, setosa

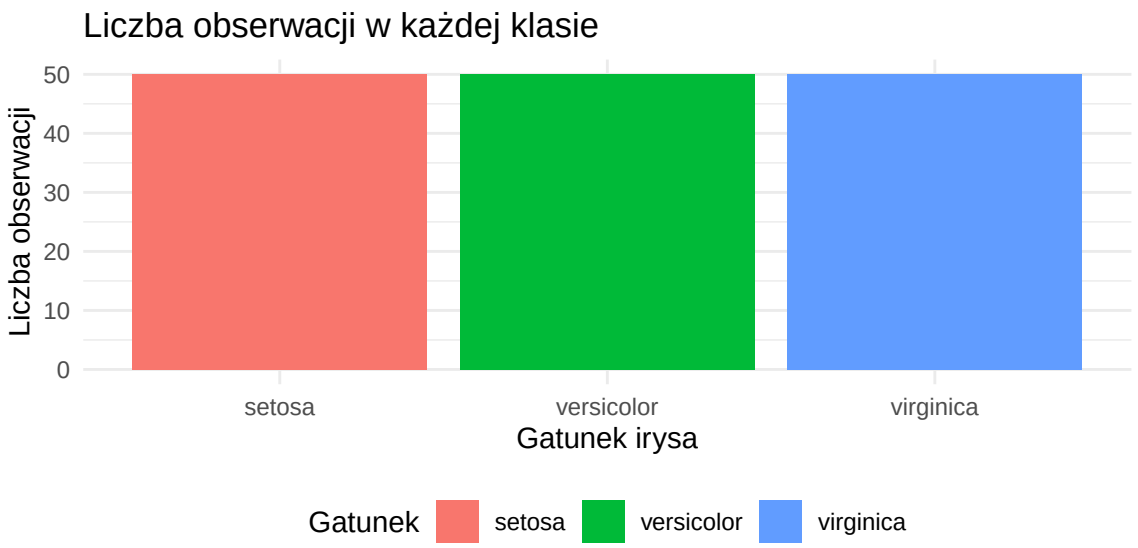
Zbiór danych `iris` zawiera pomiary o dla trzech gatunków irysów (`setosa`, `virginica`, `versicolor`). Opis poszczególnych zmiennych wraz z przykładkowymi wartościami przedstawiony jest w tabeli 1.

Celem analizy jest:

- zbudowanie modelu predykcyjnego na zbiorze uczącym,
- porównanie prognozowanych etykietek klas z rzeczywistymi etykietkami, zarówno dla zbioru uczącego, jak i testowego,
- stwierdzenie, czy wyznaczony model nadaje się do klasyfikacji zbioru `iris`.

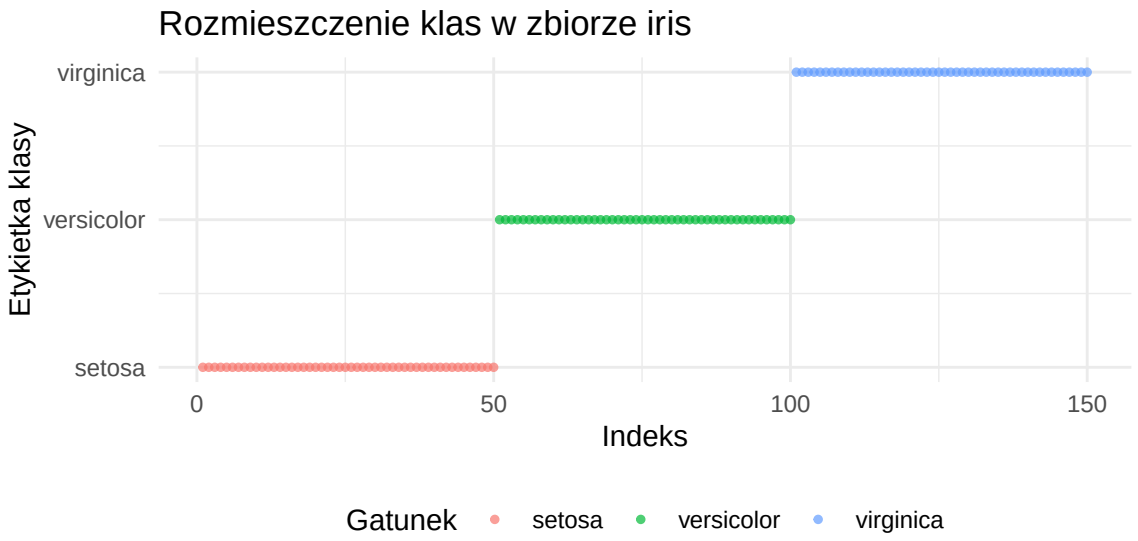
Tabela 2: Wartości brakujące w zbiorze iris

zmienna	liczba wartości brakujących
Sepal.Length	0
Sepal.Width	0
Petal.Length	0
Petal.Width	0
Species	0



Wykres 1: Wykres słupkowy, przedstawiający liczbę obserwacji dla poszczególnego gatunku

Zbiór danych `iris` nie zawiera wartości brakujących (tabela 2). Dodatkowo, liczba obserwacji dla każdego gatunku jest równa, co przedstawia wykres 1.



Wykres 2: Etykietki klas dla zbioru iris

Rozkład gatunków w zbiorze `iris` (wykres 2) jest uporządkowany.

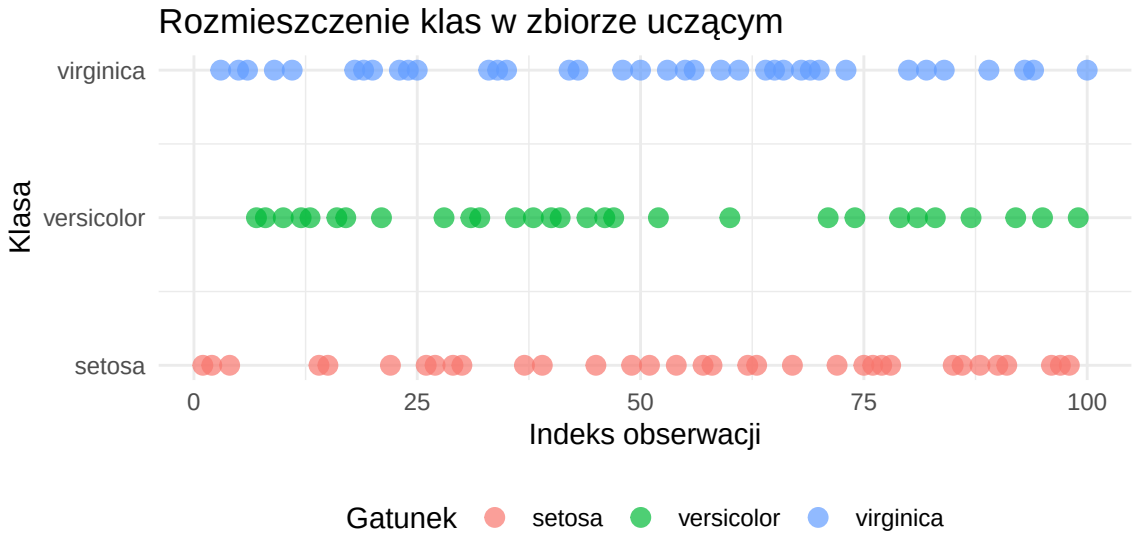
1.2 Podział danych na zbiór testowy i uczący

Dzielimy zbiór `iris` na `zbiór uczący` ($\frac{2}{3}$ zbioru `iris`) i `zbiór testowy` ($\frac{1}{3}$ zbioru `iris`).

Tabela 3: Liczba obserwacji poszczególnych gatunków w zbiorze uczącym i testowym

Gatunek	Zbiór uczący	Zbiór testowy
setosa	34.00	16.00
versicolor	29.00	21.00
virginica	37.00	13.00
Suma	100.00	50.00

Można zauważyć, że rozkład obserwacji w zależności od gatunku (tabela 3) jest dość równo rozłożony w zbiorze uczącym. W zbiorze testowym natomiast, występuje najwięcej rekordów dla gatunku `versicolor`.



Wykres 3: Rozkład obserwacji dla zbioru uczącego zbioru iris (w zależności od gatunku)

Możemy zauważyć, że po podziale zbiory nie są uporządkowane (Wizualizacja dla zbioru uczącego na wykresie 3). Może to utrudnić interpretację kolejnych wykresów. Sortujemy zatem oba zbiory względem gatunku.

1.3 Konstrukcja klasyfikatora i wyznaczenie prognoz

Prognozujemy zmienną jakościową, etykietkę klasy (do każdego ustalonego obiektu chcemy przypisać mu przynależność do danej klasy).

Przyjmujemy model liniowy regresji w postaci:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_n X_n + \epsilon,$$

gdzie Y - zmienna zależna (objaśniana), $X_i, i = 1, ..., n$ - zmienne niezależne (objaśniające), ϵ - błędy losowe o jednakowej wariancji σ^2, $\beta_j, j = 0, ..., n$ - pewne współczynniki.

Prognozę zmiennej Y wyznaczamy ze wzoru:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + ... + \hat{\beta}_n X_n,$$

gdzie $\hat{\beta}_j, j = 0, ..., n$ - estymatory nieznanych współczynników.

Tworzymy model w nastepujący sposób:

- Dodajemy kolumnę wolnych wyrazów do macierzy modelu zbioru uczącego (macierz $\mathbb{X}$).
- Tworzymy macierz wskaźnikową $\mathbb{Y}_{n \times g}$, o liczbie wierszy równej liczbie obserwacji (wartość **n**) oraz liczbie kolumn równej liczbie klas (wartość **g**), zawierającą przynależność do poszczególnych klas zakodowane za pomocą zmiennych binarnych.
- Metodą najmniejszych kwadratów wyznaczamy estymatory współczynników $\hat{\beta}_j$, korzystając ze wzoru $\hat{\mathbb{B}}_{(k+1) \times g} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{Y}$, gdzie $\mathbb{X}$ jest macierzą modelu:

$$\mathbb{X}_{n \times (k+1)} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}.$$

- Wyznaczamy prawdopodobieństwo **a posteriori** (prognozowane prawdopodobieństwo przynależności do klasy) za pomocą wzoru $\hat{\mathbb{Y}} = \mathbb{X}\hat{\mathbb{B}}$

- Do każdej obserwacji przypisujemy tę klasę, dla której mamy największe prawdopodobieństwo.

```
head(Y_prawd_uczacy)
```

```
##         [,1]      [,2]        [,3]
## 1  0.9076361 0.2911157 -0.19875178
## 2  0.9532652 0.1765026 -0.12976779
## 4  0.8981582 0.2173647 -0.11552291
## 14 0.8515397 0.3242292 -0.17576894
## 15 0.9225391 0.1149543 -0.03749339
## 22 0.9715014 0.0488415 -0.02034287
```

Możemy zauważyć, że niektóre wartości prawdopodobieństwa wykraczają poza przedział $[0, 1]$.

