Sprawozdanie 3

Eksploracja danych

Kacper Szmigielski, 282255 i Mateusz Wizner

2025-06-04

Spis treści

1 Zadanie 1

1.1 a) Analizowane dane

Zbiór danych Iris to klasyczny zestaw danych w statystyce i uczeniu maszynowym, wprowadzony przez R.A. Fishera w 1936 roku. Zawiera 150 obserwacji kwiatów z trzech gatunków (K=3 klasy) irysa: setosa, versicolor i virginica.

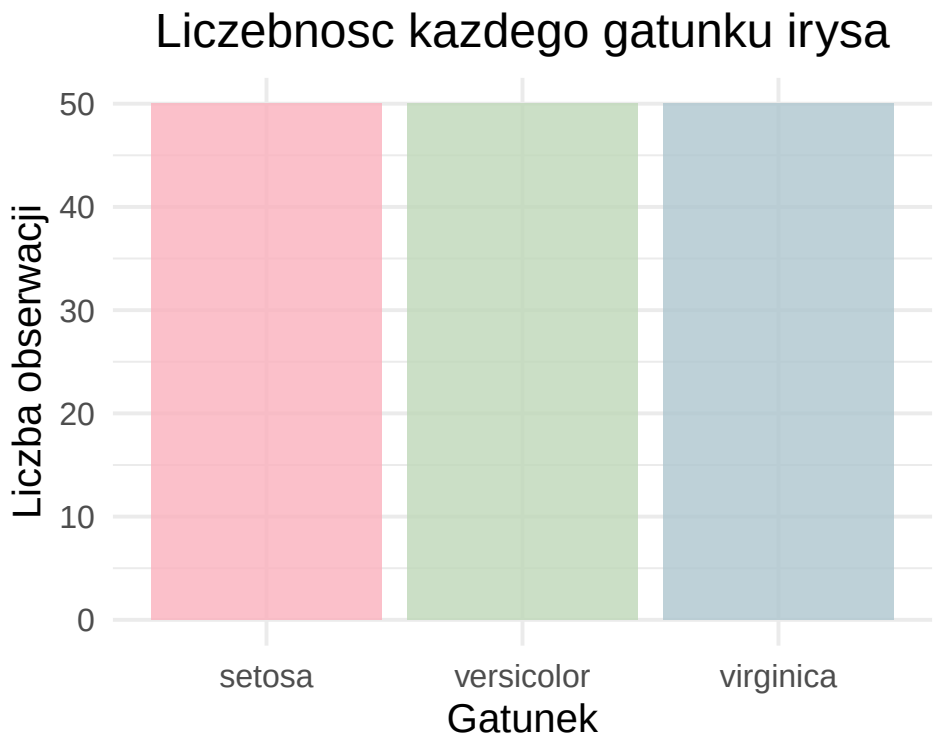
Każdy rekord opisuje pojedynczy kwiat za pomocą czterech cech numerycznych (p=4 cechy ilościowe) :

- Sepal.Length – długość działki kielicha (w cm)
- Sepal.Width – szerokość działki kielicha (w cm)
- Petal.Length – długość płatka (w cm)
- Petal.Width – szerokość płatka (w cm)

Przykładowe **3 wiersze** z danych iris

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
7.0	3.2	4.7	1.4	versicolor
6.3	3.3	6.0	2.5	virginica

Liczebność klas w danym zbiorze

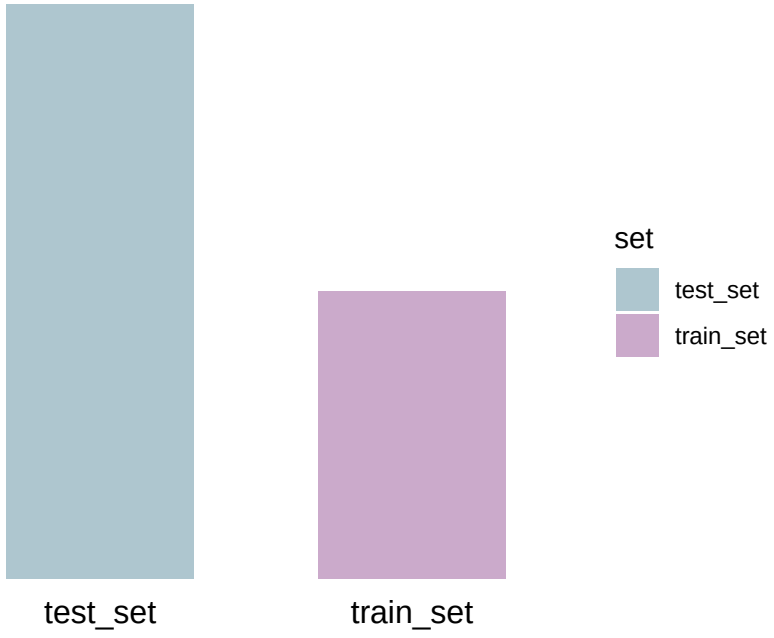


Mamy równy podział danych w zbiorze. Obserwacji każdego gatunku jest 50.

1.2 b) Podział danych na zbiór uczący i testowy

Dane zostały podzielone tak, aby zachować proporcje poszczególnych klas **(każda klasa zajmuje ok33% wszystkich danych)**, dzięki czemu zbiór uczący zawiera **reprezentatywną i równomierną próbkę wszystkich klas.**

Po takim podziale **zbiór uczący zawiera** $\frac{1}{3}$ danych, a **zbiór testowy zawiera** $\frac{2}{3}$ wszystkich danych.



