Sprawozdanie 3

Eksploracja danych

Kacper Szmigielski, 282255 i Mateusz Wizner

2025-06-03

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.

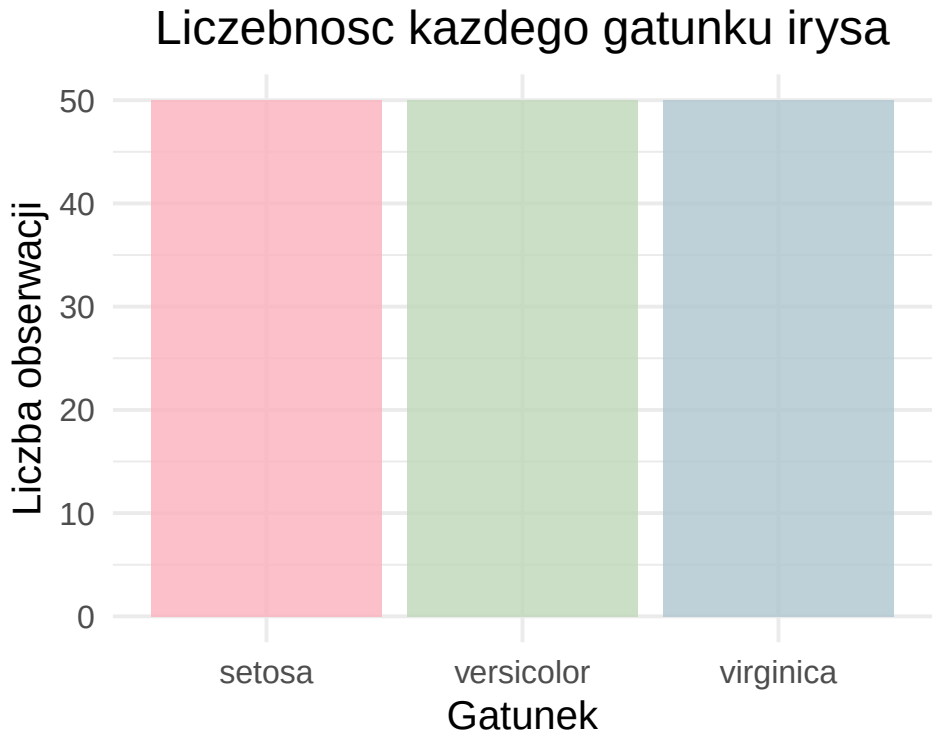
Każdy rekord opisuje pojedynczy kwiat za pomocą czterech cech numerycznych (p=4 cechy ilościowe) :

- 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

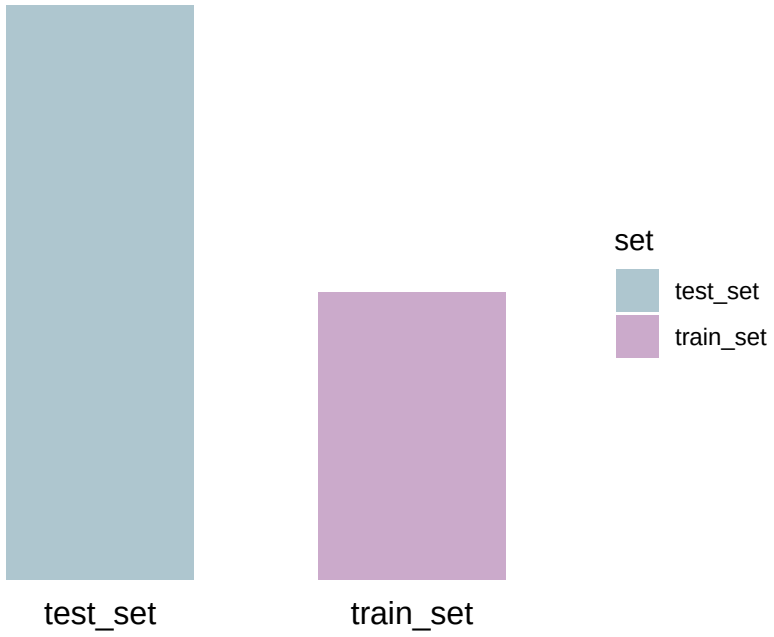


Mamy równy podział danych w zbiorze. Obserwacji każdego gatunku jest 50.

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.