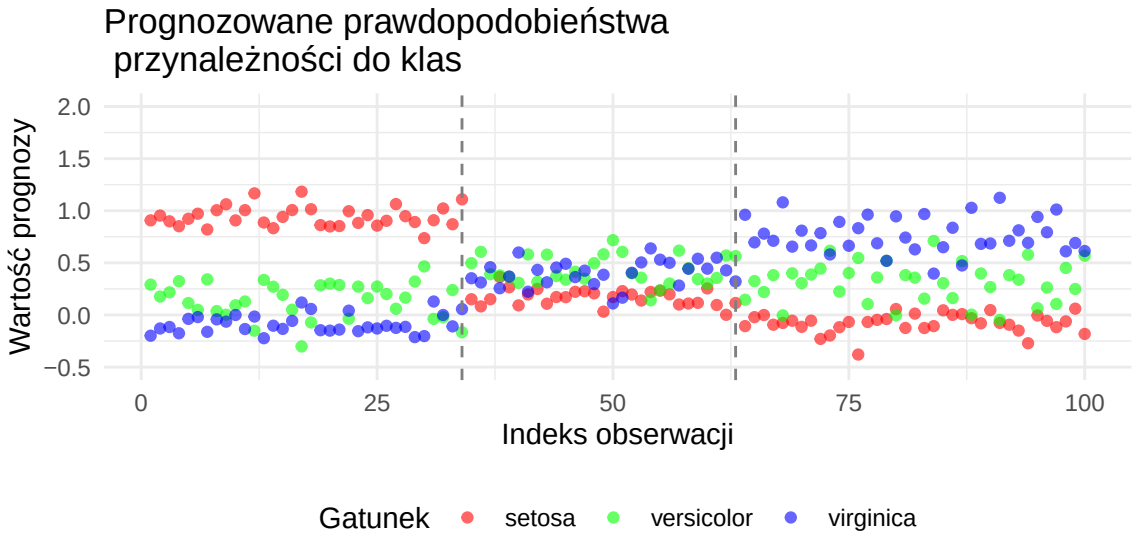
Kontrolne sprawdzenie czy prawdopodobieństwa sumują się do 1:

```
rowSums(Y_prawd_uczacy)
```

```
##   1   2   4  14  15  22  26  27  29  30  37  39  45  49  51  54  57  58  62  63
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  67  72  75  76  77  78  85  86  88  90  91  96  97  98   7   8  10  12  13  16
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  17  21  28  31  32  36  38  40  41  44  46  47  52  60  71  74  79  81  83  87
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  92  95  99   3   5   6   9  11  18  19  20  23  24  25  33  34  35  42  43  48
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
##  50  53  55  56  59  61  64  65  66  68  69  70  73  80  82  84  89  93  94 100
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
```

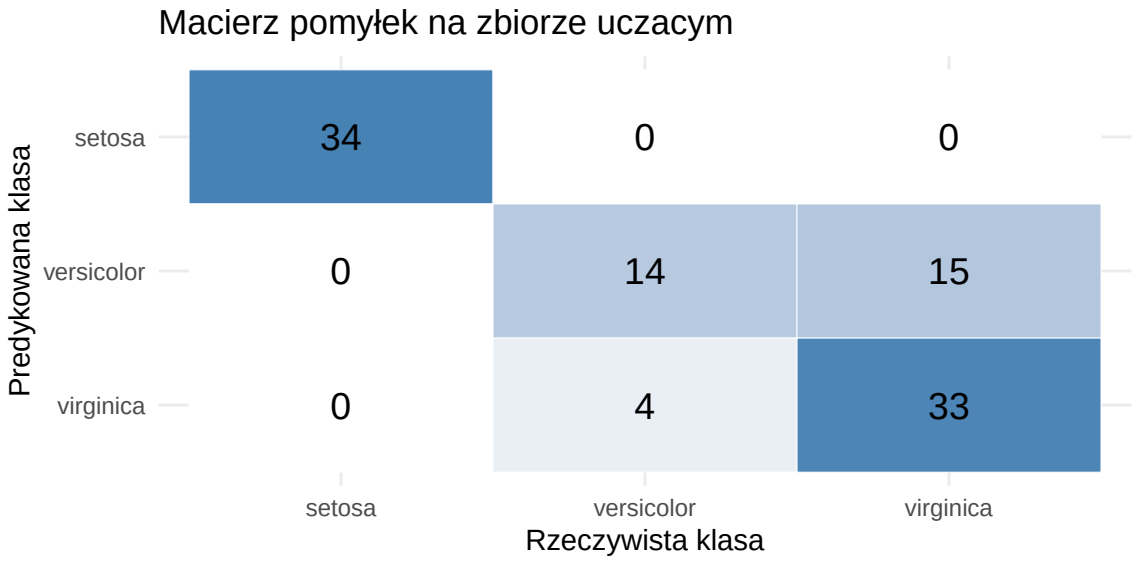
Prawdopodobieństwa dla każdej obserwacji sumują się do 1.

1.4 Ocena dokładności modelu



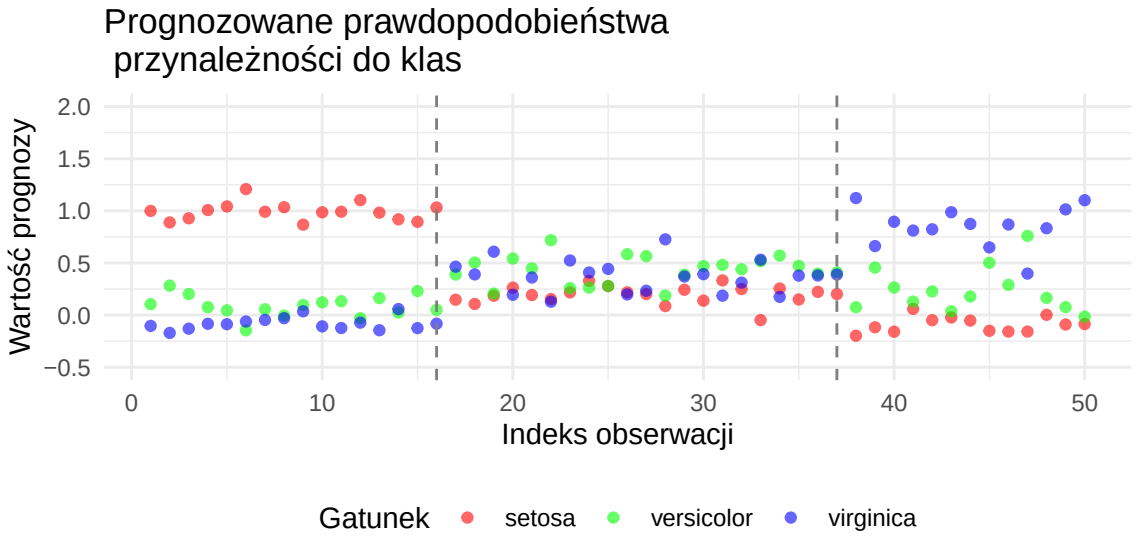
Wykres 4: Prawdopodobieństwa przynależności do poszczególnych gatunków w zbiorze uczącym wyznaczone za pomocą modelu liniowej regresji

Wykres 4 przedstawia rozkład prawdopodobieństw obserwacji dla poszczególnych klas w zbiorze uczącym. Możemy zauważyć, że w pierwszej klasie (patrząc od lewej strony wykresu) największą wartość przyjmują obiekty o etykiecie klasy `setosa`, dodatkowo są bardzo dobrze odseparowane od pozostałych obserwacji. Trochę gorsza separacja widoczna jest w trzeciej klasie. W znacznej większości dominują obserwacje z etykietką klasy `virginica`, aczkolwiek niektóre obiekty o etykiecie `versicolor` przyjmują równie wysokie prawdopodobieństwo. W drugim przedziale nie można wyróżnić żadnej z etykiet. Druga klasa (`versicolor`) jest przysłaniana przez klasę 3 (`virginica`). Wiele punktów ma prawdopodobieństwa bliskie 0.5 dla obu klas. Występuje zatem problem częściowego maskowania klas, co znacznie wpływa na dokładność klasyfikacji.



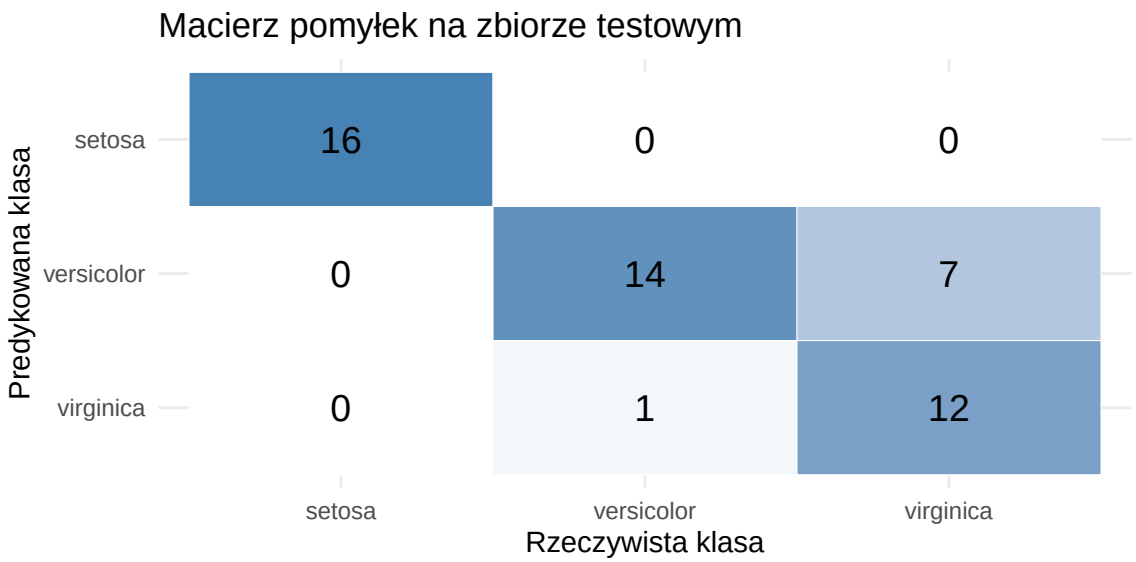
Wykres 5: Wykres liczby rzeczywistych etykietek klas względem prognozowanych w zbiorze uczącym

Wykres 5 ukazuje, że najwięcej niepoprawnych dopasowań etykietek wystąpiło dla gatunku `versicolor` (błędne przypisanie do klasy `virginica`), co zgadza się z obserwacjami z wykresu 4. Dla gatunku `setosa` wszystkie etykietki zostały poprawnie dopasowane. Dokładność klasyfikacji dla zbioru uczącego jest na poziomie **81** %.



Wykres 6: Prawdopodobieństwa przynależności do poszczególnych gatunków w zbiorze testowym wyznaczone za pomocą modelu liniowej regresji

Wykres 6 przedstawia rozkład prognozowanych prawdopodobieństw dla zbioru testowego. Możemy z niego wyciągnąć podobne wnioski jak dla wykresu 4. Ponownie możemy zauważyć bardzo dobre odseparowanie dla gatunku setosa. Również dość dobrze wyróżnia się klasa `virginica` (pierwszy przedział od prawej), jedynie pojedyncze obserwacje o etykiecie `versicolor` przyjmują wyższe wartości od tych o etykiecie `virginica`. Znowu widoczny jest brak wyraźnej dominacji jednej etykiety w środkowej części wykresu (problem częściowego maskowania klas). Może to wynikać z tego, że niektóre cechy mają podobny rozkład dla tych właśnie gatunków i częściowo na siebie nachodzą.



Wykres 7: Wykres liczby rzeczywistych etykietek klas względem prognozowanych w zbiorze testowym

Wykres 7 ponownie pokazuje, że najwięcej niepoprawnych dopasowań etykietek wystąpiło dla gatunku `versicolor`. Zgadza się z obserwacjami z wykresu 6. Dla gatunku `setosa` wszystkie etykietki zostały poprawnie dopasowane. Dokładność klasyfikacji dla zbioru uczącego jest na poziomie **84** %.

