Sprawozdanie z listy 3

Eksploracja danych

Marta Stankiewicz (282244)
Paweł Nowak (282223)

2025-05-28

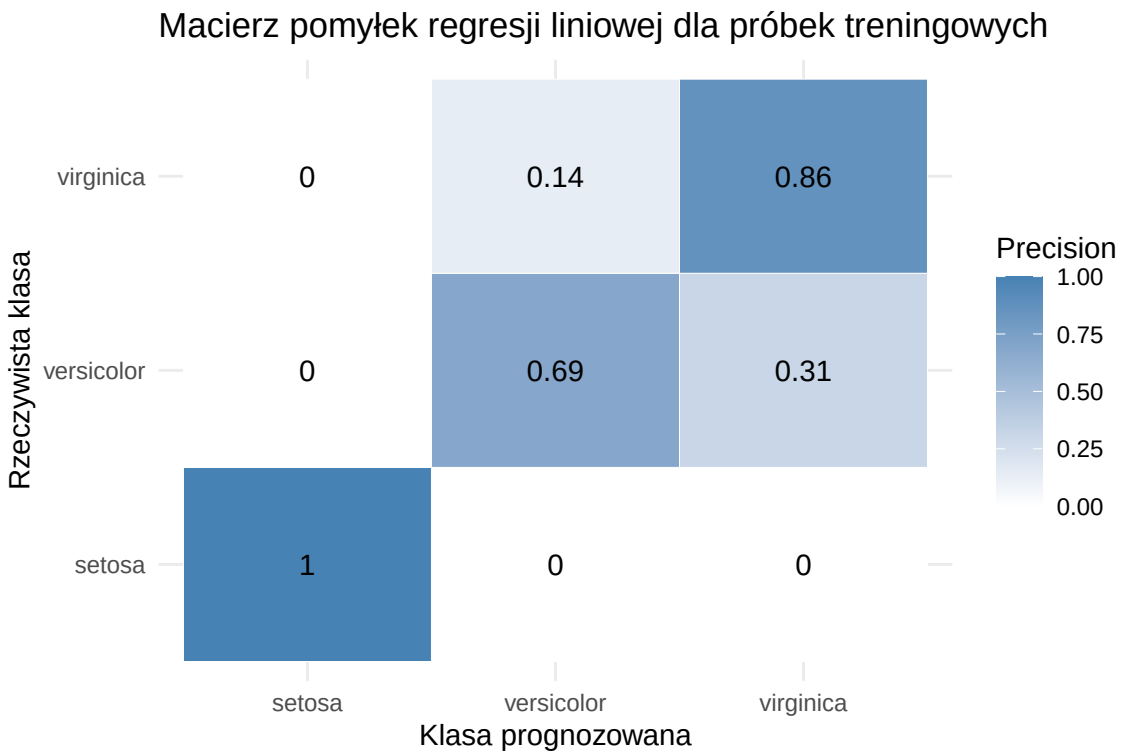
Spis treści

Spis rysunków

Spis tabel

1 Klasyfikacja na bazie modelu regresji liniowej

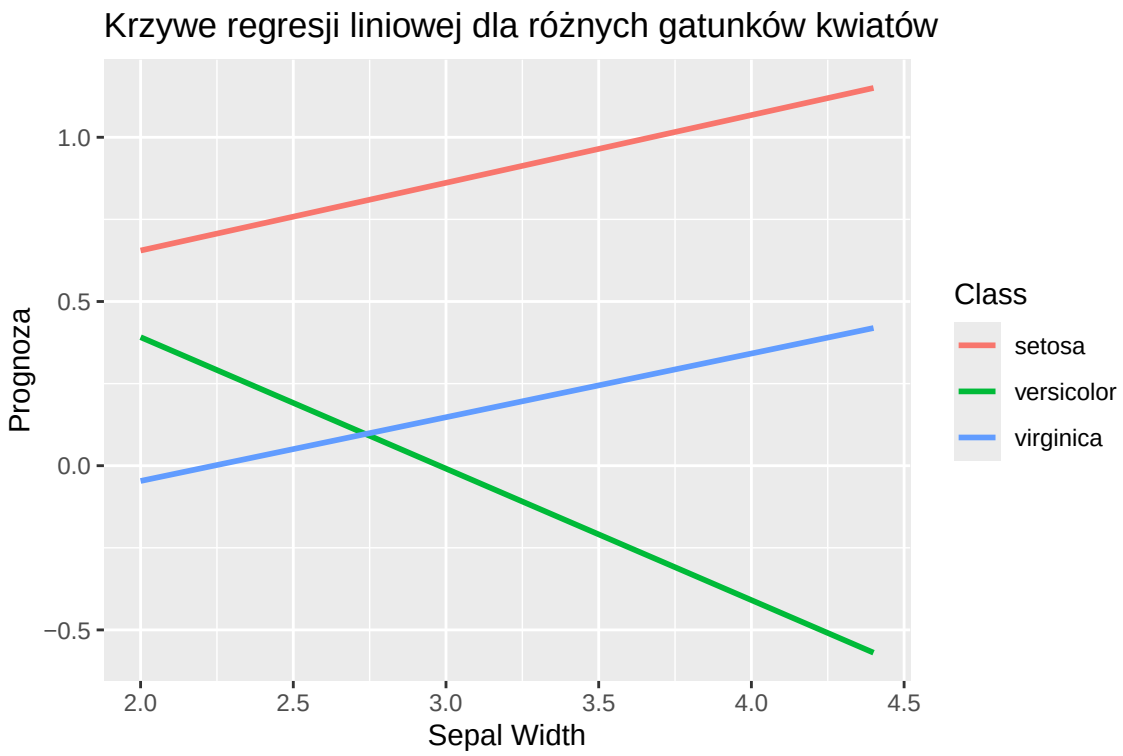
Aby ocenić skuteczność klasyfikatora opartego na modelu regresji liniowej, wykorzystamy zbiór danych iris, w którym zmienną objaśnianą jest Species, zawierająca 3 unikalnych klas. W celu uniknięcia problemu wycieku danych (ang. data leakage), przeprowadzimy podział oryginalnego zbioru na zbiór treningowy 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