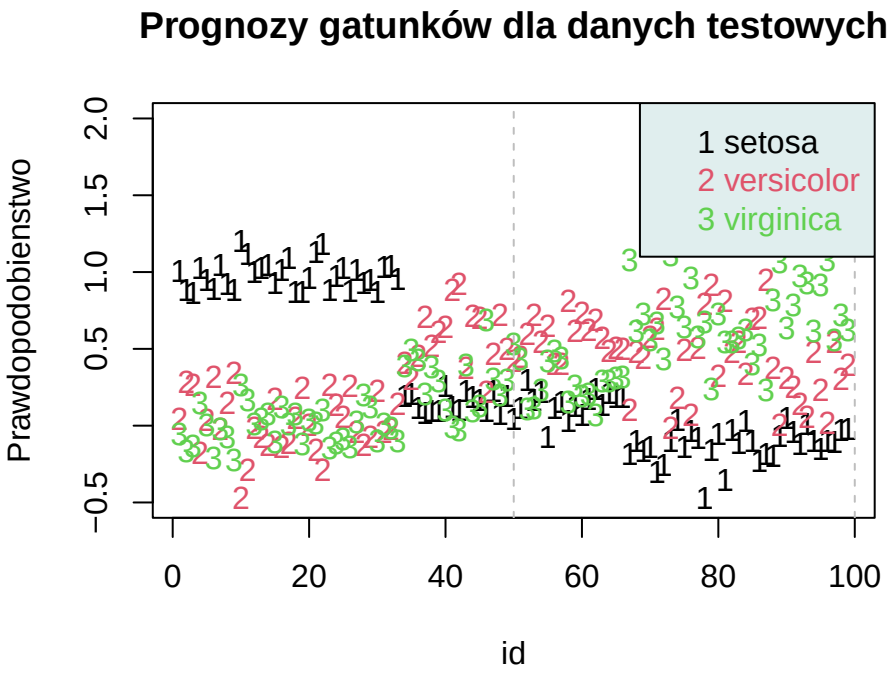
1.3.2 Estymacja współczynników i konstrukcja prognoz

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

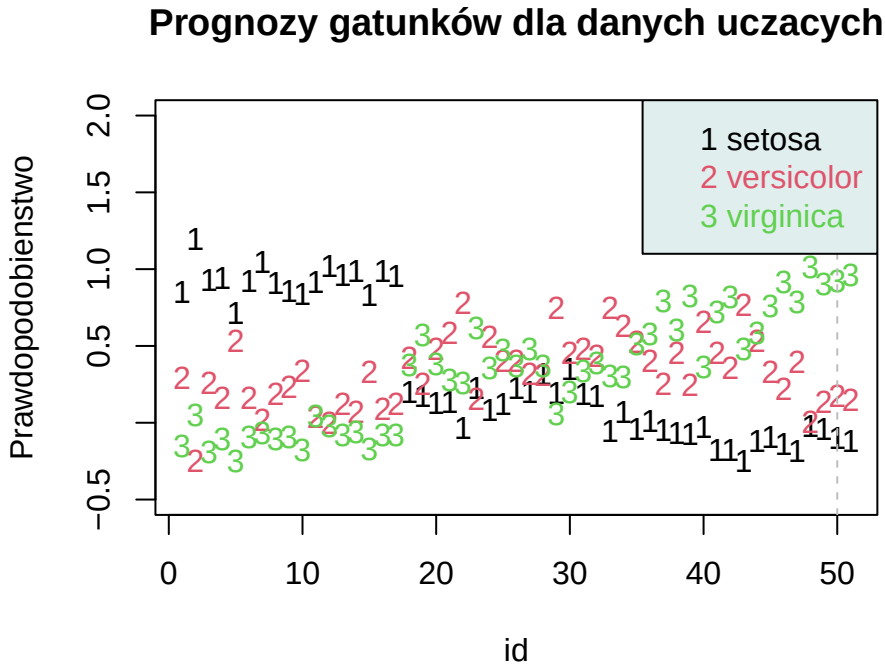
```
# Macierz estymowanych współczynników
B.hat1 <- solve(t(X2)%*%X2) %*% t(X2) %*% Y2 # X2 i Y2 są dla danych uczących
```