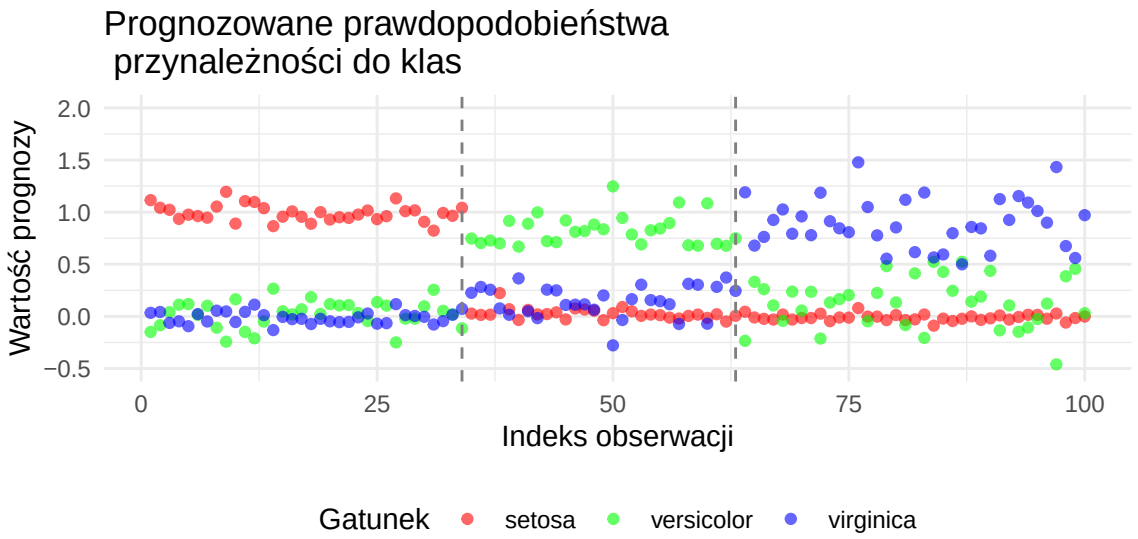
1.5 Konstrukcja modelu liniowego dla rozszerzonej przestrzeni cech

Ponownie tworzymy model regresji liniowej dla zbioru `iris` tym razem rozszerzonego o składniki wielomianowe i iloczyny zmiennych stopnia `K-1`, gdzie `K` to liczba klas, w celu teoretycznej poprawy dokładności klasyfikacji i wyeliminowania lub zmniejszenia zjawiska maskowania klas. W tym przypadku uwzględniamy wielomiany stopnia 2. Ponawiamy też wybór podzbiorów (zbiór uczący i testowych w takich samych proporcjach jak poprzednio).

```
iris_rozszerz <- iris
# Zmieniamy nazwy na krótsze
names(iris_rozszerz) <- c("SL", "SW", "PL", "PW", "S")
iris_rozszerz <- transform(iris_rozszerz, SL_SW = SL *
    SW, PL_PW = PL * PW, PL_SL = PL * SL, PL_SW = PL *
    SW, PW_SW = PW * SW, PW_SL = PW * SL, PL_2 = PL *
    PL, PW_2 = PW * PW, SL_2 = SL * SL, SW_2 = SW *
    SW)
```

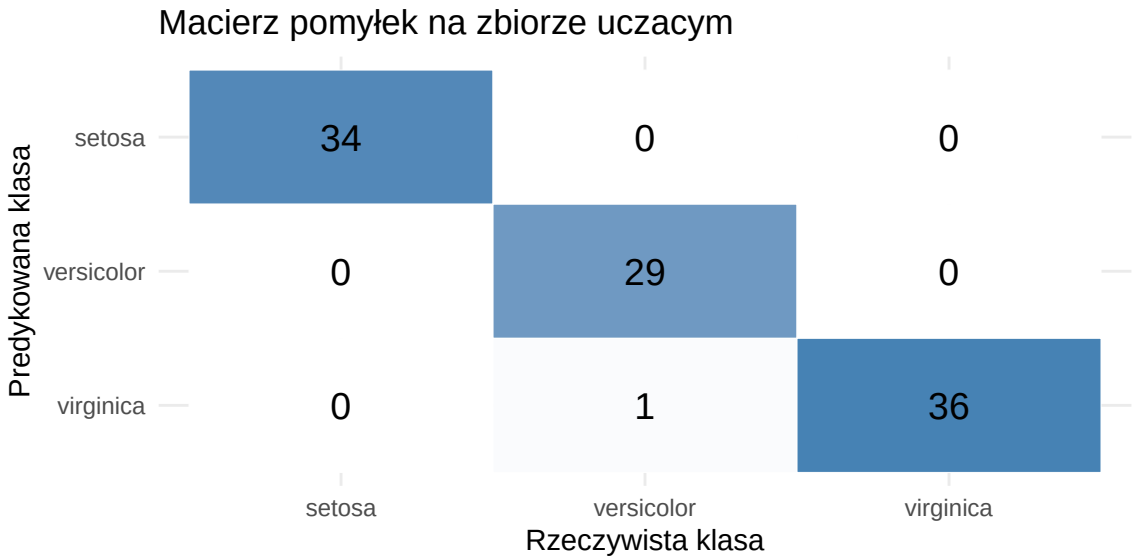
Konstrukcja modelu przebiega analogicznie. Jedną z różnic jest rozmiar macierzy $\mathbb{X}$ (zawiera teraz 15 kolumn (czyli wszystkie kombinacje zmiennych zbioru iris, które tworzą wielomiany stopnia co najwyżej 2) zamiast 5).

1.5.1 Ocena jakości modelu



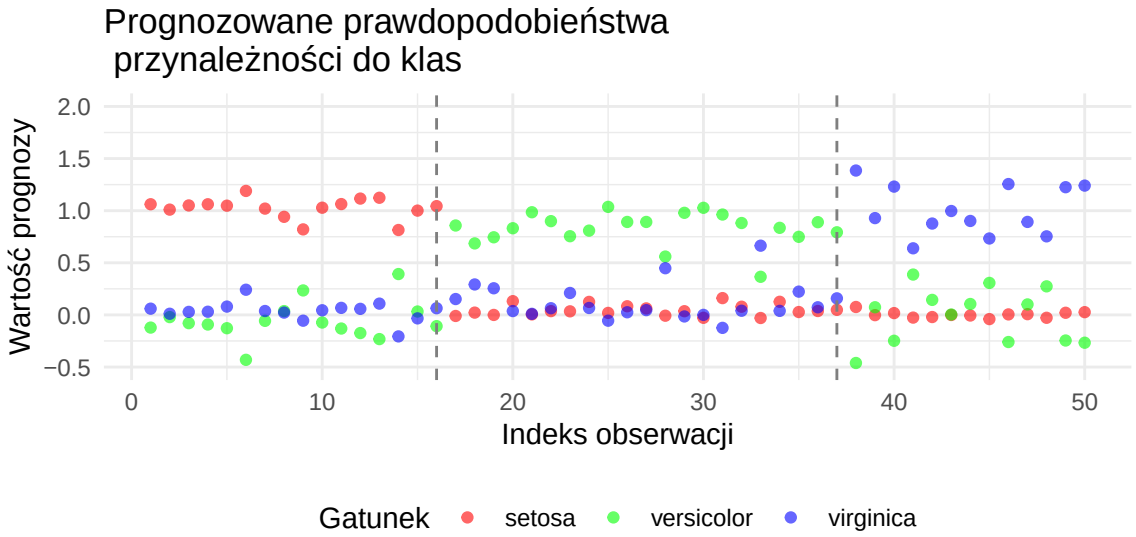
Wykres 8: Prawdopodobieństwa przynależności do poszczególnych gatunków w zbiorze uczącym wyznaczone za pomocą modelu liniowej regresji dla rozszerzonej przestrzeni cech

Po wyznaczeniu prognozowanych wartości prawdopodobieństw dla rozszerzonej przestrzeni cech, możemy zauważyć znaczną poprawę w dyskryminacji obserwacji (wykres 8). Obiekty o etykiecie `setosa` i `versicolor` są idealnie odseparowane od pozostałych obserwacji (pierwszy i drugi przedział od lewej). Natomiast dla etykietki `virginica` jedynie kilka obiektów ma prognozowaną wartość prawdopodobieństwa mniejszą niż te o etykietce `versicolor`. Możemy stąd wywnioskować, że maskowanie klas już nie występuje.



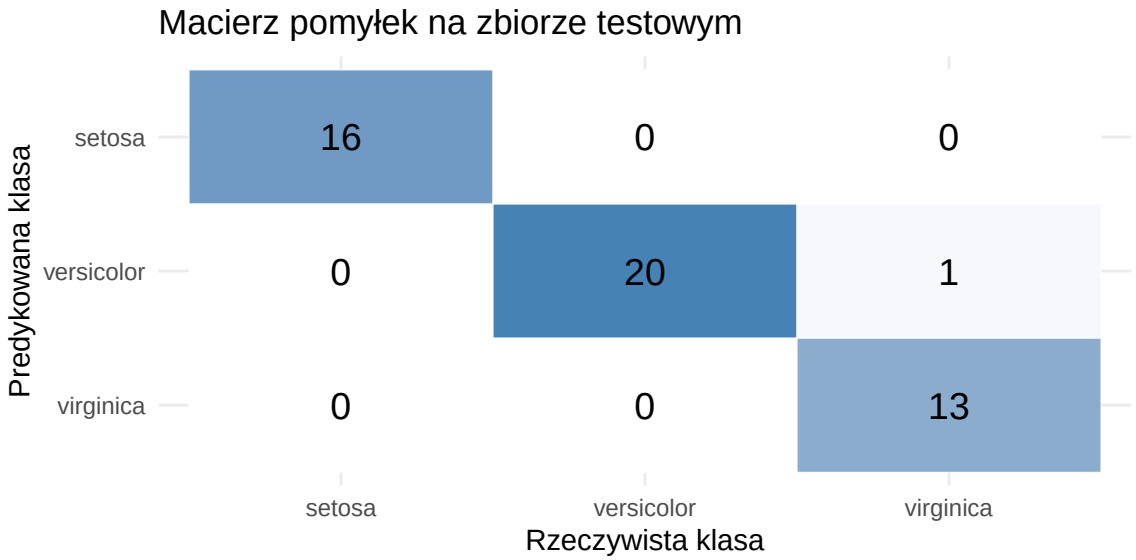
Wykres 9: Wykres liczby rzeczywistych etykietek klas względem prognozowanych w zbiorze uczącym w przestrzeni rozszerzonej

Wykres 9 potwierdza dobre przyporządkowanie klas, widoczna jest znacząca poprawa względem zbioru uczącego dla podstawowego zbioru `iris` (wykres 5). Dokładność klasyfikacji jest na poziomie **99** %. Znaczna poprawa dokładności klasyfikacji sugeruje, że problem maskowania klas został wyeliminowany.



Wykres 10: Prawdopodobieństwa przynależności do poszczególnych gatunków w zbiorze testowym wyznaczone za pomocą modelu liniowej regresji dla rozszerzonej przestrzeni cech

Dla zbioru testowego w rozszerzonej przestrzeni cech (rozkład prawdopodobieństw wykres 10), wyniki są zbliżone do tych z wykresu 8. Również widoczna jest bardzo dobra separacja etykietek dla każdego gatunku.



Wykres 11: Wykres liczby rzeczywistych etykietek klas względem prognozowanych w zbiorze testowym w przestrzeni rozszerzonej

Wykres 11 pokazuje bardzo dobrą klasyfikację. Jedynie jedna obserwacja z klasy `virginica` została źle dopasowana. Dokładność klasyfikacji jest na poziomie **98** %.

1.5.2 Przykład przeuczenia

Aby pokazać zjawisko przeuczenia modelu regresji liniowej na zbiorze `iris`, wyznaczmy dwa podzbiory zbiór uczący, lecz tym razem modyfikujemy proporcje podziału. Zbadamy dokładność klasyfikacji na zbiorze uczącym i testowym, zarówno dla domyślnego zbioru `iris`, jak i tego o rozszerzonej przestrzeni cech.

Tabela 4: Porównanie dokładności klasyfikacji dla różnych proporcji podziału

Proporcja (zb. uczący)	Iris (uczący)	Iris (testowy)	Iris rozsz. (uczący)	Iris rozsz. (testowy)
1/6	0.840	0.815	1.000	0.871
2/6	0.880	0.828	1.000	0.960
3/6	0.840	0.840	1.000	0.987
4/6	0.810	0.840	0.990	0.980
5/6	0.856	0.840	0.976	1.000

Analizując tabelę 4 możemy zauważyć, że w przypadku domyślnego zbioru `iris` dokładność klasyfikacji, zarówno dla zbioru uczącego jak i testowego, jest bliska 85% (różnice kilku punktów procentowych). Największą różnicę model osiągnął dla zbioru uczącego zawierającego $\frac{1}{3}$ wszystkich obserwacji zbioru `iris`, jednak najgorszy błąd występuje dla proporcji $\frac{1}{6}$. O wiele większe różnice w wynikach widoczne są dla zbioru o rozszerzonej przestrzeni cech. Największa różnica wystąpiła dla proporcji $\frac{1}{6}$, gdzie mamy idealną klasyfikację dla zbioru uczącego, natomiast dla zbioru testowego wynik jest o kilka punktów procentowych gorszy. Wystepuje zatem ryzyko przeuczenia modelu. Może to być spowodowane tym, że mamy zbyt mało danych do wyuczenia modelu. Natomiast wraz ze wzrostem liczby obserwacji w zbiorze uczącym dokladność klasyfikacji spada, a na zbiorze testowym rośnie. W przypadkach proporcji $\frac{3}{6}, \frac{4}{6}, \frac{5}{6}$ różnica błędów jest znacznie mniejsza (dla proporcji $\frac{5}{6}$ mamy idealną dokładność klasyfikacji).

