Sprawozdanie z listy 3

Eksploracja danych

Marta Stankiewicz (282244)
Paweł Nowak (282223)

2025-05-20

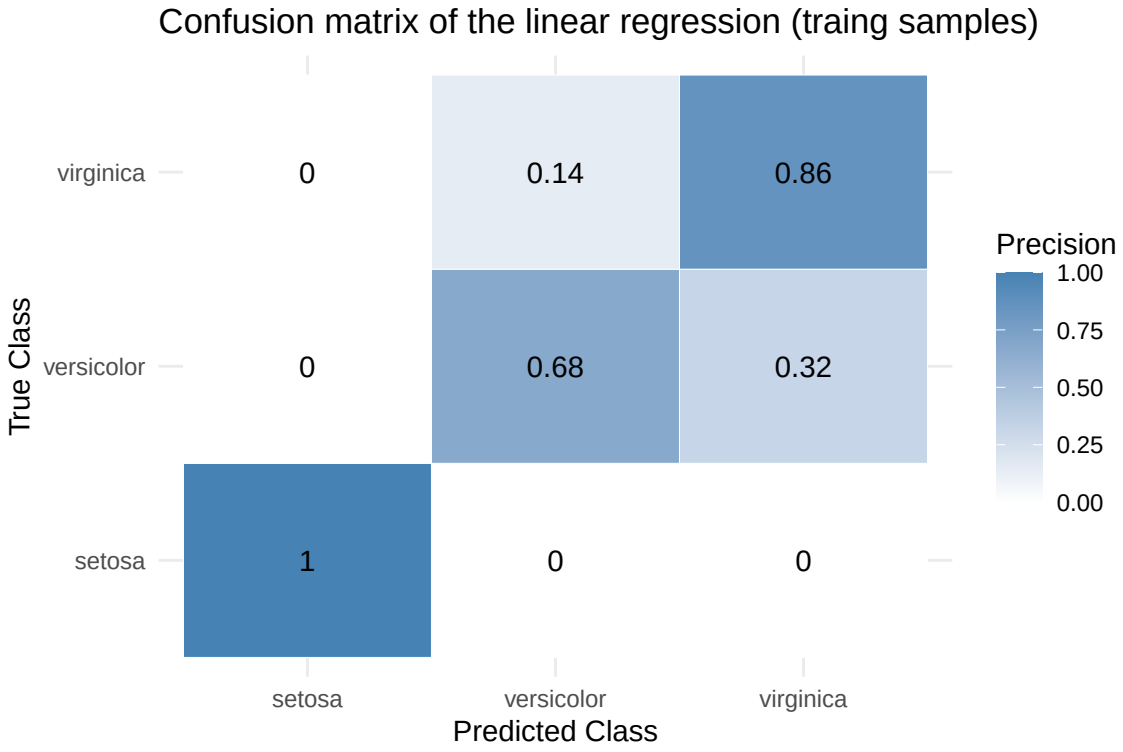
Spis treści

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.