- 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

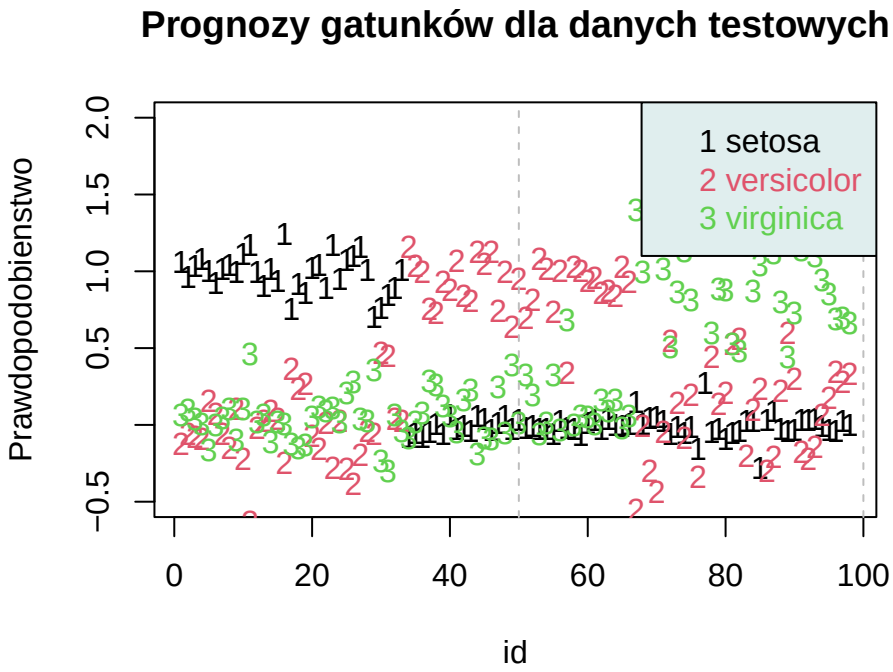
Warto zwrócić uwagę na zauważalny i istotny, a zarazem paradoksalny spadek dokładności dopasowania, mimo że etykiety przypisujemy do danych treningowych, gdzie teoretycznie oczekiwalibyśmy zgodności na poziomie 100%. **Tymczasem uzyskana wartość wynosi jedynie 84%, co oznacza spadek o około 3%.**

1.5 e) Budowa modelu liniowego dla rozszerzonej przestrzeni cech

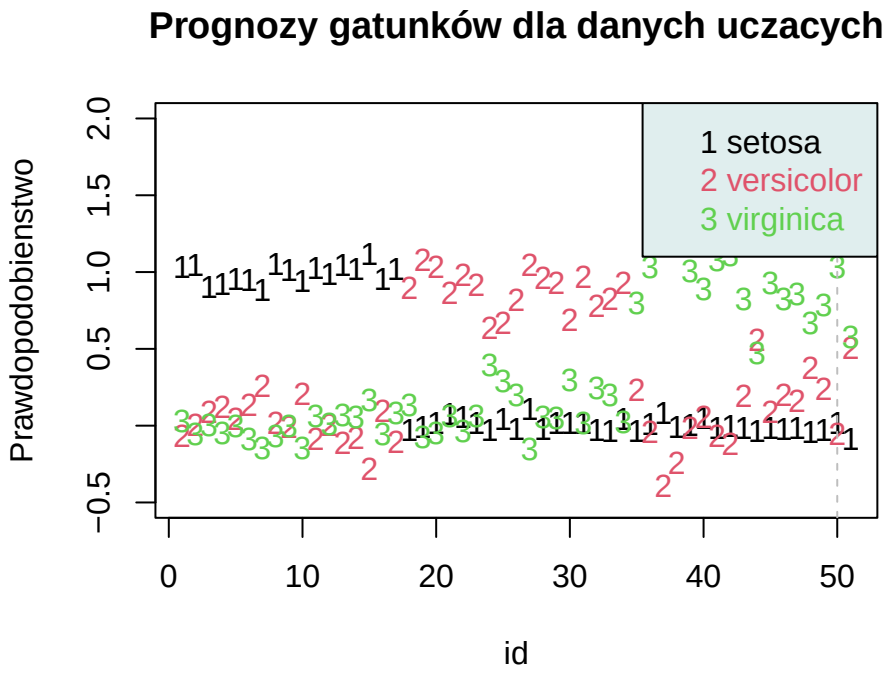
Teraz powtórzymy budowę modelu regresji, po uzupełnieniu cech o składniki wielomianowe stopnia 2 Dokładniej o $SL^2, SW^2, PL*PW, PL*SW, PL*SL, PW*SL, PW*SW, SL*SW$

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

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.