1.3 c) Konstrukcja klasyfikatora i wyznaczenie prognoz

1.3.1 Inicjalizacja klasyfikatora

- Na początek wyznaczamy macierz modelu (macierze eksperymentu), zawierającą wartości poszczególnych zmiennych (dla odpowiednio danych testowych oraz uczących)

X1 - macierz dla danych testowych X2 - macierz dla danych trenujących

```
K=3
p = 4
n1= 99
n2 = 51
# X1 - macierz eksperymentu (ang. design matrix) dla danych TESTOWYCH
X1 <- cbind(rep(1,99), test_set[,1:4])
X1 <- as.matrix(X1)

# X2 - macierz eksperymentu (ang. design matrix) dla danych UCZĄCYCH
X2 <- cbind(rep(1,51), train_set[,1:4])
X2 <- as.matrix(X2)
```

- Następnie tworzę macierz wskaźnikową Y2 wymiaru 51 **(u nas podział zbioru to 99/51)** x K, która zawiera zmienne binarne kodujące poszczególne klasy.

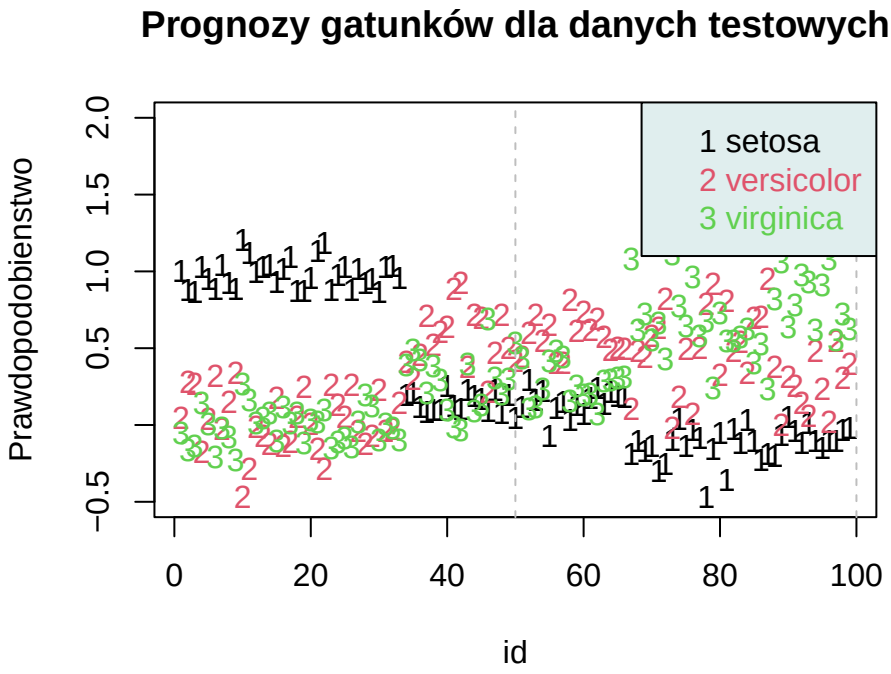
1.3.2 Estymacja współczynników i konstrukcja prognoz

- Wykorzystujemy metodę najmniejszych kwadratów (MNK) aby wyznaczyć estymatory współczynników modelu

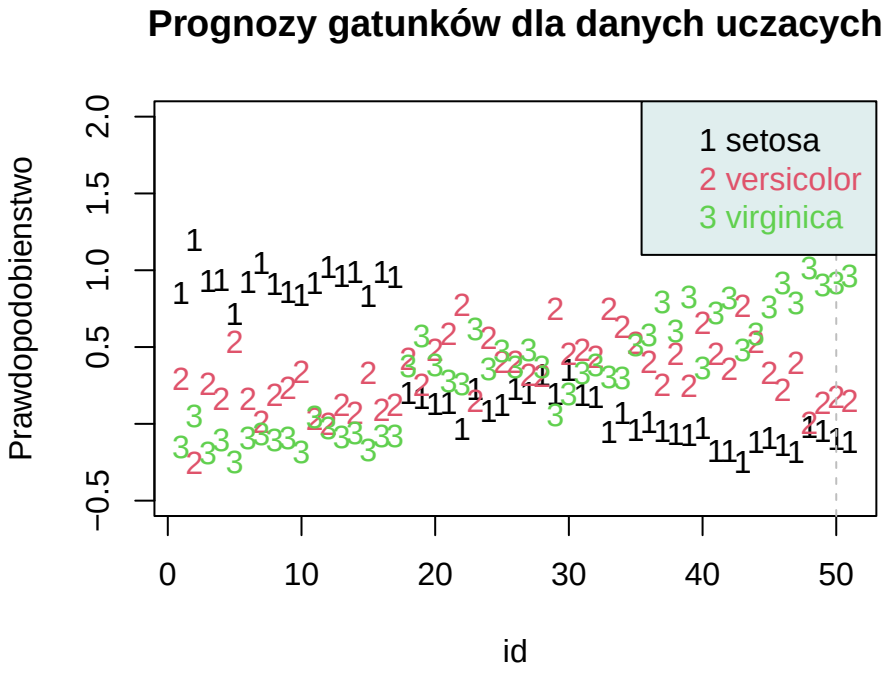
```
# Macierz estymowanych współczynników
B.hat1 <- solve(t(X2)%*%X2) %*% t(X2) %*% Y2 # X2 i Y2 są dla danych uczących
```

- Na podstawie dopasowanego modelu możemy teraz wyznaczyć wartości prognozowane dla zbioru danych testowych oraz trenujących

Wyznaczone prawdopodobieństwa możemy przedstawić na wykresach



Widać, **wyraźny podział na przedziały**, w których prawdopodobieństwo przynależności do odpowiednio grup 1,2,3 jest największe



Dla prognozowanych gatunków w zbiorze treningowym obserwujemy **podobny rozkład, z wyjątkiem środkowej grupy**. W tym przypadku występuje **zjawisko maskowania — gatunek nr 2 jest częściowo przesłaniany przez gatunek nr 3**. Na podstawie wykresu trudno jednoznacznie ocenić, który z tych dwóch gatunków ma w tym obszarze większe prawdopodobieństwo.