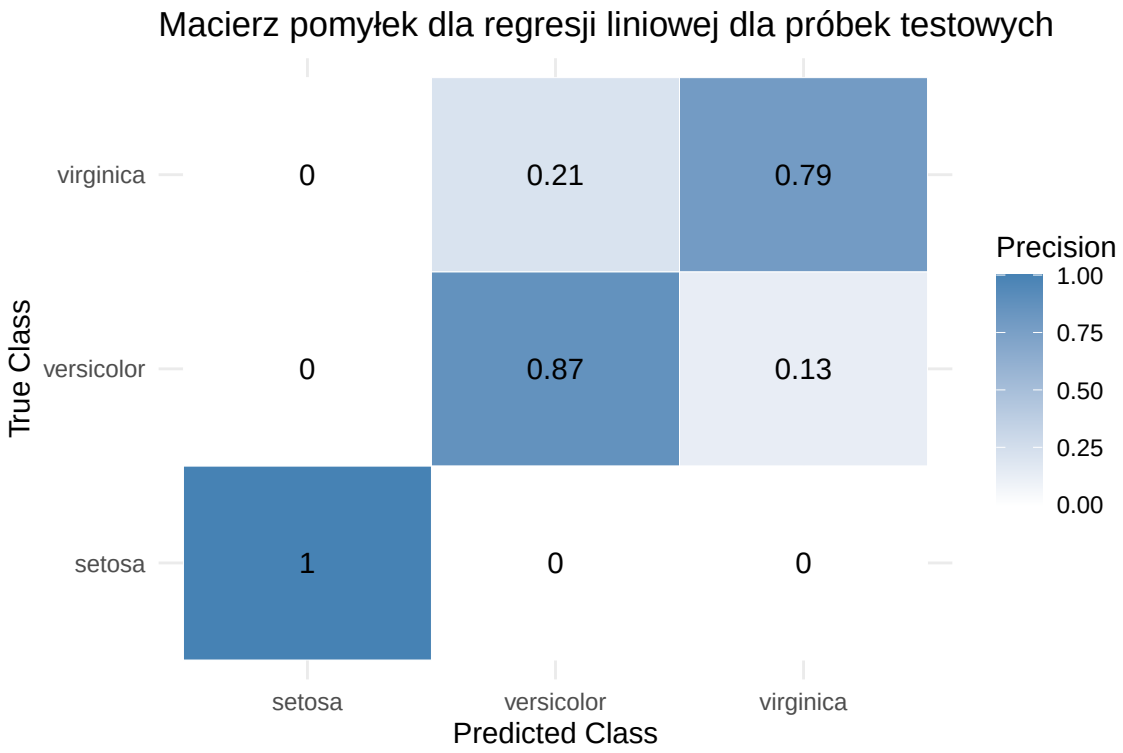
Na podstawie macierzy pomyłek przedstawionej na Rysunku 1, przyjrzeliśmy się, jak dobrze klasyfikator oparty na regresji liniowej radzi sobie z przewidywaniem poszczególnych klas. Zauważyliśmy, że skuteczność tych przewidywań różni się w zależności od tego, do której klasy należą próbki treningowe. Najlepiej 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