1.6 Wnioski

Tabela 5: Dokładność klasyfikacji dla zbioru uczącego i testowego danych iris

Zbiór	Dokładność klasyfikacji
Uczący	0.81
Testowy	0.84
Uczący (rozszerzony)	0.99
Testowy (rozszerzony)	0.98

Podsumowując przeprowadzoną analizę i tabele 4 i 5, możemy wyciągnąć następujące wnioski:

- Zmiana proporcji podziałów na zbiór testowy i uczący nie ma znacznego wpływu na dokładność klasyfikacji w przypadku podstawowego zbioru danych `iris`. Powodem może być zjawisko częściowego maskowania klas.
- Zastosowanie przestrzeni cech rozszerzonej o wielomiany stopnia 2 gwarantuje niemal doskonałą klasyfikację. Eliminujemy w ten sposób zjawisko maskowania klas.
- Podstawowy zbiór `iris` jest dość odporny na przeuczenie modelu. Biorąc nawet drastyczny podział ($\frac{1}{6}$) wyniki są podobne do wyników z podziału domyślnego.
- Klasyfikacja oparta na modelu regresji liniowej jest dość dobrym modelem dla zbioru `iris`, odpornym w dużej mierze na przeuczenie. Jedynym problemem jest częściowe maskowanie klas, któremu można w prosty sposób zaradzić, rozszerzając przestrzeń cech.

2 Zadanie 2

2.1 Wybór i zapoznanie się z danymi

Wybraliśmy zbiór `Vehicle`, który zawiera informacje opisujące kształty pojazdów. Znajduje się w nim 846 obserwacji podzielonych na cztery klasy: Bus (autobus piętrowy), Opel (Opel Manta 400), Saab (Saab 9000)

i Van (Chevrolet van).

Poniżej znajduje się tabela przedstawiająca podgląd danych.

Tabela 6: Struktura zbioru danych Vehicle

Zmienna	Opis	Typ	Przykładowe_wartości
Comp	(Compactness) - stopień zwartości kształtu	numeric	95, 91, 104
Circ	(Circularity) - bliskość kształtu do koła	numeric	48, 41, 50
D.Circ	(Distance Circularity) - odległość od okręgu referencyjnego	numeric	83, 84, 106
Rad.Ra	(Radius Ratio) - stosunek promieni	numeric	178, 141, 209
Pr.Axis.Ra	(Principal axis aspect ratio) - stosunek osi głównej do drugorzędnej	numeric	72, 57, 66
Max.L.Ra	(Max length aspect ratio) - stosunek maksymalnej długości do szerokości	numeric	10, 9, 10
Scat.Ra	(Scatter ratio) - stopień rozproszenia	numeric	162, 149, 207
Elong	(Elongatedness) - Wydłużenie kształtu	numeric	42, 45, 32
Pr.Axis.Rect	(Principal axis rectangularity) - prostokątność względem osi głównej	numeric	20, 19, 23
Max.L.Rect	(Max length rectangularity) - prostokątność względem osi maksymalnej długości	numeric	159, 143, 158
Sc.Var.Maxis	(Scaled variance along major axis) - zeskalowana wariancja wzdłuż osi głównej	numeric	176, 170, 223
Sc.Var.maxis	(Scaled variance along minor axis) - zeskalowana wariancja wzdłuż osi drugorzędnej	numeric	379, 330, 635
Ra.Gyr	(Scaled radius of gyration) - zeskalowany promień bezwładności	numeric	184, 158, 220
Skew.Maxis	(Skewness about major axis) - skośność względem osi głównej	numeric	70, 72, 73
Skew.maxis	(Skewness about minor axis) - skośność względem osi drugorzędnej	numeric	6, 9, 14
Kurt.maxis	(Kurtosis about minor axis) - kurtoza względem osi drugorzędnej	numeric	16, 14, 9
Kurt.Maxis	(Kurtosis about major axis) - kurtoza względem osi głównej	numeric	187, 189, 188
Holl.Ra	(Hollows ratio) - stosunek przestrzeni pustej do całkowitej	numeric	197, 199, 196
Class	(Class) - typ samochodu	factor	van, van, saab

Typy są przypisane poprawnie. Dane mają wymiary: 846, 19. Jest jedna zmienna typu factor "Class", która zawiera informacje o przynależności do jednej spośród czterech klas: bus, opel, saab, van. Pozostałe 18 to zmienne numeryczne. Ponadto przyglądając się danym za pomocą `View` możemy zaobserwować, że w zbiorze nie ma nietypowego kodowania zmiennych brakujących, co więcej ilość NA wynosi 0.

2.2 Cel analizy

Naszym celem jest porównać skuteczność algorytmów, które po odpowiednim wytrenowaniu, powinny być w stanie automatycznie przypisać klasę dla danej obserwacji na podstawie danych numerycznych opisujących kształt pojazdu.

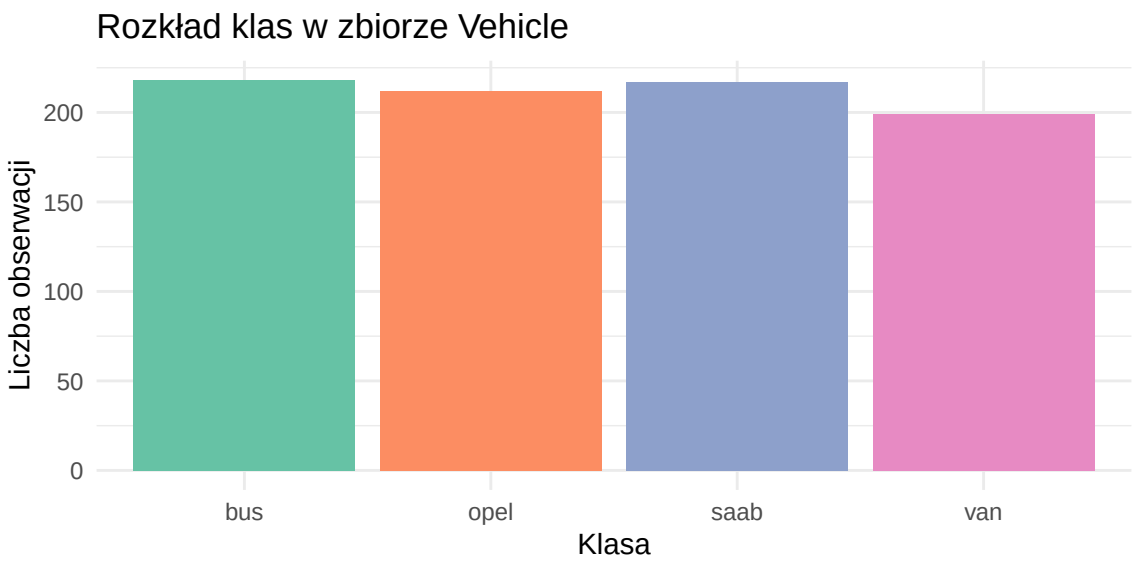
W analizie zastosujemy poznane algorytmy klasyfikacji i szczegółowo porównamy ich dokładność. Porównanie uwzględnia następujące algorytmy:

- metoda k-najbliższych sąsiadów (k-Nearest Neighbors),
- drzewa klasyfikacyjne (classification trees),
- naiwny klasyfikator bayesowski (naive Bayes classifier).

2.3 Wstępna analiza danych

Przeprowadzamy wstępną analizę danych zbioru `Vehicle` w celu wybrania cech o najlepszej dyskryminacji. W tym celu wykorzystane zostaną wykresy skrzypcowe, prezentujące rozkład poszczególnych zmiennych w zależności od typu samochodu. Przydatne może się również okazać przeprowadzenie testu ANOVA.

2.3.1 Rozkład klas



Wykres 12: Wykres słupkowy, przedstawiający ilość obserwacji dla poszczególnej klasy

Klasy są dość proporcjonalne co można zauważyć na wykresie 12. Różnica między najmniej liczną i najbardziej liczną klasą to **19** przy **846** przypadkach.

Gdybyśmy przypisali wszystkie obiekty do jednej, najczęściej występującej klasy, uzyskalibyśmy błąd klasyfikacji na poziomie **74**%, co pokazuje, że taki naiwny model byłby nieskuteczny.

2.3.2 Wariancja poszczególnych cech

Wariancja poszczególnych cech:

```
##        Comp        Circ      D.Circ      Rad.Ra   Pr.Axis.Ra     Max.L.Ra
##       8.234       6.170      15.772      33.472       7.888       4.601
##     Scat.Ra       Elong Pr.Axis.Rect  Max.L.Rect Sc.Var.Maxis Sc.Var.maxis
##      33.245       7.812       2.592      14.516      31.395      176.693
##      Ra.Gyr  Skew.Maxis  Skew.maxis  Kurt.maxis  Kurt.Maxis      Holl.Ra
##      32.546       7.487       4.918       8.931       6.164       7.439
```

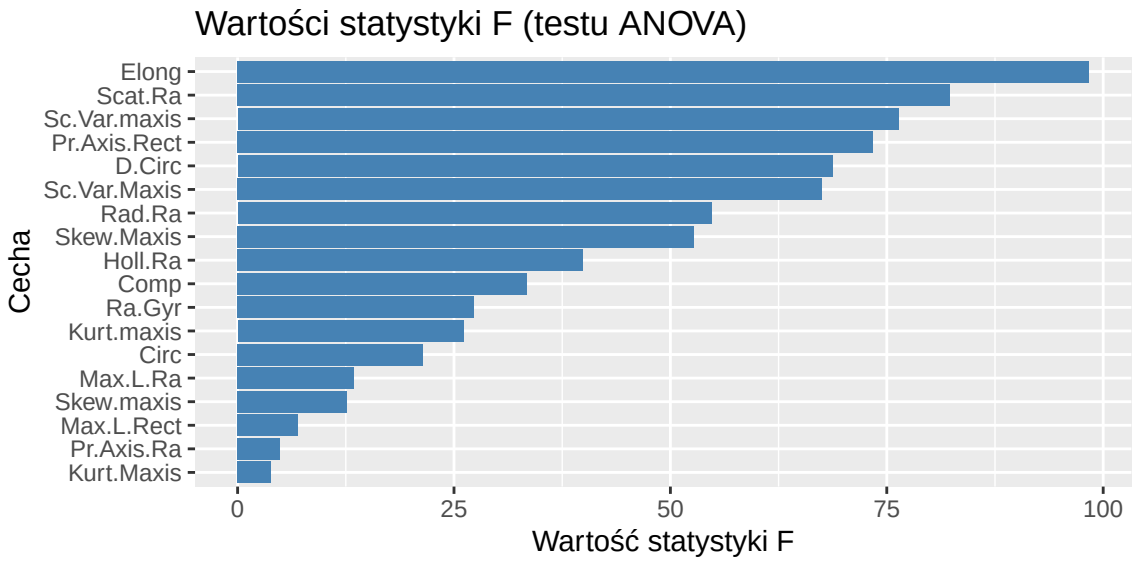
Najmniejsza wariancja to **3**, a największa to **177**. Bardzo duże różnice w wariancji poszczególnych cech. Zatem konieczne jest zastosowanie standaryzacji, w szczególności dla k-NN (w przeciwnym wypadku wpływ cech będzie niezrównoważony, co może doprowadzić do zdominowania odległości przez cechy o większym zakresie wartości). Choć dla drzew klasyfikacyjnych oraz naiwnego klasyfikatora bayesowskiego standaryzacja nie jest istotna to wykorzystamy standaryzowany zbiór dla wszystkich algorytmów w celu spójności.