1.4 d) Ocena jakości modelu

Tworzymy macierz pomyłek, wygenerowanych etykiet odpowiednio dla:

- **Danych trenujących**

	setosa	versicolor	virginica
setosa	33	0	0
versicolor	0	27	6
virginica	0	8	25

- **Danych uczących**

	setosa	versicolor	virginica
setosa	17	0	0
versicolor	0	12	5
virginica	0	3	14

Teraz liczymy **dokładność naszego modelu dla danych testowych**

Dokładność
0.8585859

Widzimy **dokładność na poziomie ok 87% dla danych trenujących**, jest to dokładność na dobrym poziomie

Następnie dla danych uczących

Dokładność
0.8431373

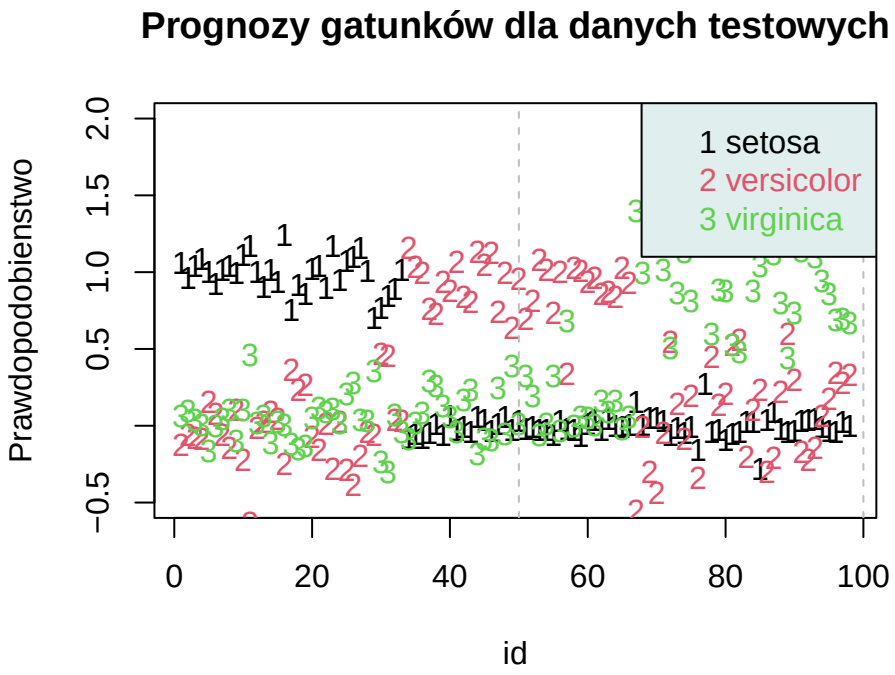
Warto zwrócić uwagę na zauważalny i istotny, a zarazem paradoksalny spadek dokładności dopasowania, mimo że etykiety przypisujemy do danych treningowych, gdzie teoretycznie oczekiwalibyśmy zgodności na poziomie 100%. **Tymczasem uzyskana wartość wynosi jedynie 84%, co oznacza spadek o około 3%.**

1.5 e) Budowa modelu liniowego dla rozszerzonej przestrzeni cech

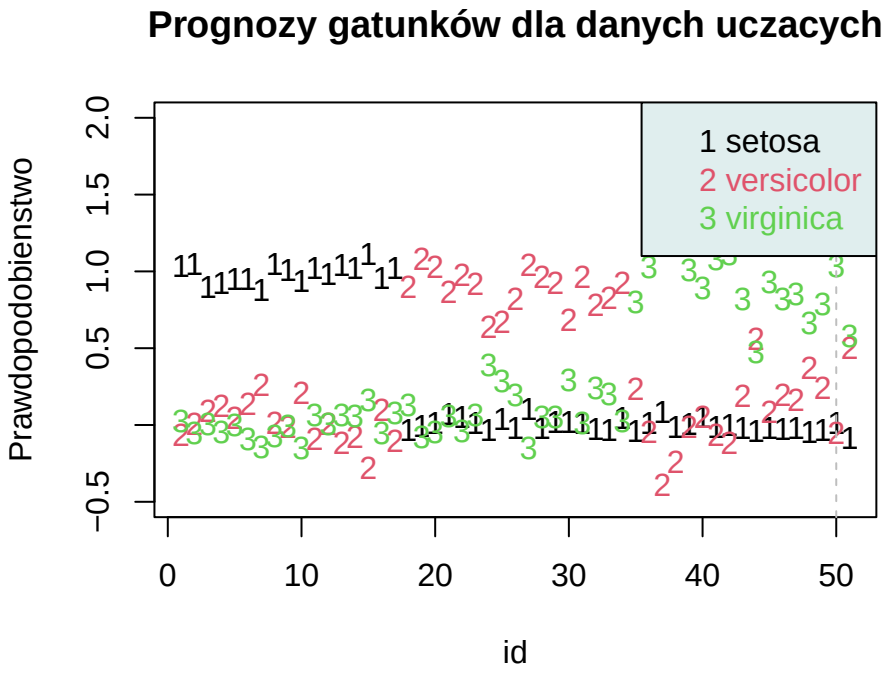
Teraz powtórzymy budowę modelu regresji, po uzupełnieniu cech o składniki wielomianowe stopnia 2 Dokładniej o $SL^2, SW^2, PL*PW, PL*SW, PL*SL, PW*SL, PW*SW, SL*SW$

Kroki b) oraz c) przebiegają analogicznie — tworzymy model regresji w ten sam sposób, z tą różnicą, że **teraz uwzględniamy dodatkowe cechy**

Wyznaczone w nowy sposób prawdopodobieństwa przypisań możemy ponownie przedstawić za pomocą wykresów.



