Sprawozdanie z listy 3

Eksploracja danych

Marta Stankiewicz (282244)
Paweł Nowak (282223)

2025-05-12

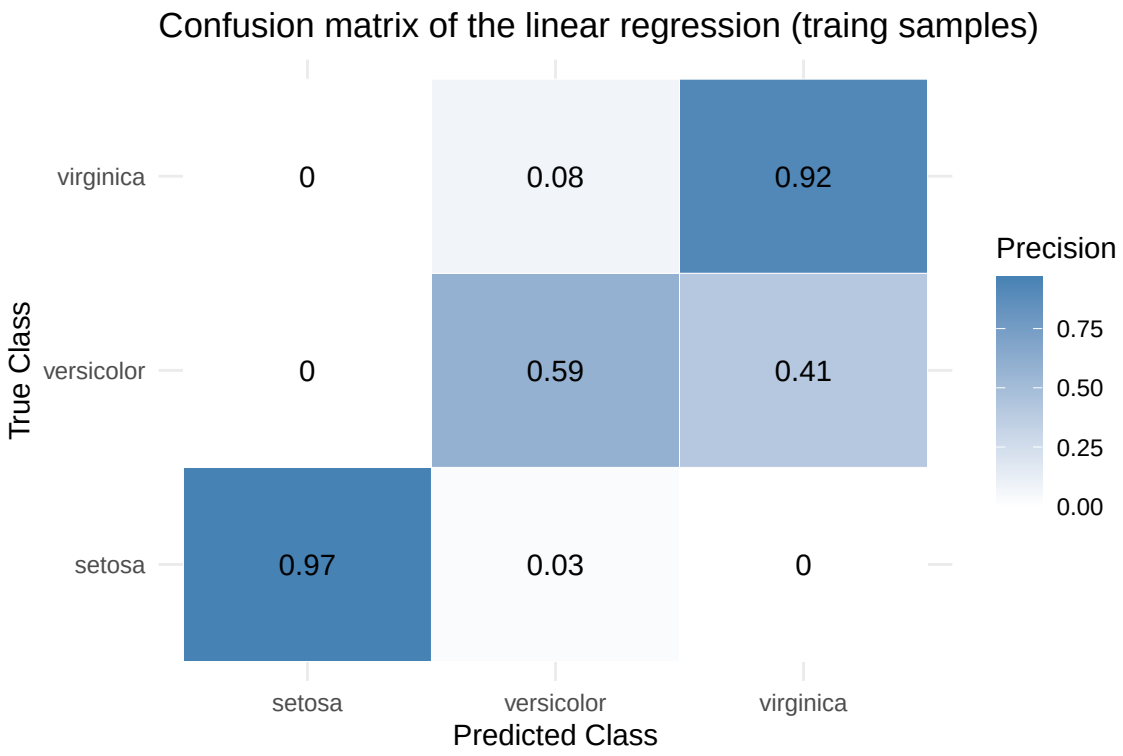
Spis treści

Spis rysunków

Spis tabel

1 Klasyfikacja na bazie modelu regresji liniowej

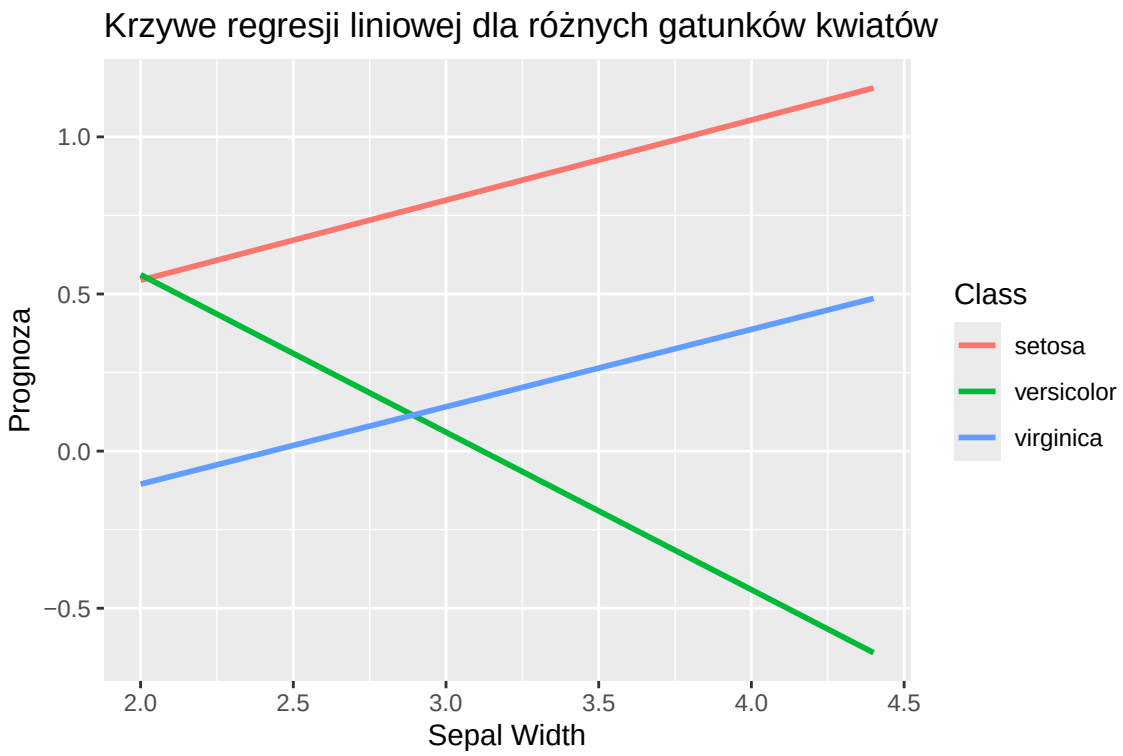
Aby ocenić skuteczność klasyfikatora opartego na modelu regresji liniowej, wykorzystamy zbiór danych iris, w którym zmienną objaśnianą jest Species, zawierająca 3 unikalnych klas. W celu uniknięcia problemu wycieku danych (ang. data leakage), przeprowadzimy podział oryginalnego zbioru na zbiór treningowy oraz zbiór testowy, przy czym zbiory te będą zawierały odpowiednio 70 obserwacji z danych iris. Po wytrenowaniu modelu na danych treningowych, przeprowadzimy ewaluację jego skuteczności na podstawie zbioru testowego. Wyniki klasyfikacji zaprezentujemy za pomocą znormalizowanej macierzy pomyłek, w której wartości w każdej kolumnie zostaną podzielone przez sumę elementów tej kolumny, co pozwoli na lepszą interpretację skuteczności klasyfikacji dla poszczególnych klas.