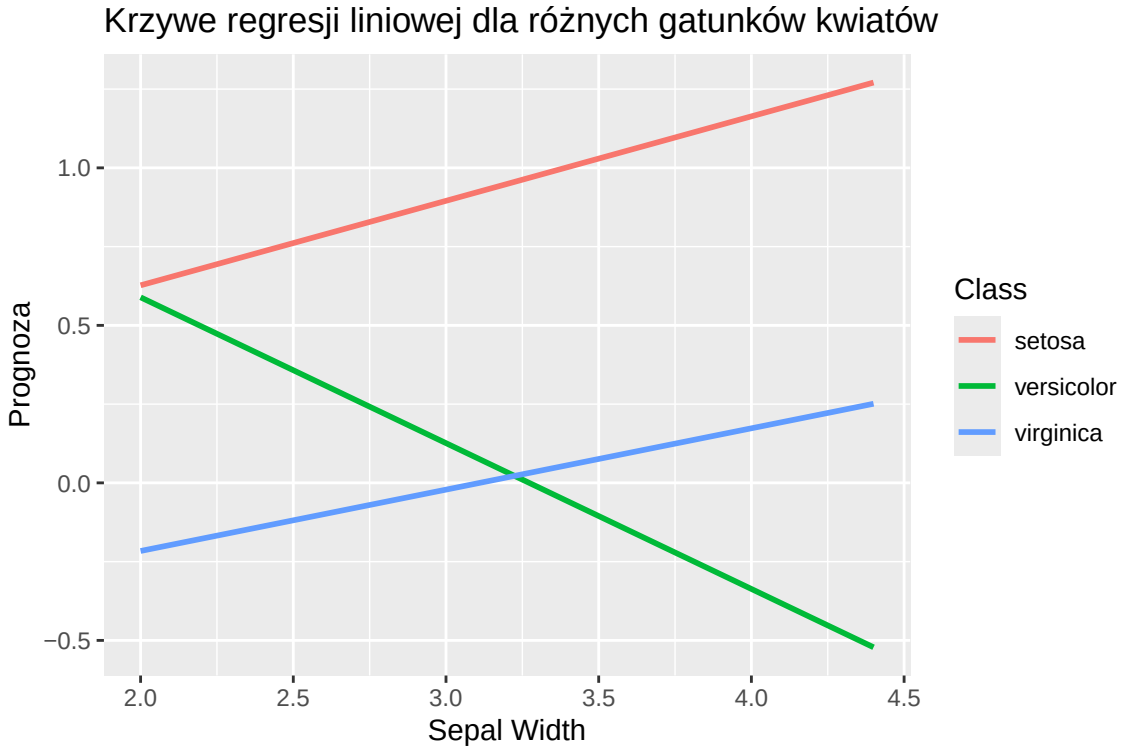


Rysunek 1: Macierz pomyłek regresji liniowej dla obserwacji treningowych

1.1 Analiza skuteczności klasyfikacji dla zbioru treningowego

Na podstawie macierzy pomyłek przedstawionej na Rysunku 1, 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

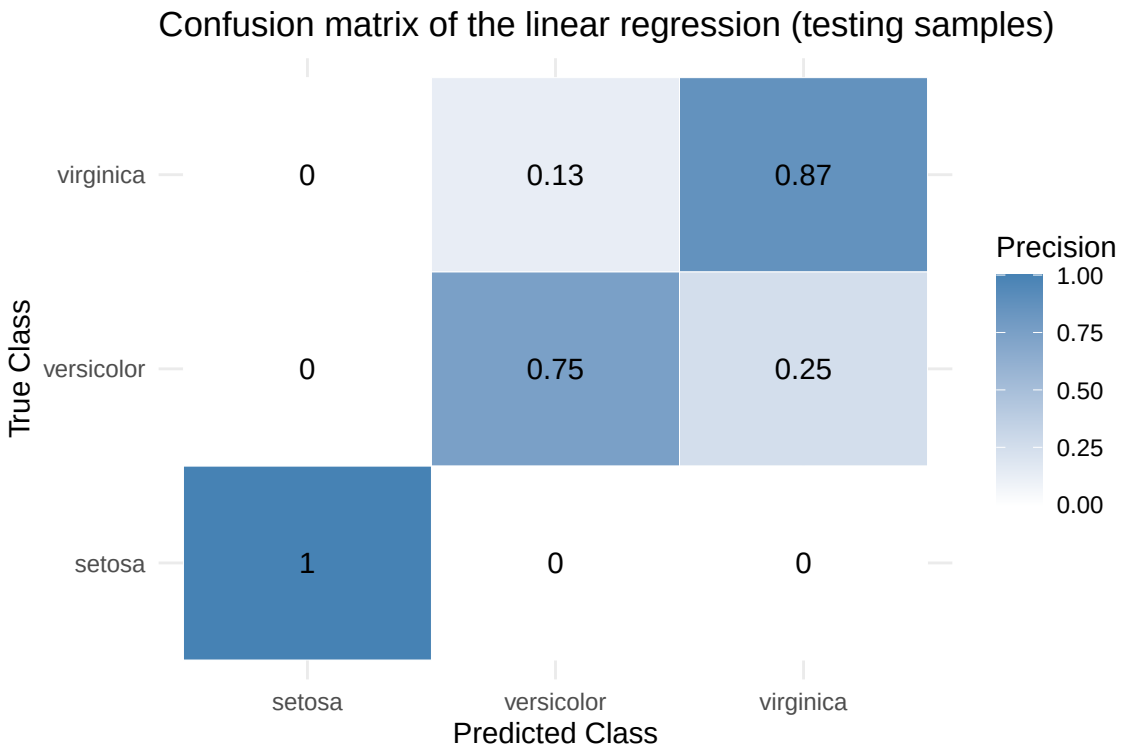
dobrze rozpoznaje te dwie klasy w zbiorze treningowym i rzadko się myli, przypisując im inne próbki. Jednak w przypadku klasy versicolor skuteczność przewidywania wyraźnie spadła, osiągając tylko 71% precyzji. To sugeruje, że klasyfikator ma większy problem z prawidłowym rozpoznawaniem próbek należących 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 2 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