Widać jeszcze wyraźniejszy podział na grupy niż w przypadku predykcji bez dodatkowych cech. Szczególnie dobrze widać dla których przedziałów dominują prawdopodobieństwa poszczególnych grup.



Dla danych uczących obserwujemy podobne cechy jak na poprzednim wykresie — nie dostrzegamy żadnych oznak maskowania. Wyraźnie zaznaczają się przedziały największych prawdopodobieństw dla poszczególnych grup. Całość cechuje się znacznie większą przejrzystością niż w przypadku regresji liniowej bez dodatkowych cech.

1.5.0.1 Ocena jakości nowego modelu

Tworzymy macierz pomyłek, wygenerowanych etykiet odpowiednio dla:

*Danych trenujących

	setosa	versicolor	virginica
setosa	33	0	0
versicolor	0	32	1
virginica	0	4	28

- Danych uczących

	setosa	versicolor	virginica
setosa	17	0	0
versicolor	0	17	0
virginica	0	1	16

	setosa	versicolor	virginica

I liczymy dokładność naszego modelu dla danych testowych

Dokładność
0.9489796

W przypadku danych treningowych obserwujemy dokładność na poziomie około 95%, co stanowi znakomity wynik — to aż **o 8 punktów procentowych więcej niż w przypadku standardowego modelu regresji liniowej.**

Następnie dla danych uczących

Dokładność
0.9803922

W nowym modelu obserwujemy wzrost dokładności dla danych treningowych — **osiąga ona bardzo wysoki poziom 98%**. To o 1 punkt procentowy więcej niż dla danych testowych oraz aż o 13 punktów procentowych więcej w porównaniu do modelu bez dodatkowych cech.

1.6 Wnioski

Dodanie dodatkowych cech znacząco poprawia dokładność modelu regresji liniowej, jednocześnie zmniejszając jego podatność na efekt maskowania między klasami.

2 Zadanie 2

2.1 a) Wybór i zapoznanie się z danymi

Opis zmiennych w zbiorze danych **Wine**

Kolumna	Nazwa zmiennej	Opis
V1	Alcohol	Zawartość alkoholu (%)
V2	Malic acid	Zawartość kwasu jabłkowego (g/l)
V3	Ash	Zawartość popiołu (g/l)
V4	Alcalinity of ash	Zasadowość popiołu (g/l)
V5	Magnesium	Zawartość magnezu (mg/l)
V6	Total phenols	Zawartość fenoli ogółem (g/l)
V7	Flavanoids	Zawartość flawonoidów (g/l)

Kolumna	Nazwa zmiennej	Opis
V8	Nonflavanoid phenols	Zawartość fenoli nienależących do flawonoidów (g/l)
V9	Proanthocyanins	Zawartość proantocyjaninów (g/l)
V10	Color intensity	Intensywność koloru (od 0 do 13)
V11	Hue	Odcień barwy
V12	OD280/OD315 of diluted wines	Absorbancja przy długości fali 280 nm do 315 nm (rozcieńczone wino)
V13	Proline	Zawartość proliny (mg/l)

Pierwsze 10 rekordów zbioru danych.

class	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13
1	14.23	1.71	2.43	15.6	127	2.80	3.06	0.28	2.29	5.64	1.04	3.92	1065
1	13.20	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.40	1050
1	13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	2.81	5.68	1.03	3.17	1185
1	14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	2.18	7.80	0.86	3.45	1480
1	13.24	2.59	2.87	21.0	118	2.80	2.69	0.39	1.82	4.32	1.04	2.93	735
1	14.20	1.76	2.45	15.2	112	3.27	3.39	0.34	1.97	6.75	1.05	2.85	1450
1	14.39	1.87	2.45	14.6	96	2.50	2.52	0.30	1.98	5.25	1.02	3.58	1290
1	14.06	2.15	2.61	17.6	121	2.60	2.51	0.31	1.25	5.05	1.06	3.58	1295
1	14.83	1.64	2.17	14.0	97	2.80	2.98	0.29	1.98	5.20	1.08	2.85	1045
1	13.86	1.35	2.27	16.0	98	2.98	3.15	0.22	1.85	7.22	1.01	3.55	1045

Zbiór danych ma **178** przypadków i **14** zmiennych

Tabela 11: Liczba unikalnych wartości w każdej zmiennej

	Liczba unikalnych
class	3
V1	126
V2	133
V3	79
V4	63
V5	53
V6	97
V7	132
V8	39
V9	101
V10	132

	Liczba unikalnych
V11	78
V12	122
V13	121

Zmienna **class** pełni rolę etykiety klas, informując o przynależności każdego obiektu do jednej z trzech grup. Świadczy o tym zarówno jej nazwa, jak i liczba unikalnych wartości, które przyjmuje — są to trzy klasy: ***1***, ***2*** i ***3***.