2.3.3 Test ANOVA

Przeprowadzimy test ANOVA, który pomaga w rozstrzygnięciu, które cechy dobrze rozróżniają grupy pojazdów. Test porównuje dwie wariancje: międzygrupową i wewnątrzgrupową, a potem oblicza statystykę:

$$F = \frac{\text{wariancja międzygrupowa}}{\text{wariancja wewnątrzgrupowa}}$$

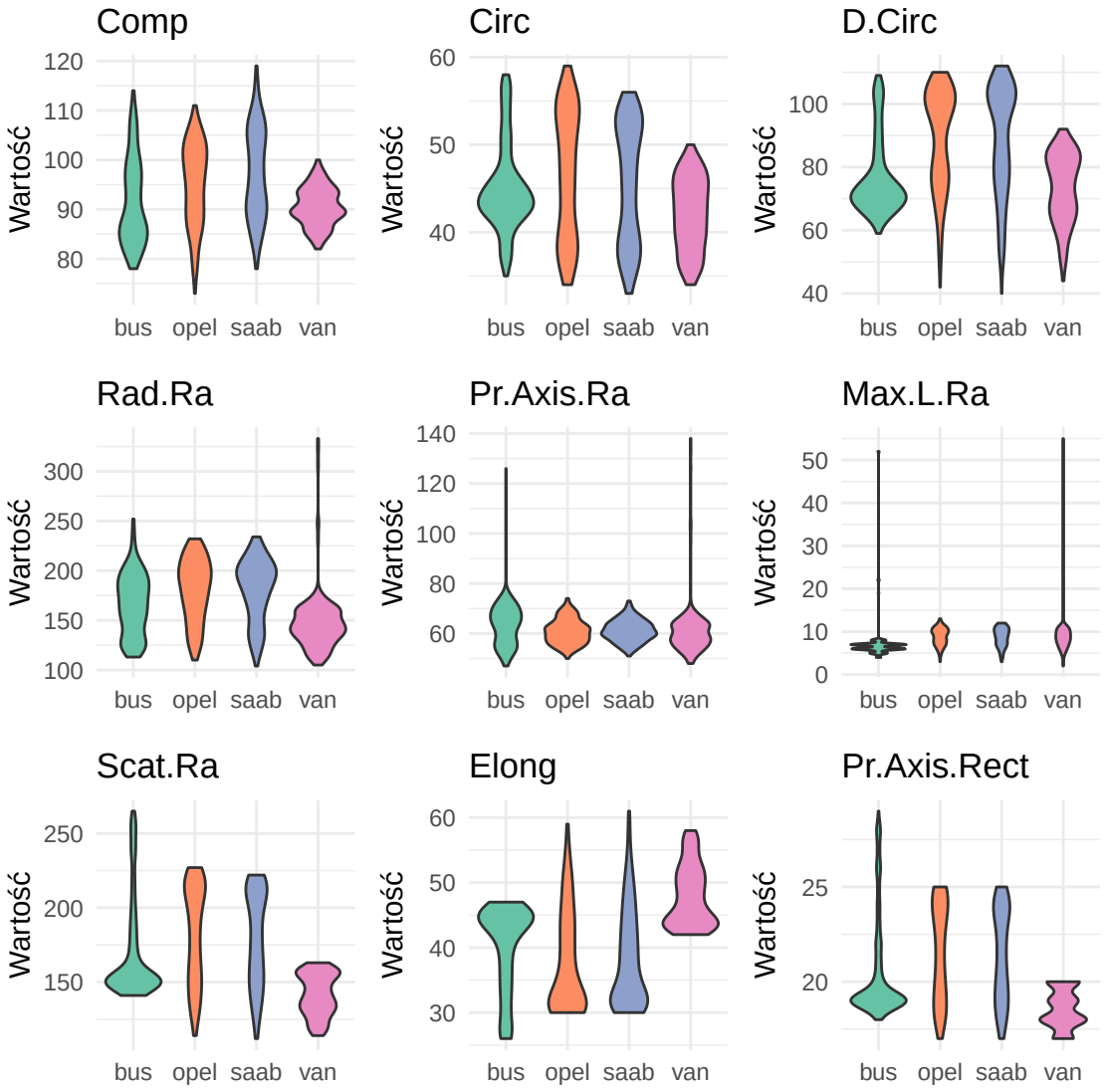
Duże F Świadczy o istotnych różnicach średnich między grupami.



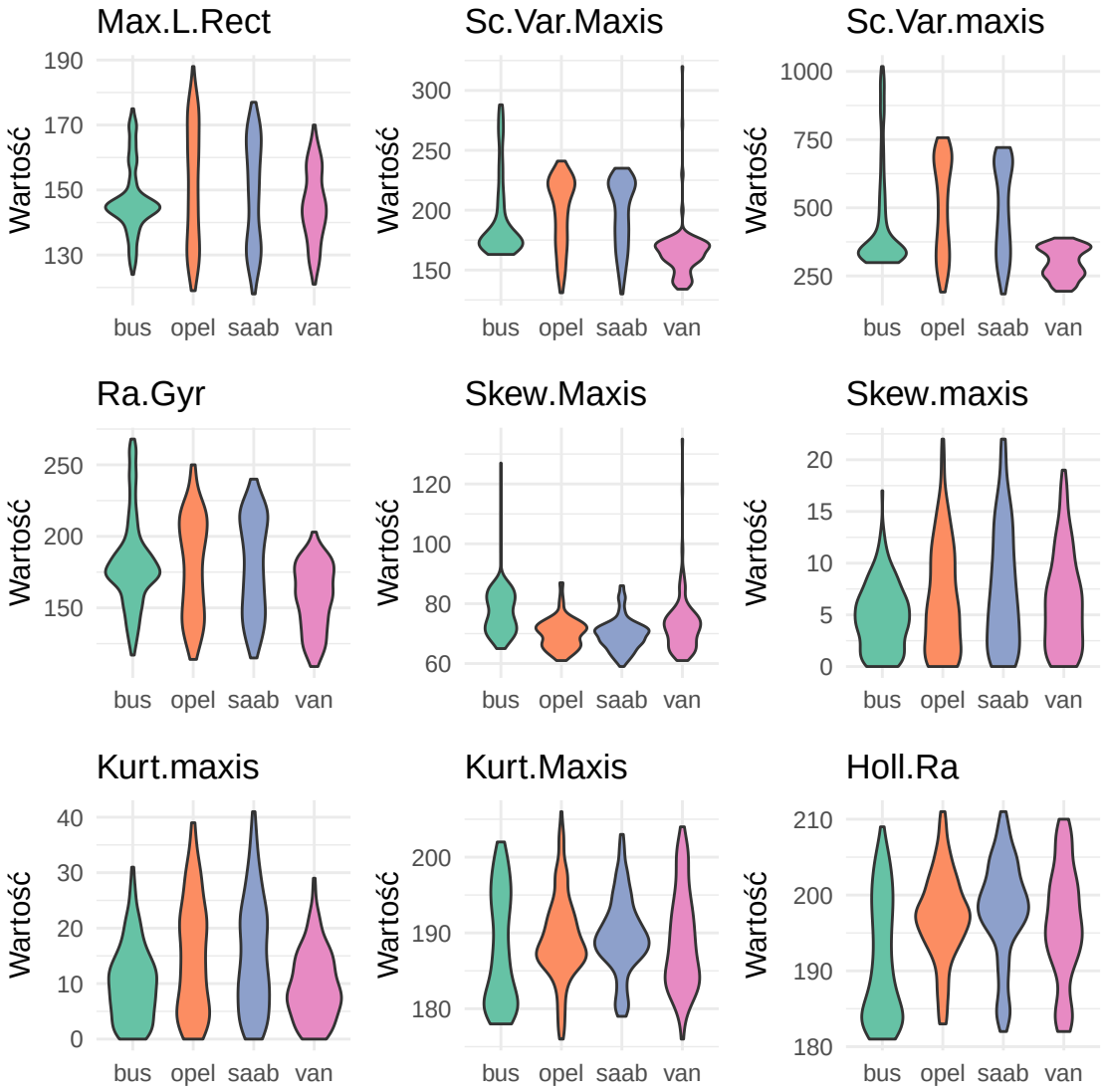
Wykres 13: Wykres słupkowy wartości statystyki F otrzymane za pomocą testu ANOVA dla poszczególnych cech zbioru Vehicle

Na wykresie 13 widzimy, że `Elong`, `Scat.Ra`, `Sc.Var.Maxis`, `D.Circ`, `Pr.Axis.Rect`, `Sc.Var.maxis` mają potencjał do bycia cechami, które dobrze rozróżniają grupy pojazdów.

2.3.4 Wykresy skrzypcowe



Wykres 14: Wybrane wykresy skrzypcowe pomagające w oszacowaniu zdolności dyskryminacyjnych poszczególnych cech



Wykres 15: Wybrane wykresy skrzypcowe pomagające w oszacowaniu zdolności dyskryminacyjnych poszczególnych cech

Na podstawie wykresu 14 oraz 15, najlepszymi zdolnościami dyskryminacyjnymi cechują się zmienne:

- `Comp`, ponieważ rozkład zmiennej dla każdej klasy jest choć trochę przesunięte względem pozostałych,
- `D.Circ`, ponieważ klasy Bus i Van są dość dobrze oddzielone od Opel i Saab,
- `Elong` dość dobrze odróżnia Vana,
- `Scat.Ra` odróżnia Vana od reszty,
- `Holl.Ra` odróżnia Busa.

Brak cechy, która wyraźnie odróżnia Opla Mante 400 od Saaba 9000, ale to prawdopodobnie dlatego, że te samochody mają bardzo podobny kształt. Według testu ANOVA `Holl.Ra` i `Comp` są dość przeciętne pod względem różnicowania klas, ale weźniemy je pod uwagę, ponieważ wykresy skrzypcowe zdradzają ich ukryty potencjał.

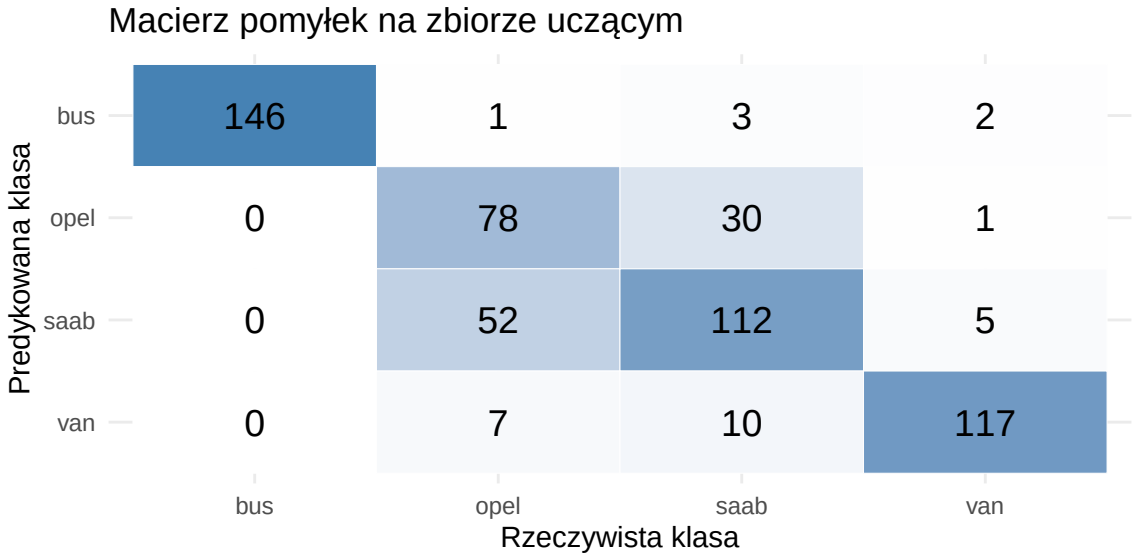
2.4 Ocena dokładności klasyfikacji i porównanie metod

W niniejszym podrozdziale zajmiemy się sprawdzeniem błędów klasyfikacji dla wszystkich 3 algorytmów na całym zbiorze danych. Porównamy błąd otrzymany na zbiorze uczącym i testowym, dodatkowo spojrzymy na macierze pomyłek.

2.4.1 Metoda k-Nearest Neighbors

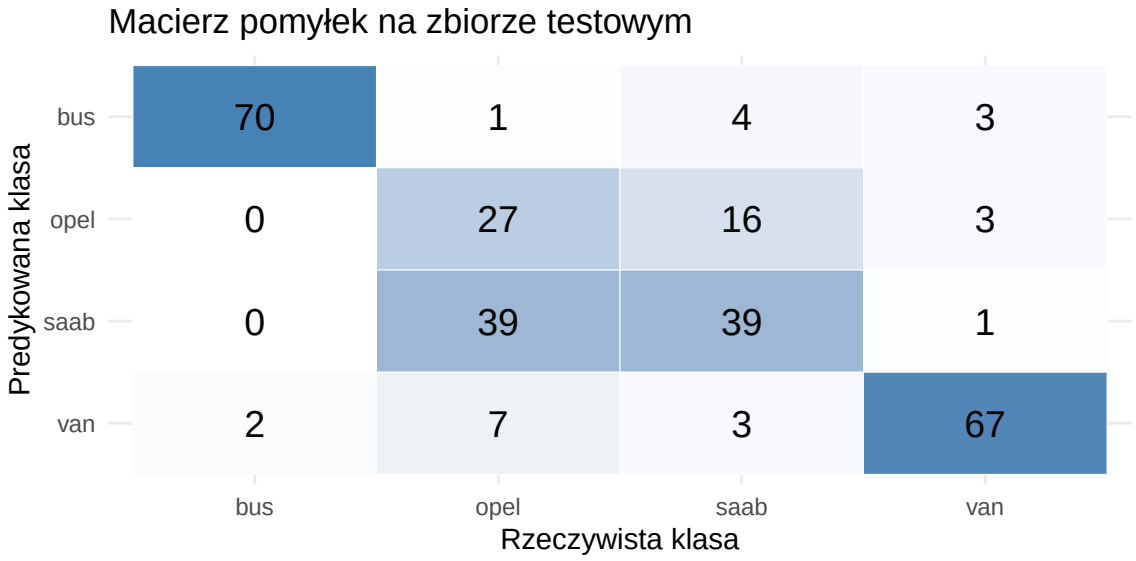
Metoda k-NN (k-Nearest Neighbors) polega na klasyfikowaniu obiektów na podstawie ich k najbliższych sąsiadów w przestrzeni cech. Klasa nowego punktu jest zazwyczaj określana większością głosów spośród tych sąsiadów.