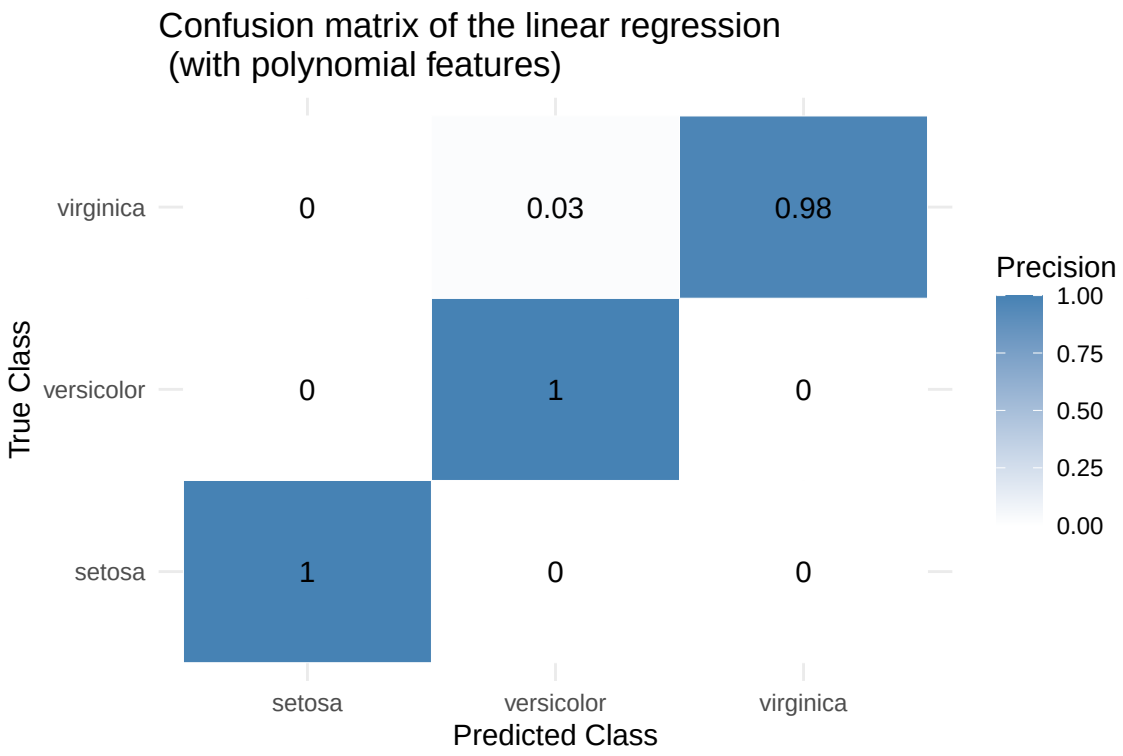
Rysunek 3: Macierz pomyłek regresji liniowej dla obserwacji testowych

Podobne wnioski można wyciągnąć z oceny skuteczności modelu regresji na zbiorze testowym, co ilustruje macierz pomyłek na rysunku 3. Klasa 'versicolor' wykazuje relatywnie wysoką częstotliwość błędnych klasyfikacji, co jest prawdopodobnie konsekwencją wspomnianego problemu maskowania klas.

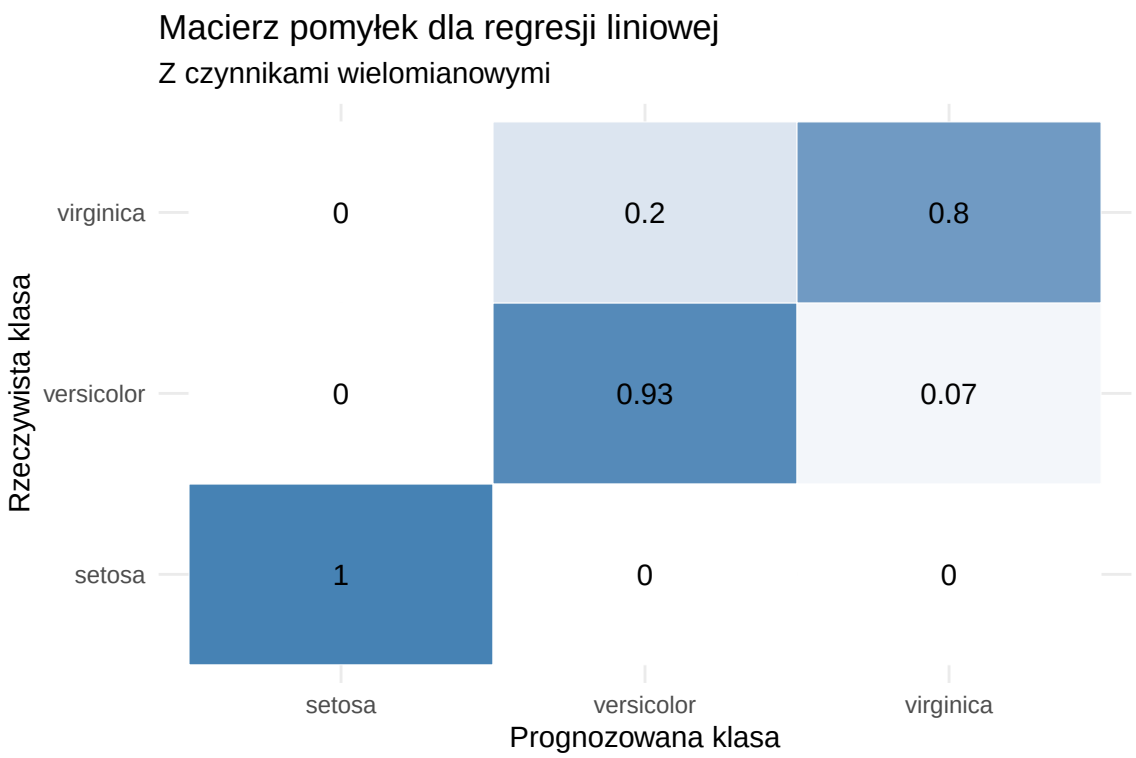
2 Klasyfikacja na bazie modelu regresji liniowej z czynnikami wielomianowymi

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: ...