Ilość danych oznaczonych jako **Na** w danych kolumnach.

```
## class    V1    V2    V3    V4    V5    V6    V7    V8    V9   V10   V11   V12
##     0     0     0     0     0     0     0     0     0     0     0     0     0
##   V13
##     0
```

Analizując dane, można zauważyć, że **zbiór nie zawiera żadnych braków** — ani oznaczonych jako **NA**, ani zapisanych w inny sposób. **Wszystkie obserwacje** wydają się być **poprawnie wprowadzone**.

Jeśli chodzi o **wartości nietypowe**, to w kolumnie **V10** znajduje się **jedna obserwacja** o wartości **9.899999**, która ma **aż sześć miejsc po przecinku**.

Dla porównania, **pozostałe wartości** w tej kolumnie mają **najwyżej dwie cyfry po przecinku**, co może **sugerować błąd w zapisie** tej konkretnej danej.

W kolumnie **V5** pojawia się również wartość **162**, która **znacząco odstaje od reszty obserwacji** i może być wynikiem **błędu pomiaru** lub **wprowadzenia danych**.

Typ danych jaki przyjmują wartości z danej kolumny

	Typ danych
class	integer
V1	numeric
V2	numeric
V3	numeric

	Typ danych
V4	numeric
V5	integer
V6	numeric
V7	numeric
V8	numeric
V9	numeric
V10	numeric
V11	numeric
V12	numeric
V13	integer

Widać, że **wszystkie zmienne mają prawidłowo przypisane typy danych**, z wyjątkiem **naszej etykiety klas** — kolumny **class**.

Obecnie ma ona typ **integer**, co jest zrozumiałe, ponieważ wartości to liczby całkowite **{1, 2, 3}**.

Jednak **dla poprawnej analizy** i właściwego traktowania tej zmiennej jako **zbioru kategorii**, powinna zostać **przekonwertowana na typ factor**.

2.2 b) Wstępna analiza danych

2.2.1 Rozkład klas w zbiorze

Ilość rekordów przypisanych do odpowiedniej klasy.

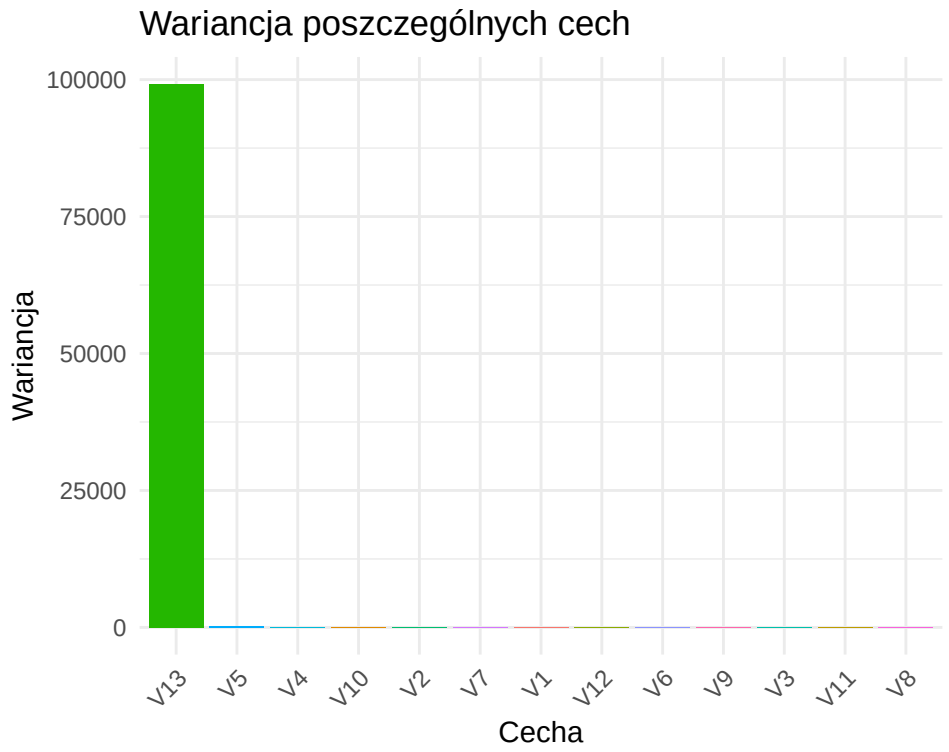
Klasa	Ilość
1	59
2	71
3	48

W zbiorze danych widoczne są wyraźne dysproporcje w liczebności poszczególnych klas. Najliczniejsza jest klasa 2, która stanowi około 40% wszystkich obserwacji. Następnie plasuje się klasa 1 z udziałem na poziomie około 33%. Najmniej reprezentowana jest klasa 3, obejmująca pozostałe przypadki.

Jeżeli przypisalibyśmy wszystkie obiekty do jednej dominującej klasy **(klasa 2)**, to aż 107 win byłoby przyporządkowanych źle. Wszystkich trunków mamy 178, więc mielibyśmy błąd klasyfikacji na poziome $\frac{107}{178} \approx 60\%$

2.2.2 Wariancje poszczególnych cech

Dla danych możemy zwizualizować wariancje poszczególnych cech na wykresie słupkowym:



Wariancje (posortowane malejąco)

```
##       V13        V5        V4       V10        V2        V7        V1       V12
## 99166.72    203.99     11.15      5.37      1.25      1.00      0.66      0.50
##        V6        V9        V3       V11        V8
##      0.39      0.33      0.08      0.05      0.02
```