Rysunek 1: Macierz pomyłek regresji liniowej dla obserwacji treningowych

1.1 Analiza skuteczności klasyfikacji dla zbioru treningowego

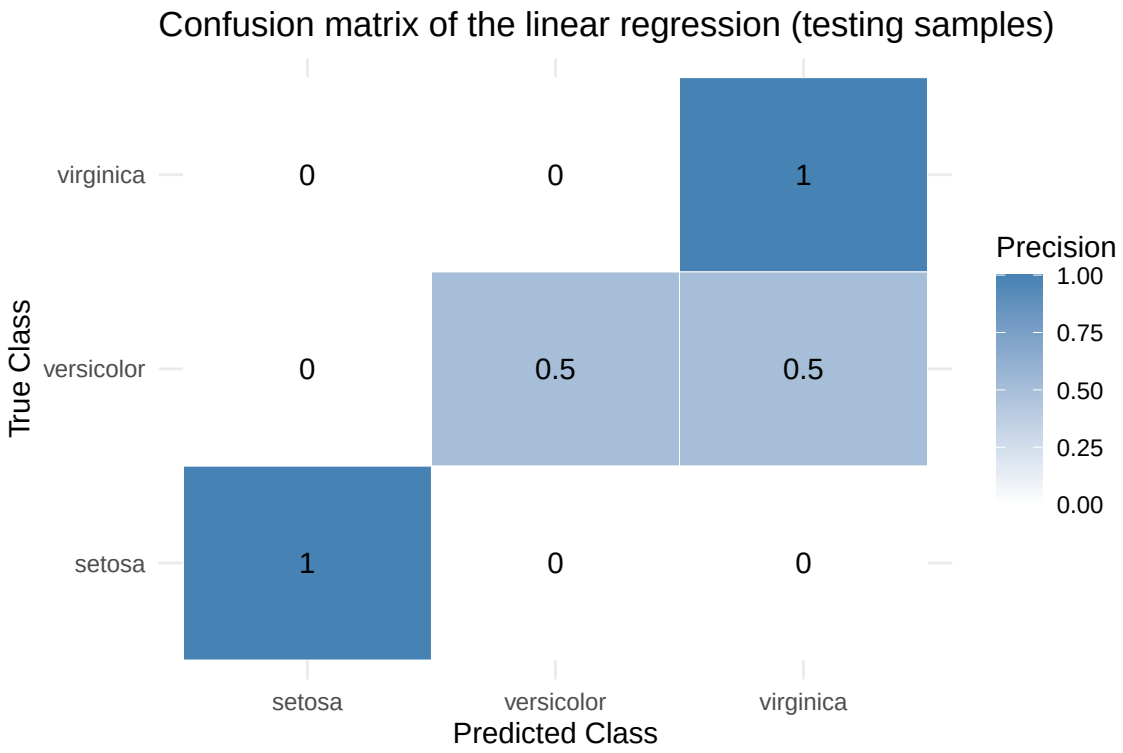
Na podstawie macierzy pomyłek przedstawionej na Rysunku **??**, przyjrzeliśmy się, jak dobrze klasyfikator oparty na regresji liniowej radzi sobie z przewidywaniem poszczególnych klas. Zauważyliśmy, że skuteczność tych przewidywań różni się w zależności od tego, do której klasy należą próbki treningowe. Najlepiej klasyfikator poradził sobie z klasą setosa oraz virginica - dla obu tych klas precyzja wyniosła ponad 92%. To pokazuje, że model bardzo dobrze rozpoznaje te dwie klasy w zbiorze treningowym i rzadko się myli, przypisując im inne próbki. Jednak w przypadku klasy versicolor skuteczność przewidywania wyraźnie spadła, osiągając tylko 71% precyzji. To sugeruje, że klasyfikator ma większy problem z prawidłowym rozpoznawaniem próbek należących do tej klasy. Możemy przypuszczać, że powodem gorszych wyników dla gatunku versicolor jest tak zwany problem maskowania klasy (class masking problem). Chodzi o to, że cechy charakterystyczne dla klasy versicolor mogą być podobne do cech innych klas, co utrudnia modelowi regresji liniowej jednoznaczne przypisanie próbek do właściwej kategorii. Aby sprawdzić, czy tak jest, przeanalizujemy teraz kolejny wykres.



Rysunek 2: Krzywe regresji liniowej dla różnych gatunków kwiatów

Analiza przedstawionych na rysunku **??** prostych regresji liniowych ujawnia problem maskowania klas w odniesieniu do kategorii 'Versicolor'. W obszarze niskich wartości predyktora Sepal.Width, charakteryzujących się największym prawdopodobieństwem obserwacji, krzywa regresji odpowiadająca gatunkowi Iris versicolor przebiega pomiędzy krzywymi pozostałych klas. Taka konfiguracja przestrzenna implikuje, iż w zakresie wspomnianych wartości predyktora, klasyfikator oparty na bezpośrednim porównaniu wartości regresji liniowej może systematycznie pomijać przynależność obserwacji do klasy 'Versicolor', prowadząc do potencjalnych błędów klasyfikacji.

1.2 Analiza skuteczności klasyfikacji dla zbioru testowego



Rysunek 3: Macierz pomyłek regresji liniowej dla obserwacji testowych

Podobne wnioski można wyciągnąć z oceny skuteczności modelu regresji na zbiorze testowym, co ilustruje macierz pomyłek na rysunku **??**. Klasa 'versicolor' wykazuje relatywnie wysoką częstotliwość błędnych klasyfikacji, co jest prawdopodobnie konsekwencją wspomnianego problemu maskowania klas.

2 Klasyfikacja na bazie modelu regresji liniowej z czynnikami wielomianowymi

Trudności związane z maskowaniem klas znacząco utrudniają stworzenie efektywnego klasyfikatora opartego na modelu regresji liniowej. W celu zminimalizowania tego problemu i poprawy jakości klasyfikacji, wykorzystamy predyktory do wygenerowania czynników wielomianowych, czyli wyrażeń w formie: ...

$$X_1^{t_1} X_2^{t_2} \dots X_p^{t_p},$$
$$\text{gdzie } \sum_{i=1}^{p} t_i = 2 \quad \text{oraz} \quad \forall i \in \{1, \dots, p\} \quad t_i \geq 0$$