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



Rysunek 4: Macierz pomyłek regresji liniowej z cechami wielomianowymi



Rysunek 5: Macierz pomyłek regresji liniowej z cechami wielomianowymi

2.1 Wnioski

Analiza map cieplnych macierzy pomyłek wykazała znaczącą poprawę skuteczności klasyfikacji po zastosowaniu modelu regresji liniowej z uwzględnieniem cech wielomianowych. W odróżnieniu od modelu bazowego, w którym zaobserwowano problem maskowania klasy oraz niską skuteczność klasyfikacji próbek należących do klasy ‘versicolor’, rozszerzony model z cechami wielomianowymi charakteryzuje się niemal całkowitym wyeliminowaniem tych niekorzystnych zjawisk.

Wykresy macierzy pomyłek jednoznacznie wskazują, że wprowadzenie czynników wielomianowych przyczyniło się do lepszego rozdzielenia przestrzeni cech, co w konsekwencji umożliwiło modelowi regresji liniowej dokładniejsze przypisanie próbek do właściwych klas. Zanik “pomijania” klasy ‘versicolor’ oraz ogólnie wyższa koncentracja wartości na głównej diagonali macierzy pomyłek dla modelu z cechami wielomianowymi stanowią silne argumenty przemawiające za istotnością rozszerzenia zestawu cech o komponenty wielomianowe.

Na podstawie przeprowadzonych obserwacji można zatem wnioskować, że dodanie czynników wielomianowych do modelu regresji liniowej jest uzasadnione i korzystnie wpływa na zdolności klasyfikacyjne modelu w analizowanym problemie. Rozszerzenie przestrzeni cech o interakcje i potęgi oryginalnych cech dostarcza modelowi dodatkowych informacji, które pozwalają na tworzenie bardziej złożonych i dokładnych granic decyzyjnych między klasami.

3 Porównanie metod klasyfikacji

W tym rozdziale skupimy się na porównaniu wybranych metod klasyfikacyjnych: algorytmu K-najbliższych sąsiadów, drzewa decyzyjnego oraz klasyfikatora bayesowskiego. Reguły decyzyjne zostaną wytrenowane i przetestowane na zbiorze danych PimaIndiansDiabetes2, który jest dostępny w pakiecie mlbench w języku R. Przed przystąpieniem do budowy modeli podzielimy dane na zbiór treningowy i testowy, aby zapobiec zjawisku wycieku danych oraz zapewnić wiarygodną ocenę skuteczności klasyfikatorów.

Zbiór danych PimaIndiansDiabetes2 zawiera 768 przypadków oraz 9 zmiennych, z czego 8 stanowi potencjalne cechy predykcyjne, a jedna kolumna – diabetes – pełni rolę zmiennej objaśnianej. Określa ona, czy dana osoba – kobieta pochodząca z rdzennego plemienia Pima, zamieszkującego stan Arizona w USA – należy do grupy osób chorujących na cukrzycę typu 2. W celu uproszczenia analizy, poziomy zmiennej diabetes zostały zmienione z "pos" i "neg" na "1" i "0", przy czym zmienna zachowała typ czynnika (ang. factor).

Wstępna analiza opisowa zbioru danych wykazała, że brakujące wartości zostały oznaczone w sposób zgodny z powszechną konwencją, czyli przy użyciu symbolu NA. W danych nie występują nietypowe lub niepoprawne sposoby kodowania braków, takie jak wartość 0 w kolumnie insulin, co czasem spotyka się w gorzej przygotowanych zbiorach. Zmienna docelowa diabetes jest prawidłowo przechowywana jako czynnik (ang. factor), co umożliwia bezpośrednie jej wykorzystanie w zadaniach klasyfikacyjnych.