Widać bardzo istotne różnice w zmienności poszczególnych grup, wariancja cechy V13 stanowi główny element wykresu. Jeżeli przyjrzymy się im dokładniej, to okaże się, że wariancja V13 jest większa 10^4 raz od drugiej co do wielkosci wariancji **(Zmiennej V5)**

	cecha	wariancja
V1	V1	6.590623e-01
V2	V2	1.248015e+00
V3	V3	7.526460e-02
V4	V4	1.115269e+01
V5	V5	2.039893e+02
V6	V6	3.916895e-01

	cecha	wariancja
V7	V7	9.977187e-01
V8	V8	1.548860e-02
V9	V9	3.275947e-01
V10	V10	5.374449e+00
V11	V11	5.224500e-02
V12	V12	5.040864e-01
V13	V13	9.916672e+04

Konieczność standaryzacji w niektórych algorytmach wynika z tego, że cechy danych mają różne zakresy wartości. Bez standaryzacji cechy o większych skalach mogą zdominować obliczenia, co prowadzi do nieoptymalnych wyników

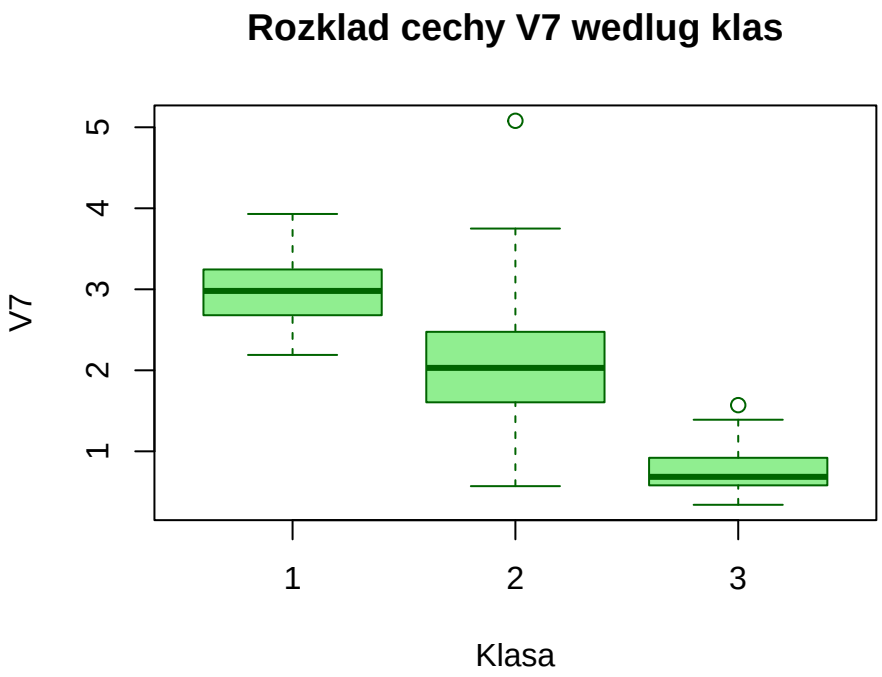
2.2.3 Cechy o najlepszej zdolności dyskryminacyjnej

Teraz obliczamy wartość dyskryminacyjne testem anova.

Tabela 15: Wartości p z testu ANOVA dla każdej cechy

	Cecha	p.wartość
V7	V7	3.599e-50
V13	V13	5.783e-47
V12	V12	1.393e-44
V1	V1	3.320e-36
V10	V10	1.162e-33
V11	V11	5.918e-30
V6	V6	2.138e-28
V2	V2	4.127e-14
V4	V4	9.444e-14
V9	V9	5.125e-12
V8	V8	3.888e-11
V3	V3	4.150e-06
V5	V5	8.963e-06

Na podstawie testu Anova, widać że zmienna **V7 ma najlepszą zdolność dyskryminacyjną.**



2.3 c) Ocena dokładności klasyfikacji

2.3.1 Pojedynczy podział na zbiór uczący i testowy

Dla porównania metod klasyfikacyjnych podzielono dane na zbiór uczący/treningowy (70%) czyli 126 rekordów i testowy (30%) czyli 52 rekordy. Oceniono skuteczność klasyfikatorów na zbiorze testowym oraz uczącym.

2.3.2 Metoda k-najbliższych sąsiadów

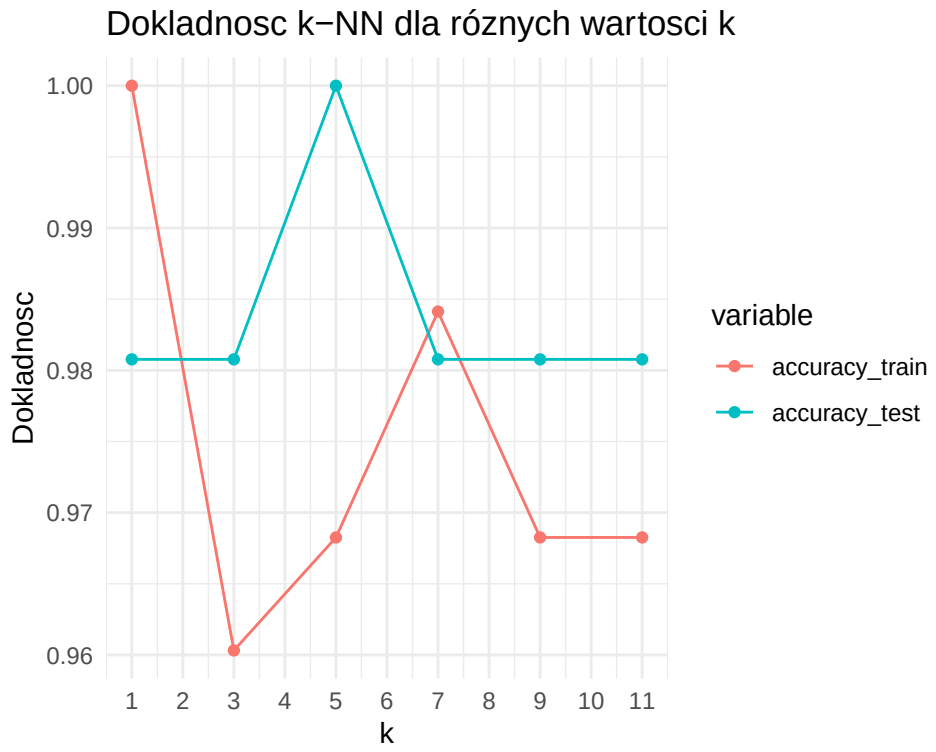
Tabela 16: Wyniki k-NN dla różnych wartości k