2.4.1.1 Jednorazowy podział Standaryzowany zbiór `Vehicle` dzielimy na uczący ($\frac{2}{3}$ zbioru) i testowy ($\frac{1}{3}$ zbioru) przy użyciu funkcji `sample`. Liczbę sąsiadów ustalamy na `k = 5`.



Wykres 16: Macierz pomyłek przedstawiająca błędy w klasyfikacji na zbiorze uczącym dla metody k-NN

Błąd klasyfikacji dla zbioru uczącego wynosi 20%.



Wykres 17: Macierz pomyłek przedstawiająca błędy w klasyfikacji na zbiorze testowym dla metody k-NN

Błąd klasyfikacji dla zbioru testowego wynosi 28%.

Wnioski:

- Błąd na zbiorze uczącym 20% oznacza, że klasyfikator nie dopasował się perfekcyjnie, ale też nie jest przeuczony,
- Błąd na zbiorze testowym 28% jest wyższy niż na uczącym i może oznaczać, że dane są trudne do klasyfikacji lub wybraliśmy nieoptymalny parametr `k`, czyli liczbę sąsiadów.

Uwaga: Wyniki mogą być przypadkowe, ponieważ przeprowadziliśmy klasyfikację tylko na jednym konkretnym podziale danych.

2.4.1.2 Zaawansowane schematy oceny dokładności

W celu bardziej rzetelnej oceny jakości klasyfikatora k-NN stosujemy trzy metody walidacji wielokrotnej: walidację krzyżową, bootstrap oraz poprawiony bootstrap `.632+`.

Oszacowane błędy klasyfikacji dla standaryzowanego zbioru danych `Vehicle` wynoszą:

- Walidacja krzyżowa: 27.9%
- Bootstrap: 31.5%
- Metoda `.632+`: 26.8%

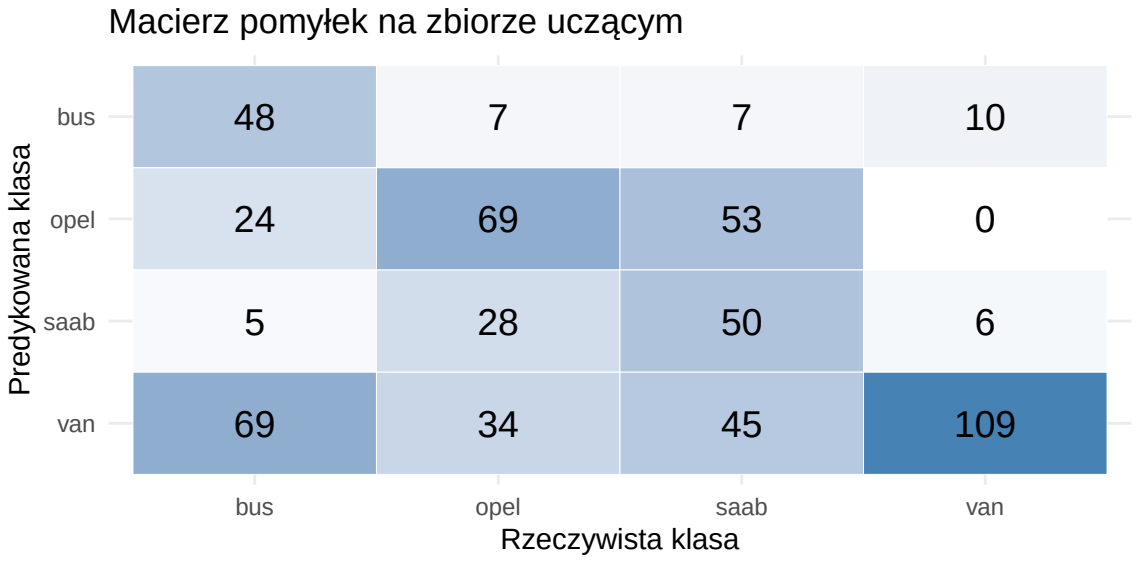
Najniższy błąd uzyskujemy metodą `.632+`, co sugeruje, że rzeczywisty błąd klasyfikatora wynosi około 26.8%. Wynik ten jest zbliżony do wcześniej uzyskanego błędu na zbiorze testowym (28%), co potwierdza stabilność klasyfikatora k-NN. Wyższy błąd w metodzie bootstrap wskazuje na jej tendencję do zawyżania oszacowań.

2.4.2 Naiwny klasyfikator bayesowski

Metoda Naiwnego Bayesa to klasyfikator oparty na twierdzeniu Bayesa, zakładający niezależność cech. Szacuje prawdopodobieństwo przynależności obiektu do danej klasy, wybierając tę z największym prawdopodobieństwem posteriori.

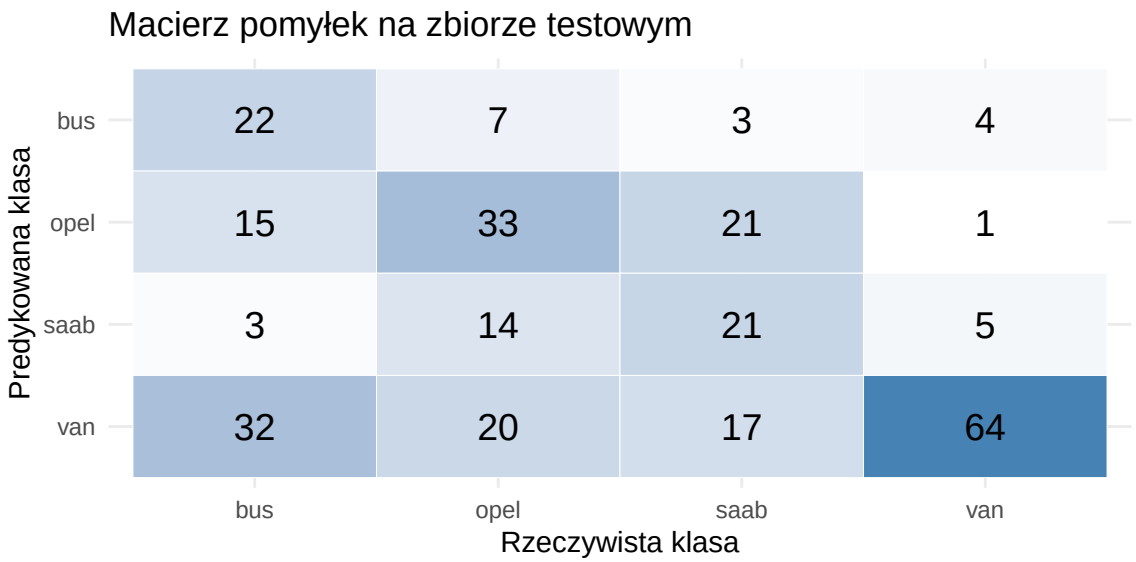
2.4.2.1 Jednorazowy podział

Standaryzowany zbiór `Vehicle` dzielimy na uczący ($\frac{2}{3}$ zbioru) i testowy ($\frac{1}{3}$ zbioru) przy użyciu funkcji `sample`.



Wykres 18: Macierz pomyłek przedstawiająca błędy w klasyfikacji na zbiorze uczącym dla klasyfikatora bayesowskiego

Błąd klasyfikacji dla zbioru uczącego wynosi 51%



Wykres 19: Macierz pomyłek przedstawiająca błędy w klasyfikacji na zbiorze testowym dla klasyfikatora bayesowskiego

Błąd klasyfikacji dla zbioru testowego wynosi 50%

Wnioski:

- Bardzo duży błąd na zbiorze uczącym jak i testowym oznacza, że model nie dopasował się do danych. Prawdopodobnie wynika to z tego, że cechy nie są warunkowo niezależne co łamie założenia tej metody.
- Model potrafi skutecznie odróżnić tylko klasę `Van` co może oznaczać, że jest ona najbardziej odrębna od pozostałych klas.

2.4.2.2 Zaawansowane schematy oceny dokładności W celu bardziej rzetelnej oceny jakości naiwnego klasyfikatora bayesowskiego stosujemy trzy metody walidacji wielokrotnej: walidację krzyżową, bootstrap oraz poprawiony bootstrap `.632+`.

Oszacowane błędy klasyfikacji dla standaryzowanego zbioru danych `Vehicle` wynoszą:

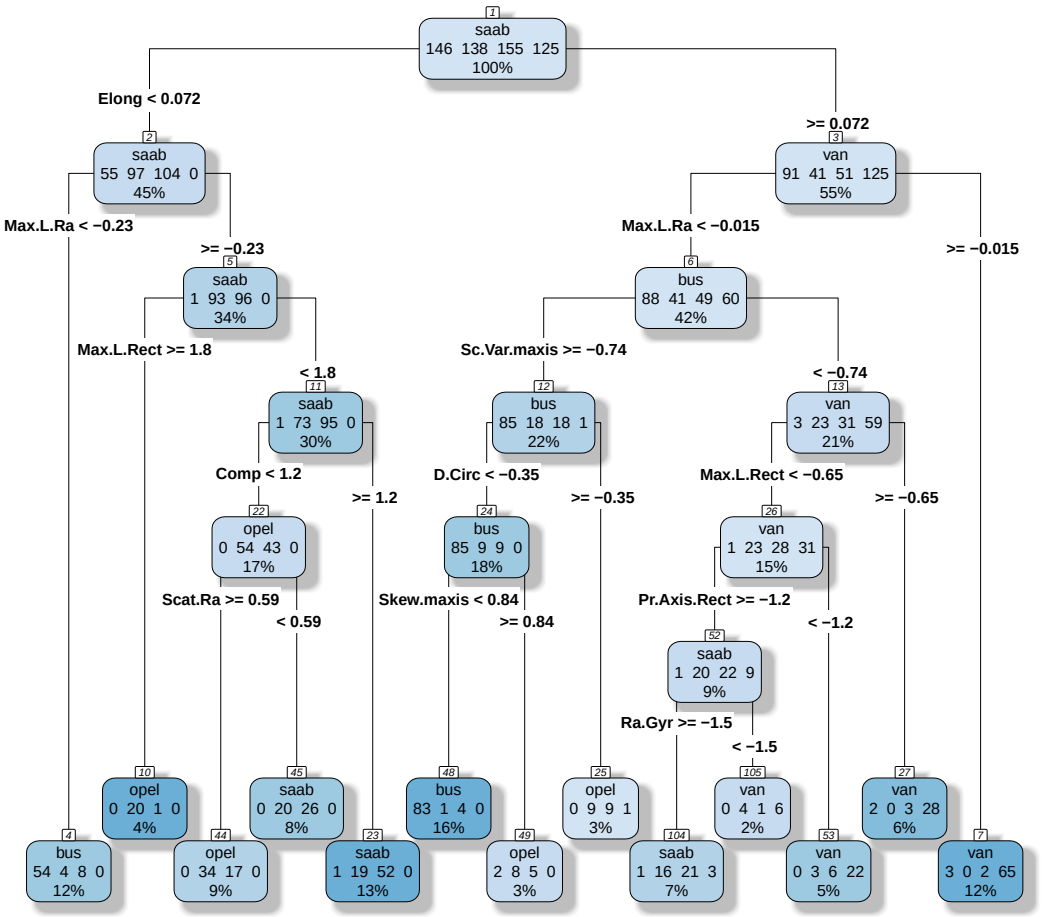
- Walidacja krzyżowa: 54.4%
- Bootstrap: 54.9%
- Metoda `.632+`: 54.5%

Otrzymaliśmy jeszcze większe błędy co potwierdza, że naiwny klasyfikator bayesowski nie nadaje się do użycia na zbiorze danych `Vehicle`.