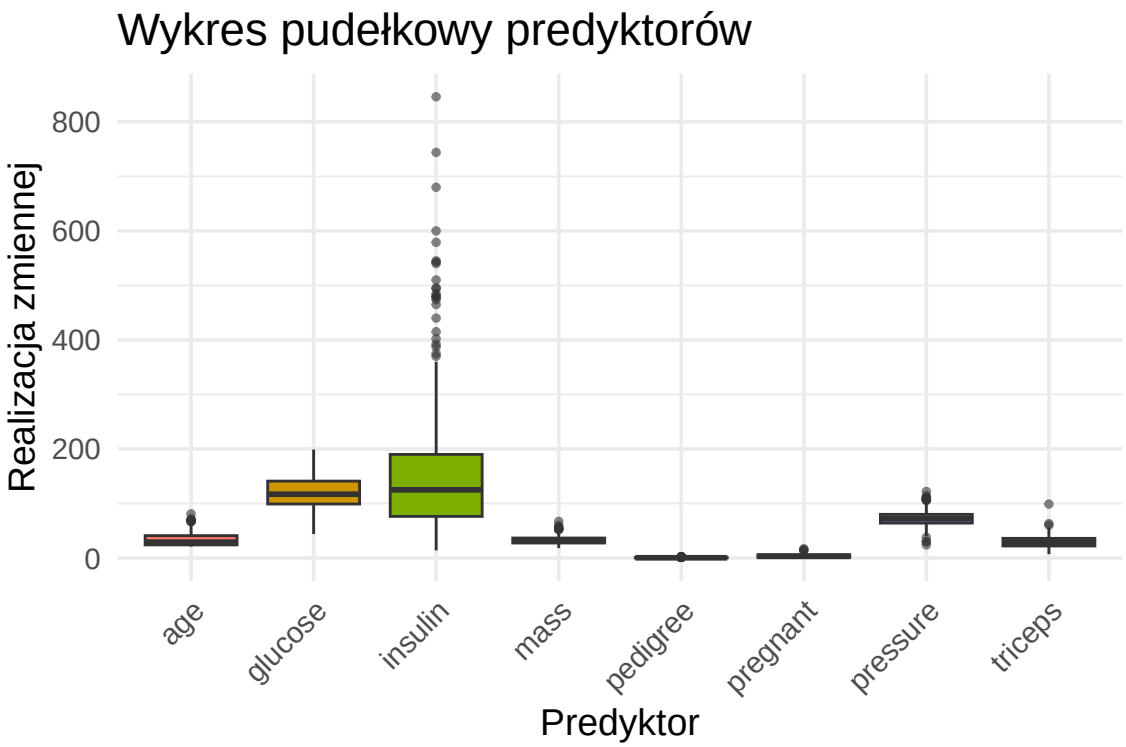
3.1 Wstępna analiza danych.

Na początku przeprowadzona zostanie analiza rozkładu klas zmiennej docelowej. Jest to ważny krok, gdyż nierównomierne rozłożenie klas (tzw. problem niezbalansowanych danych) może znacząco wpłynąć na ocenę skuteczności modeli.

Rysunek 6: Rozkład etykiet zmiennej celu diabetes

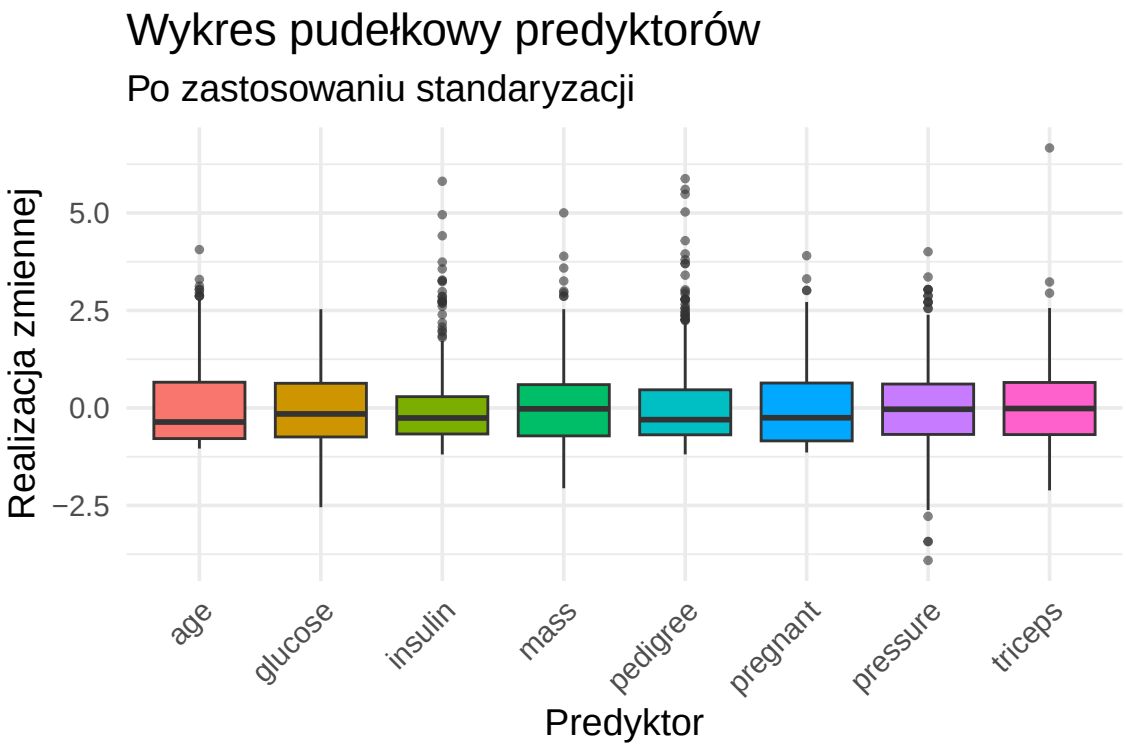
Z rysunku 6 wyraźnie wynika, że mamy do czynienia z problemem niezbalansowanych danych — liczba osób niechorujących na cukrzycę typu 2 jest niemal dwukrotnie większa niż liczba osób chorych. Teoretycznie, gdybyśmy zastosowali prosty, „naiwny" klasyfikator przypisujący każdą obserwację do klasy dominującej, uzyskalibyśmy wysoką ogólną skuteczność. Jednak po bliższej analizie metryk oceniających dokładność dla obu klas okazałoby się, że taki model w praktyce jest nieskuteczny, ponieważ całkowity błąd klasyfikacji dla klasy mniejszościowej wyniósłby 100%.