k	accuracy_train	accuracy_test
1	1.0000	0.9808
3	0.9603	0.9808
5	0.9683	1.0000
7	0.9841	0.9808
9	0.9683	0.9808
11	0.9683	0.9808

Tabela 17: Macierz pomyłek – metoda KNN dla k = 5

	Przewidziana: 1	Przewidziana: 2	Przewidziana: 3
Prawdziwa: 1	17	0	0
Prawdziwa: 2	0	21	0
Prawdziwa: 3	0	0	14

Na przykład dokładność metody kNN dla k=5 wynosi *1* a błąd klasyfikacji wynosi *0*.



2.3.3 Metoda drzewa klasyfikacyjnego

Tabela 18: Wyniki drzew dla różnych wartości cp

cp	accuracy_train	accuracy_test
0.001	0.9603	0.9808
0.010	0.9603	0.9808
0.050	0.9603	0.9808
0.100	0.9206	0.9038

Tabela 19: Macierz pomyłek – metoda drzewa klasyfikacyjnego dla cp=0.01

	Przewidziana: 1	Przewidziana: 2	Przewidziana: 3
Prawdziwa: 1	16	0	0
Prawdziwa: 2	1	21	0
Prawdziwa: 3	0	0	14

Na przykład dokładność metody drzewa klasyfikacyjnego dla cp=0.01 wynosi *0.9808* a błąd klasyfikacji wynosi *0.0192*.

2.3.4 Metoda naiwnego Bayes'a

Tabela 20: Macierz pomyłek – metoda naiwnego Bayes'a

	Przewidziana: 1	Przewidziana: 2	Przewidziana: 3
Prawdziwa: 1	17	0	0
Prawdziwa: 2	0	21	0
Prawdziwa: 3	0	0	14

Tabela 21: Macierz pomyłek – metoda naiwnego Bayes'a na ustandaryzowanych danych

	Przewidziana: 1	Przewidziana: 2	Przewidziana: 3
Prawdziwa: 1	17	0	0
Prawdziwa: 2	0	21	0
Prawdziwa: 3	0	0	14

Jak widać metoda naiwnego bayesa ma 100% dokładność na naszych danych testowych.

2.4 d) Różne parametry i różne podzbiory cech

Dokładność podczas testowania ze wszystkimi cechami i z tylko top5 cechami dyskryminującymi

Tabela 22: Porównanie wyników dla różnych zbiorów cech

Metoda	Wszystkie_cechy	Top5_cech
k-NN	1.0000	0.9808
Drzewo	0.9808	0.9231
Naiwny Bayes	1.0000	1.0000

Cross-Validation

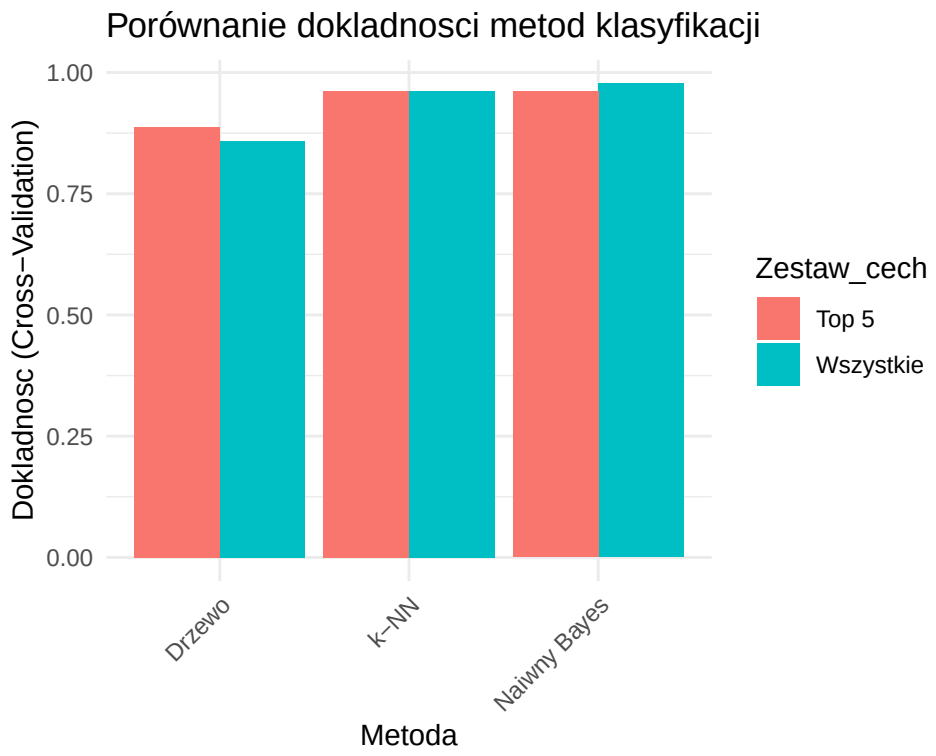
Cross-validation dla wszystkich cech:

Metoda	Srednia	Odchylenie_std
knn_k3	0.9556	0.0683
knn_k5	0.9611	0.0457
knn_k7	0.9608	0.0459
tree_cp001	0.8595	0.0397
tree_cp01	0.8595	0.0397
nb	0.9775	0.0291

Cross-validation dla top 5 cech:

Metoda	Srednia	Odchylenie_std
knn_k3	0.9722	0.0293
knn_k5	0.9611	0.0375
knn_k7	0.9611	0.0527
tree_cp001	0.8873	0.0536
tree_cp01	0.8873	0.0536
nb	0.9608	0.0377

Podsumowanie najlepszych wyników



GŁÓWNE WNIOSKI

1. Najlepszą metodą okazał się Naiwny Bayes.
2. Selekcja cech (top 5) wpłynęła na wyniki pozytywnie
3. Różnica między zbiorem treningowym a testowym wskazuje na dobra generalizację
4. Cross-validation dało bardziej wiarygodne wyniki niż pojedynczy podział
5. Wszystkie metody osiągnęły wysoką dokładność (>90%) na tym zbiorze danych

2.5 e) Wnioski końcowe

2.5.1 Najlepsze podzbiory zmiennych i parametry

2.5.1.1 A) Podzbiory zmiennych

- **Wszystkie cechy (13 zmiennych)**: Średnia dokładność CV = **0.929**
- **Top 5 cech dyskryminacyjnych**: Średnia dokładność CV = **0.9383**

Wniosek: Selekcja cech (**top 5**) poprawiła wyniki o **0.93 punktu procentowego**.

Najlepsze cechy dyskryminacyjne:

1. **V7**
2. **V12**
3. **V13**
4. **V11**
5. **V1**

2.5.1.2 B) Optymalne parametry

- **k-NN**: Najlepsze k = **5** (dokładność testowa: **1**)
- **Drzewo klasyfikacyjne**: Najlepsze cp = **0.001** (dokładność testowa: **0.9808**)
- **Naiwny Bayes**: Standaryzacja **nie wpłynęła na wyniki** (**100% w obu przypadkach**)

2.5.2 Ranking metod klasyfikacyjnych

Ranking metod (według cross-validation na wszystkich cechach):

1. **Naiwny Bayes** – dokładność: **0.9775 ± 0.0291**
2. **k-NN (k=5)** – dokładność: **0.9611 ± 0.0457**
3. **k-NN (k=7)** – dokładność: **0.9608 ± 0.0459**
4. **k-NN (k=3)** – dokładność: **0.9556 ± 0.0683**
5. **Drzewo (cp=0.001)** – dokładność: **0.8595 ± 0.0397**
6. **Drzewo (cp=0.01)** – dokładność: **0.8595 ± 0.0397**

> **Najlepsza metoda**: **Naiwny Bayes Najgorsza metoda**: **Drzewo (cp=0.01) Różnica**: **11.8 punktu procentowego**

Stabilność metod (odchylenie standardowe w CV):

- **Najbardziej stabilna**: **Naiwny Bayes** *(std = 0.0291)*
- **Najmniej stabilna**: **k-NN (k=3)** *(std = 0.0683)*

2.5.3 Wpływ schematu oceny na wnioski

2.5.3.1 A) Porównanie pojedynczy podział vs cross-validation

Metoda	Pojedynczy podział	Cross-validation	Różnica
k-NN (k=5)	**1.0000**	**0.9611**	**-0.0389**
Drzewo (cp=0.01)	**0.9808**	**0.8595**	**-0.1213**
Naiwny Bayes	**1.0000**	**0.9775**	**-0.0225**

2.5.3.2 B) Analiza overfittingu (zbiór treningowy vs testowy)

- **k-NN**: Różnica treningowy-testowy = **0%**
- **Drzewo**: Różnica treningowy-testowy = **-2.05%**
- **Naiwny Bayes**: **100% na obu zbiorach** → Brak overfittingu

---

2.5.4 Kluczowe wnioski końcowe

2.5.4.1 Najlepsza konfiguracja

- **Metoda**: **Naiwny Bayes**
- **Zestaw cech**: **Top 5 cech**
- **Osiągnięta dokładność**: **0.9775**

2.5.4.2 Hierarchia skuteczności metod

- **Naiwny Bayes → NAJSKUTECZNIEJSZY**
- **Metoda odporna na przekleństwo wymiarowości**
- **Doskonała wydajność na surowych i standaryzowanych danych**

2.5.4.3 Wpływ selekcji cech

- **Selekcja cech ZNACZĄCO poprawiła wyniki**
- **Najważniejsze cechy**: **V7, V12, V13**

2.5.4.4 Znaczenie schematu oceny

- **Cross-validation dało bardziej WIARYGODNE wyniki**
- **Pojedynczy podział może prowadzić do ZNACZĄCO RÓŻNYCH wniosków**
- **Standardowe odchylenia w CV wskazują na stabilność metod**

2.5.4.5 Charakterystyka zbioru danych Wine

- **Stosunkowo łatwy do klasyfikacji** *(wszystkie metody >90%)*
- **Cecha V7 (Flavanoids) ma najlepszą zdolność dyskryminacyjną**
- **Wymagana standaryzacja ze względu na różne skale** *(V13 ma wariancję* 10^4 *większą)*