2.4.3 Drzewo klasyfikacyjne

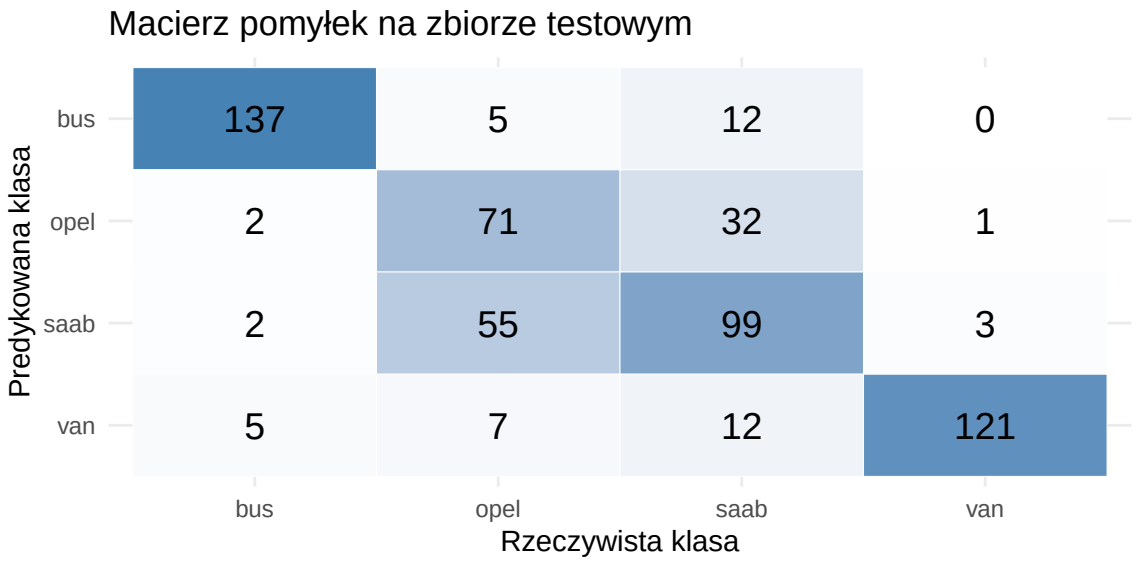
Drzewo klasyfikacyjne to model predykcyjny, który podejmuje decyzje na podstawie sekwencji warunków logicznych, dzieląc dane na podzbiory według cech. Struktura drzewa umożliwia łatwą interpretację i wizualizację procesu klasyfikacji.

2.4.3.1 Jednorazowy podział Standaryzowany zbiór `Vehicle` dzielimy na uczący ($\frac{2}{3}$ zbioru) i testowy ($\frac{1}{3}$ zbioru) przy użyciu funkcji `sample`. Na początku nie przycinamy zbytnio drzewa `cp=0.01`, jego rozmiar wynosi 13.



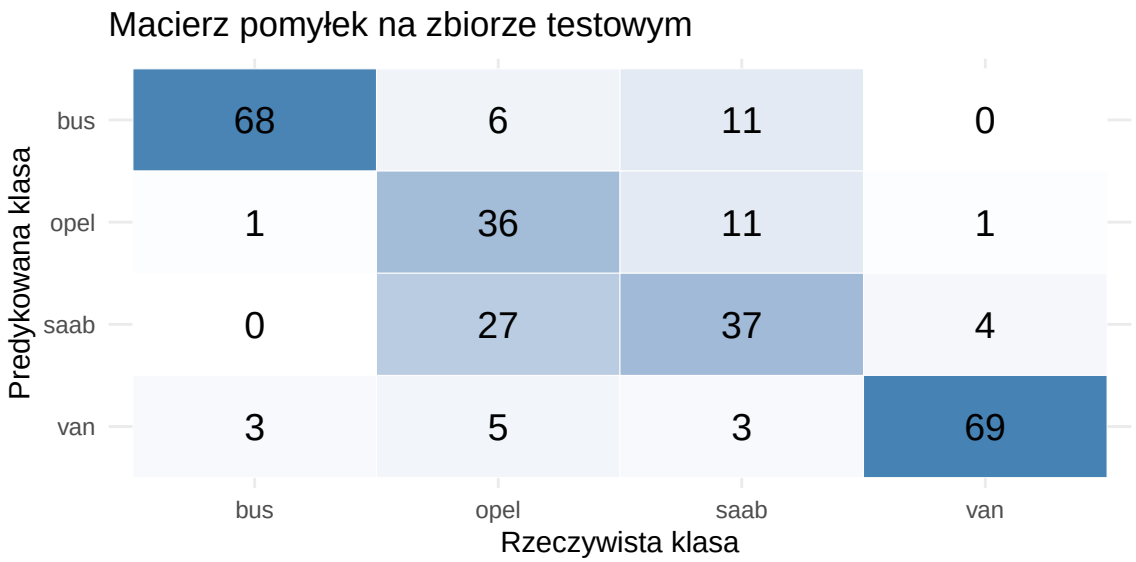
Wykres 20: Wizualizacja drzewa z podstawowymi parametrami dla cp=0.01

Wykres 20 umożliwia bezpośredni wgląd w decyzje podejmowane przez model. Największe role odgrywają zmienne `Elong` i `Max.L.Ra`, gdyż definiują początkowe podziały. Przyglądając się liściom, możemy zauważyć, że `van` i `bus` charakteryzują się największą czystością węzłów. Największym problemem w klasyfikacji jest oddzielenie typu `opel` od `saab` (W wielu liściach jest zbliżona liczba obserwacji z obu tych klas). Zmienne takie jak `Skew.maxis` i `Ra.Gyr` nie wnoszą wiele do poprawy klasyfikacji, zatem widoczny jest potencjał do przycięcia drzewa i uproszczenia modelu.



Wykres 21: Macierz pomyłek przedstawiająca błędy w klasyfikacji na zbiorze uczącym dla drzewa klasyfikacyjnego

Błąd klasyfikacji dla zbioru uczącego wynosi 24%



Wykres 22: Macierz pomyłek przedstawiająca błędy w klasyfikacji na zbiorze testowym dla drzewa klasyfikacyjnego

Błąd klasyfikacji dla zbioru testowego wynosi 26%

Wnioski:

- Bardzo podobny błąd na zbiorze uczącym (24%) i testowym (26%) oznacza, że klasyfikator nie jest przeuczony. Ponadto błąd klasyfikacji na zbiorze testowym jest podobny do metody k-NN (28%)

2.4.3.2 Zaawansowane schematy oceny dokładności W celu bardziej rzetelnej oceny jakości drzewa klasyfikacyjnego możemy zastosować trzy metody walidacji wielokrotnej: walidację krzyżową, bootstrap oraz

poprawiony bootstrap `.632+`.

Oszacowane błędy klasyfikacji dla standaryzowanego zbioru danych `Vehicle` wynoszą:

- Walidacja krzyżowa: 30.9%
- Bootstrap: 32.5%
- Metoda `.632+`: 30.2%

Wyniki uzyskane zaawansowanymi metodami są trochę gorsze niż przy jednokrotnym podziale. Mamy okazję zauważyć, że warto stosować bardziej zaawansowane metody oceny dokładności, ponieważ przy jednokrotnym podziale mogliśmy po prostu mieć szczęście i dlatego wynik był lepszy.

2.5 Różne parametry i różne podzbiory cech

W tym podrozdziale sprawdzimy jak wybranie podzbioru danych na podstawie wykresów skrzypcowych 14, 15 i testu ANOVA 13 wpływa na błąd klasyfikacji. Ponadto dla zbioru, który będzie dawać najlepszy wynik pozmieniamy parametry algorytmów i zobaczymy jak wpływają one na wyniki.

2.5.1 Metoda k-Nearest Neighbors

Ten podrozdział zaczniemy od obserwacji jak wybór różnych podzbiorów `Vehicle` wpływa na błąd klasyfikatora dla algorytmu k-NN.

Przypomnijmy wyniki tego klasyfikatora gdy braliśmy po uwagę wszystkie zmienne numeryczne:

- Walidacja krzyżowa: 27.9%
- Bootstrap: 31.5%
- Metoda `.632+`: 26.8%

Zaczniemy od wyznaczenia błędu k-NN gdy podzbiór wybieramy kierując się wykresami skrzypcowymi na 14 oraz 15 . Na tej podstawie najbardziej obecującymi cechami są: `Comp`, `D.Circ`, `Elong`, `Scat.Ra`, `Holl.Ra`. W celu łatwiejszego odwoływania się do tego podzioru, nazwijmy go jako `model 1`.

- Walidacja krzyżowa: 31.1%
- Bootstrap: 33.7%
- Metoda `.632+`: 30.6%

Wybierając ten podzbiór tylko pogorszyliśmy wyniki.

Wybierzmy więc inny kierując się również metodą ANOVA 13 wybierzmy podzbiór (`model 2`): `Comp`, `D.Circ`, `Scat.Ra`, `Elong`, `Holl.Ra`, `Pr.Axis.Rect`, `Sc.Var.maxis`, `Sc.Var.Maxis`

- Walidacja krzyżowa: 32%
- Bootstrap: 33.3%
- Metoda `.632+`: 30.1%

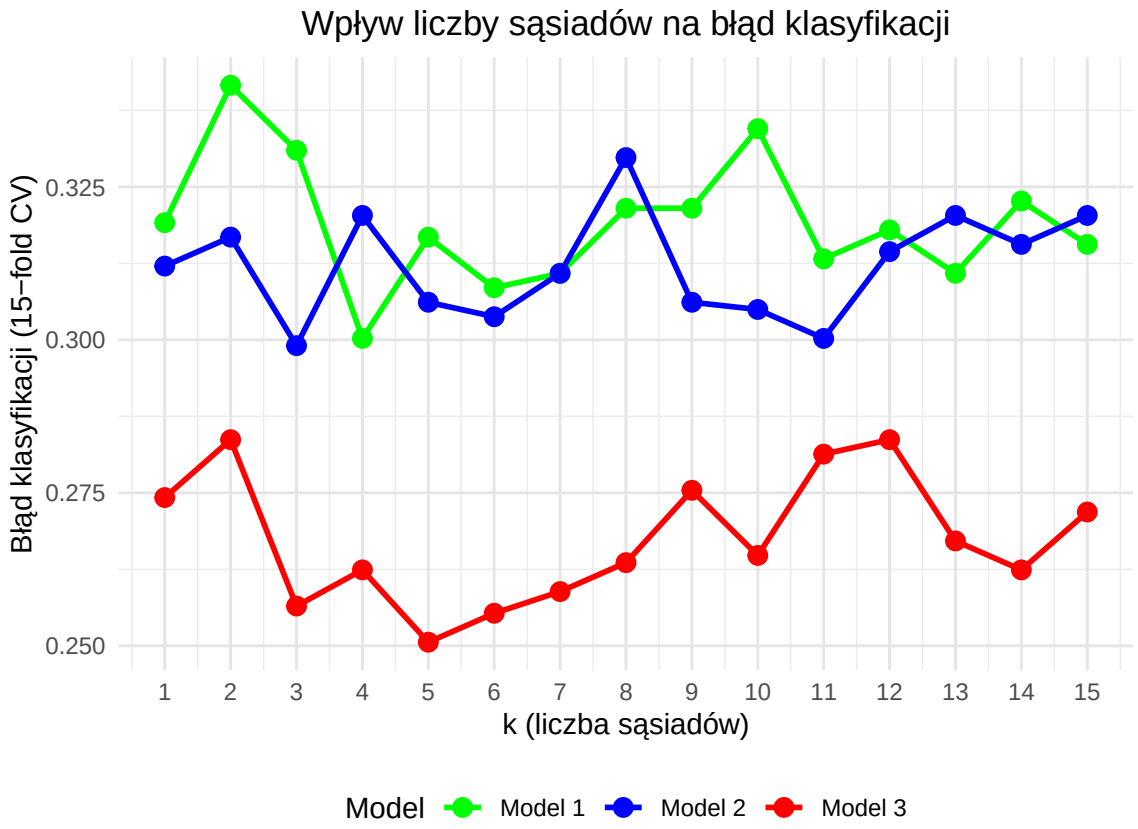
Wyniki wciąż są gorsze niż dla całego zbioru.