Przeanalizujemy teraz rozkłady ciągłych predyktorów. Jest to bardzo istotna kwestia, zwłaszcza w kontekście klasyfikatora KNN, gdzie różne skale pomiarowe różnych zmiennych mogą znacząco zanizać wpływ zmiennych charakteryzujących się zwięzłym zakresem wartości.



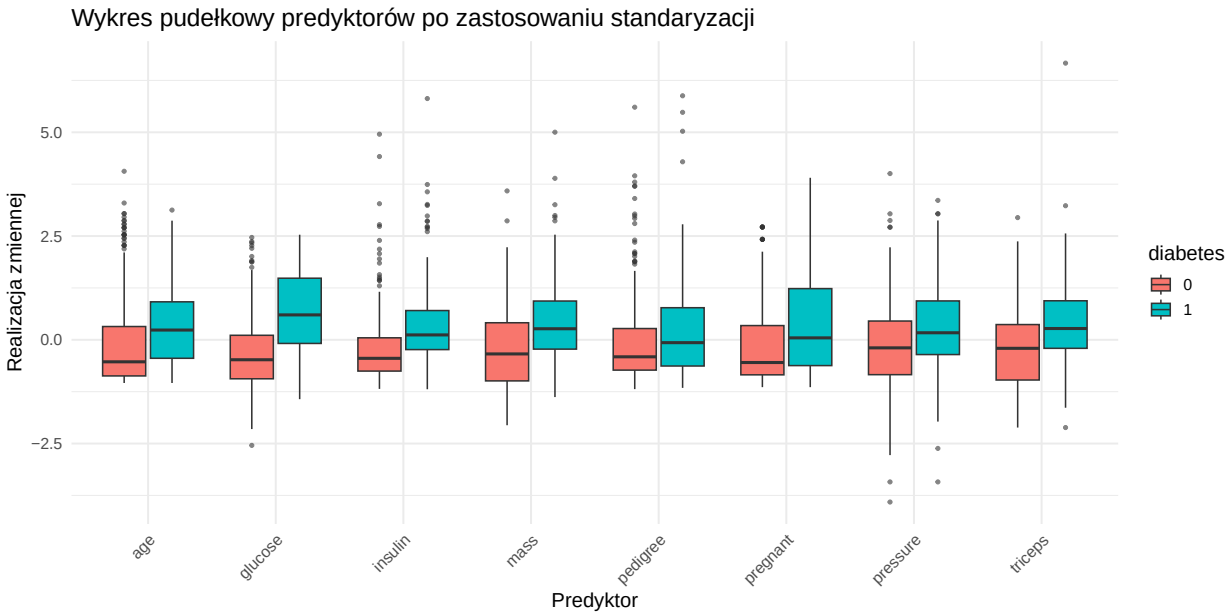
Rysunek 7: Porównanie wariancji predyktorów

Z rysunku 7 jasno wynika, że standaryzacja predyktorów jest niezbędna. Zmienne różnią się istotnie zarówno pod względem wariancji, jak i wartości tendencji centralnej, co szczególnie widoczne jest na przykładzie porównania wykresów pudełkowych zmiennych „insulin" oraz „triceps". Dokonajmy zatem standaryzacji i spójrzmy na rezultaty, które przedstawiono na rysunku 8



Rysunek 8: Porównanie wariancji predyktorów po zastosowaniu standaryzacji

Mając już ujednoliconą skalę pomiarową dla wszystkich predyktorów, możemy przejść do oceny ich zdolności dyskryminacyjnej. Ten etap pozwoli nam zidentyfikować zmienne najlepiej rozróżniające klasy i lepiej zrozumieć ich wpływ na działanie modelu.

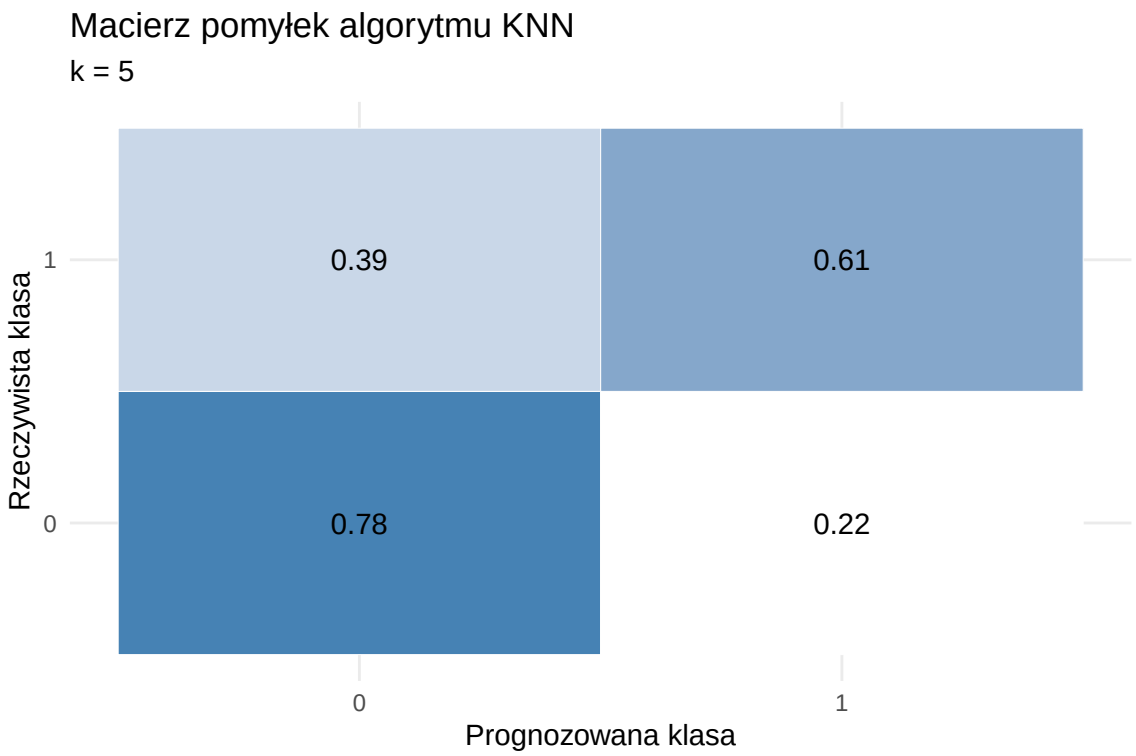


Rysunek 9: Porównanie zdolności dyskryminacyjnych predyktorów

Jak pokazano na rysunku 9, żadna ze zmiennych nie rozdziela klas docelowych w sposób idealny. Mimo to, można wyróżnić predyktory o wyraźnie silniejszych zdolnościach dyskryminacyjnych. Do takich atrybutów należą zmienne: ‘glucose’, ‘insulin’, ‘triceps’ oraz ‘mass’. Natomiast najsłabszą zdolność separacji klas wykazują zmienne ‘pedigree’ oraz ‘pressure’.

4 Ocena dokładności klasyfikacji i porównanie metod

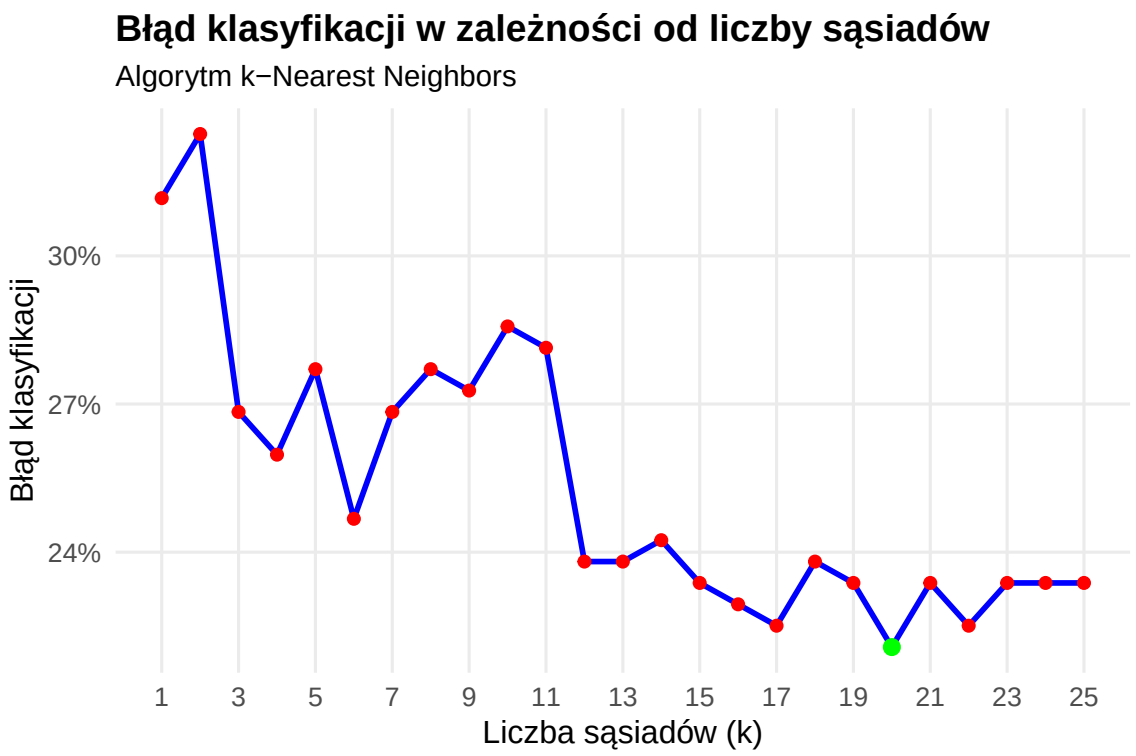
Celem niniejszej części analizy jest szczegółowe porównanie trzech wybranych algorytmów klasyfikacyjnych: K-najbliższych sąsiadów (KNN), naiwnego klasyfikatora Bayesowskiego oraz drzewa decyzyjnego. Aby zapewnić obiektywną i rzetelną ocenę skuteczności poszczególnych metod, wszystkie modele zostaną wytrenowane oraz przetestowane na tym samym podziale danych. Dzięki temu możliwe będzie bezpośrednie porównanie ich wyników, przy zachowaniu jednolitych warunków eksperymentu.