Metodą prób i błędów możemy testować różne podzbiory. Na przykład: `Comp`, `Circ`, `D.Circ`, `Pr.Axis.Ra`, `Scat.Ra`, `Elong`, `Max.L.Ra`, `Sc.Var.maxis`, `Max.L.Rect`, `Pr.Axis.Rect`, `Ra.Gyr`. Nazwijmy ten podzbiór jako `model 3`

- Walidacja krzyżowa: 25.5%
- Bootstrap: 28.6%
- Metoda `.632+`: 24.3%

Taki wybór znacznie poprawia wynik, względem tego który uzyskaliśmy dla całego zbioru. Widać, że ani wykresy skrzypcowe ani ANOVA nie pomogły nam w wyborze podzbioru, który daje lepszy wynik. W przypadku tego zbioru danych najlepiej wyznaczyć podzbiór ręcznie.

Spróbujmy poprawić wynik dla tego podzbioru dobierając odpowiednią liczbę sąsiadów



Wykres 23: Wykres przedstawiający zależność błędu od liczby sąsiadów w metodzie k-NN (metoda walidacji krzyżowej)

Okazuje się, że k = 5 sąsiadów to najlepszy wybór dla modelu 3, więc pozostawiamy k bez zmian.

Ostatecznie:

Tabela 7: Dokładność klasyfikacji metodą k-NN dla różnych podzbiorów zmiennych i k = 5

model	cross-validation	bootstrap	.632+
model1	0.31	0.34	0.31
model2	0.32	0.33	0.30
model3	0.26	0.29	0.24

Wykorzystując liczbę sąsiadów k = 5 i model 3 uzyskujemy najniższy błąd klasyfikacji dla wszystkich zastosowanych schematów.

2.5.2 Naiwny klasyfikator bayesowski

W przypadku naiwnego klasyfikatora bayesowskiego porównamy go dla 3 różnych podzbiorów, które określiliśmy wcześniej. Co ciekawe podzbiór najlepszy dla k-NN daje największy błąd przy użyciu "bayesa".

Tabela 8: Dokładność klasyfikacji metodą naiwnego klasyfikatora Bayesowskiego dla różnych podzbiorów zmiennych

model	cross-validation	bootstrap	.632+
model1	0.55	0.57	0.55
model2	0.56	0.56	0.56
model3	0.57	0.57	0.57

2.5.3 Drzewa klasyfikacyjne

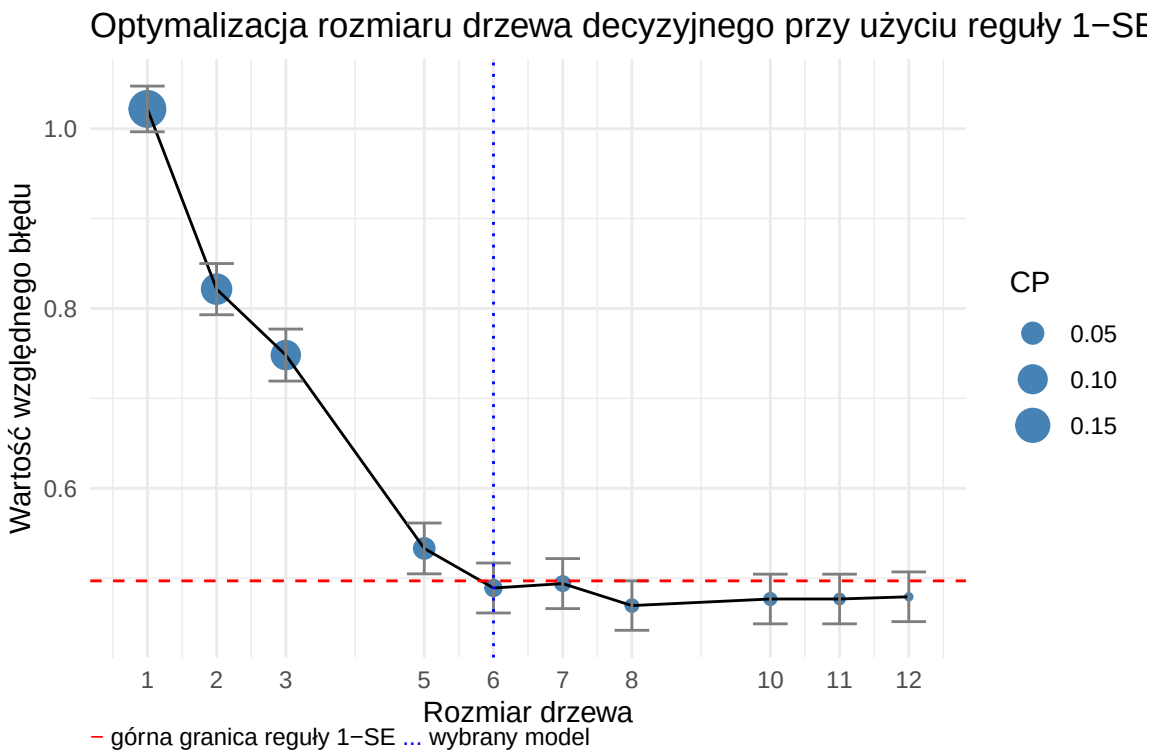
Najpierw sprawdzamy jak zmienia się błąd w zależności od wybranego podzbioru (parametr `cp = 0.01` bez zmian)

Tabela 9: Dokładność klasyfikacji metodą drzew klasyfikacyjnych dla różnych podzbiorów zmiennych

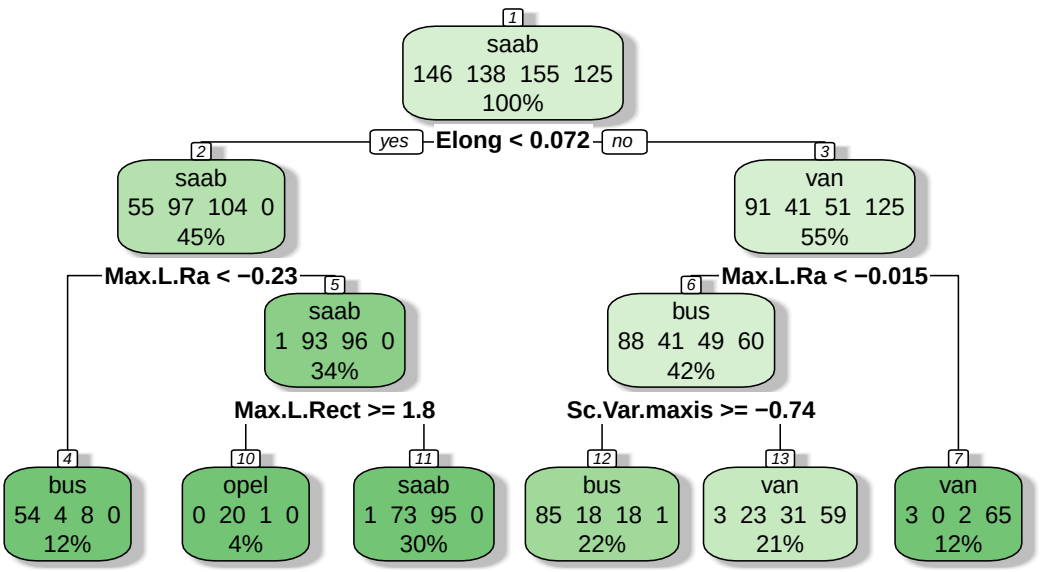
model	cross-validation	bootstrap	.632+
model1	0.35	0.36	0.33
model2	0.32	0.32	0.31
model3	0.31	0.32	0.31

Model 3 jest najskuteczniejszy podobnie jak w k-NN. Jednak tym razem nie poprawia on wyniku względem całego zbioru, błąd pozostaje podobny.

W drzewie klasyfikacyjnym możemy również sprawdzić jak jego rozmiar wpływa na błąd. Będziemy tutaj wykorzystywać podzbiór o nazwie model 3.



Wykres 24: Zależność błędu klasyfikacji od rozmiaru drzewa (opartego na modelu 3)



Wykres 25: Wizualizacja drzewa z podstawowymi parametrami po przycięciu

Wybieramy rozmiar drzewa równy 6 (czyli nsplit = 5), ponieważ jest to najmniejszy możliwy rozmiar drzewa, którego błąd walidacji krzyżowej mieści się w granicy wyznaczonej przez regułę 1-SE, czyli minimum błędu + odchylenie standardowe. Taki wybór pozwala uniknąć przeuczenia i zapewnia lepszą zdolność uogólniania niż bardziej złożone drzewa o podobnym błędzie.

Tabela 10: Dokładność klasyfikacji metodą drzew klasyfikacyjnych dla różnych podzbiorów zmiennych

model	cross-validation	bootstrap	.632+
model1	0.36	0.35	0.33
model2	0.34	0.35	0.33
model3	0.34	0.35	0.33

Co ciekawe po przycięciu wyniki są podobne dla wszystkich trzech modeli. Błąd przed przycięciem dla całego zbioru uzyskany metodą `.632+` wynosił 30.2%, a po przycięciu i po użyciu podzbioru (model 3) uzyskujemy wynik 33.3%, który nie jest dużo gorszy, a czytelność drzewa jak i zdolność generalizacji mogły się poprawić. Jest to akceptowalny kompromis między złożonością a skutecznością.

2.6 Podsumowanie

Dla jakiego podzbioru zmiennych predykcyjnych i dla jakich parametrów poszczególnych metod otrzymujemy najlepsze wyniki?

- Podzbiór wyznaczony metodą prób i błędów "model 3" daje najlepszy wynik dla `k-NN` i `drzew klasyfikacyjnych`, ale dla `naiwnego klasyfikatora bayesowskiego` błąd nieznacznie się zwiększa.
- W metodzie `k-NN` dla modelu 3 najlepiej użyć k = 5 sąsiadów.
- Drzewo klasyfikacyjne dobrze jest przyciąć, żeby było czytelniejsze, do rozmiaru 6, ale nieznacznie zwiększa to błąd. Najlepszy wynik uzyskaliśmy bez przycinania.

Która z metod klasyfikacyjnych daje lepsze, a które gorsze rezultaty w przypadku analizowanych danych?

- W przypadku danych `Vehicle` metoda `naiwnego klasyfikatora bayesowskiego` daje najgorsze rezultaty.
- Metoda `k-NN` po znalezieniu odpowiedniego podzbioru daje najlepszy wynik 24.3% (sprawdzany metodą `.632+`).
- Drzewo klasyfikacyjne jest niewiele gorsze od `k-NN`, daje nawet wynik: 33.3% (sprawdzany metodą `.632+`) być może gdybyśmy sprawdzili inne kombinacje zmiennych w podzbiorach to ta metoda mogłaby być lepsza od `k-NN`.

Czy wybór schematu oceny dokładności miał istotny wpływ na wnioski dotyczące skuteczności metod?

Chociaż metoda bootstrap często zawyża błąd, a .632+ daje najbardziej optymistyczne wyniki, to stosowanie różnych technik walidacji jest uzasadnione. Każda z nich ocenia model z innej perspektywy, dzięki czemu porównania między klasyfikatorami są bardziej wiarygodne i odporne na przypadkowość pojedynczego oszacowania. Różne metody mają różne właściwości i mogą inaczej oceniać błąd, więc pokazanie ich wszystkich jest cenne. Uwzględnienie trzech schematów pozwala na bardziej zrównoważoną ocenę skuteczności klasyfikatora i zwiększa wiarygodność porównań między modelami.

Spostrzeżenia

- Opel jak i Saab są najtrudniejsze do rozróżnienia nawet po dobraniu odpowienich parametrów w najskuteczniejszej metodzie `k-NN`. Wynika to oczywiście z podobieństwa kształtu tych pojazdów. Prawdopodobine potrzebowalibyśmy dużo więcej danych lub bardziej zaawansowanego algorytmu aby rozróżnić tak podobne pojazdy. Bus piętrowy i Van są dużo łatwiejsze do odróżnienia. Na macierzach pomyłek możemy zaobserwować, że są bardzo skutecznie sklasyfikowane. Co więcej Van jest na tyle odrębny, że nawet Naive Bayes potrafi dobrze odróżnić Vana od pozostałych pojazdów.
- W analizie zbioru `Vehicle` najlepiej sprawdził się algorytm `k-NN` z odpowiednio dobranym podzbiorem cech (`model 3`). Natomiast `naiwny klasyfikator bayesowski` wykazał się niską skutecznością, co może wskazywać na silne współzależności cech.
- Warto zauważyć, że wybór cech na podstawie ANOVA i wykresów skrzypcowych nie zawsze prowadzi do najlepszego podzbioru.