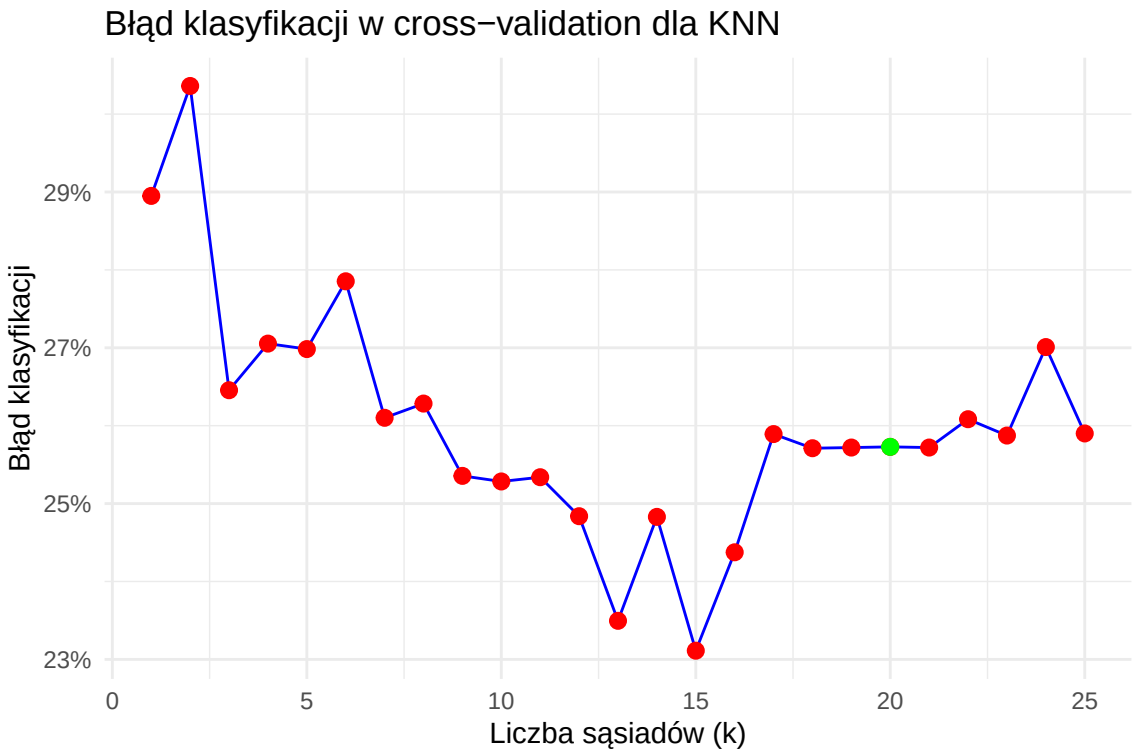
Rysunek 10: Macierz pomyłek dla algorytmu KNN, k = 5

Jak przedstawiono na rysunku 10, metoda KNN przy liczbie sąsiadów ‘k = 5’ nie osiąga zadowalających rezultatów. Wartość precyzji dla klasy 0 wynosi 0.78, natomiast dla klasy 1 0.61. Uzyskane wyniki wskazują na niewystarczającą skuteczność modelu w obecnej konfiguracji, co skłania do dalszej analizy wpływu hiperparametru ’k’na jakość klasyfikacji.

W tym celu porównamy błędy klasyfikacji algorytmu dla różnych wartości ‘k’, najpierw stosując jednokrotny podział danych na zbiór treningowy i testowy, a następnie wykorzystując walidację krzyżową. Podejście to pozwoli na bardziej obiektywną ocenę efektywności modelu w zależności od doboru liczby sąsiadów.



Rysunek 11: Porównanie błędu klasyfikacji KNN dla różnych wartości hiperparametru k



Rysunek 12: Porównanie błędu klasyfikacji KNN dla różnych wartości hiperparametru k z uwzględnieniem walidacji krzyżowej

Rysunki 11 oraz 12 ilustrują estymowaną wartość błędu klasyfikacji w zależności od liczby sąsiadów k w algorytmie KNN. Na podstawie wykresu dla jednokrotnego podziału zbioru danych wnioskujemy, że najniższy błąd osiągany jest dla k = `20`. Z kolei analiza wyników uzyskanych metodą walidacji krzyżowej wskazuje, że optymalną wartością hiperparametru jest k = `15`. Obserwacje te potwierdzają znaczenie odpowiedniego doboru parametru k oraz pokazują, że walidacja krzyżowa może dostarczyć bardziej wiarygodnych oszacowań błędu generalizacji.