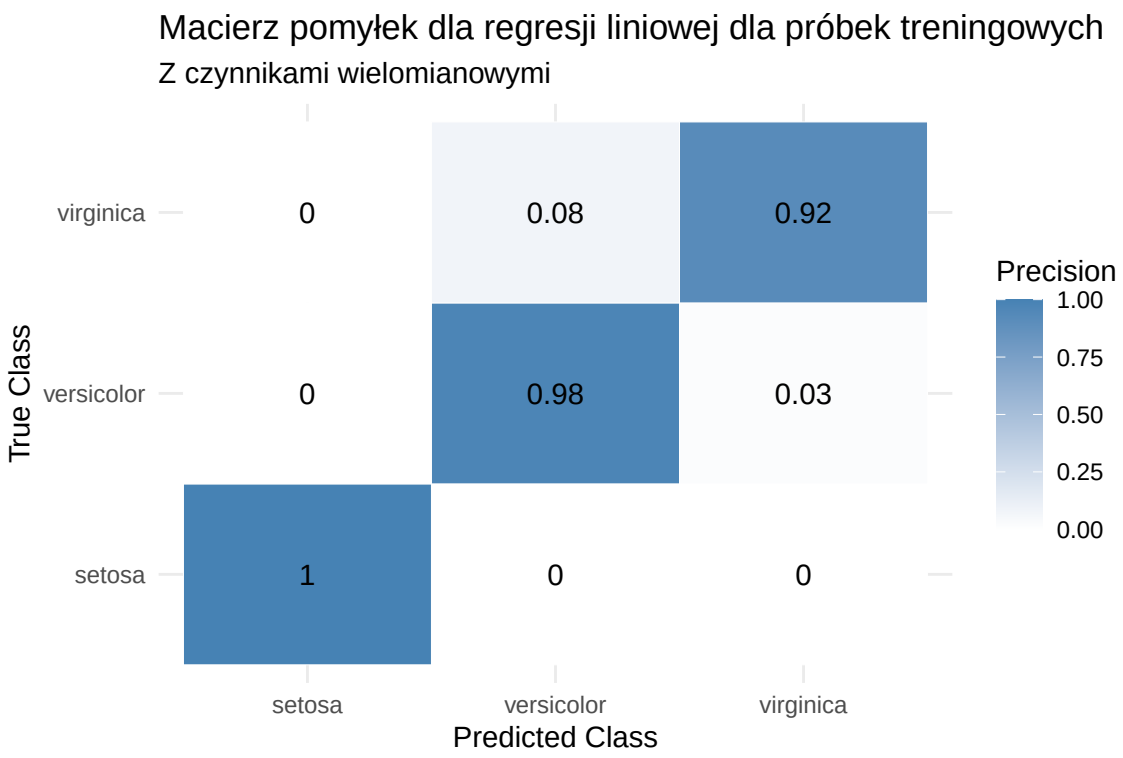
Rysunek 3: Macierz pomyłek regresji liniowej dla obserwacji testowych

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.

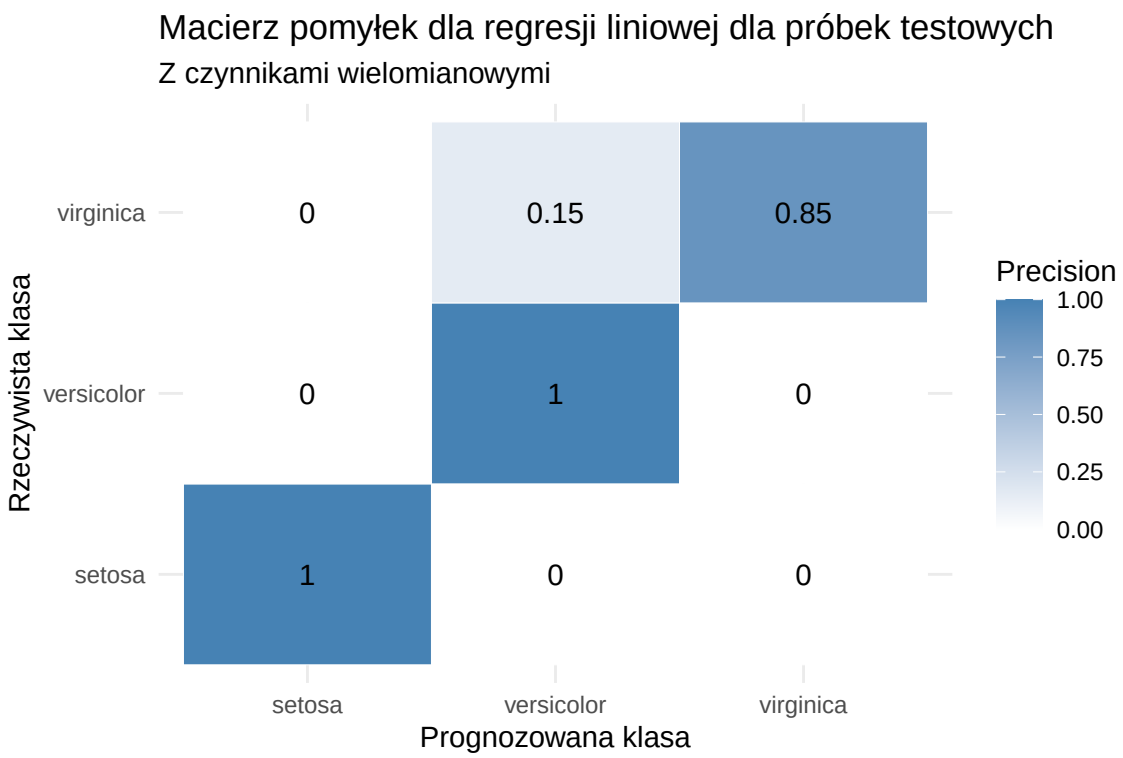
2 Klasyfikacja na bazie modelu regresji liniowej z czynnikami wielomianowymi

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.

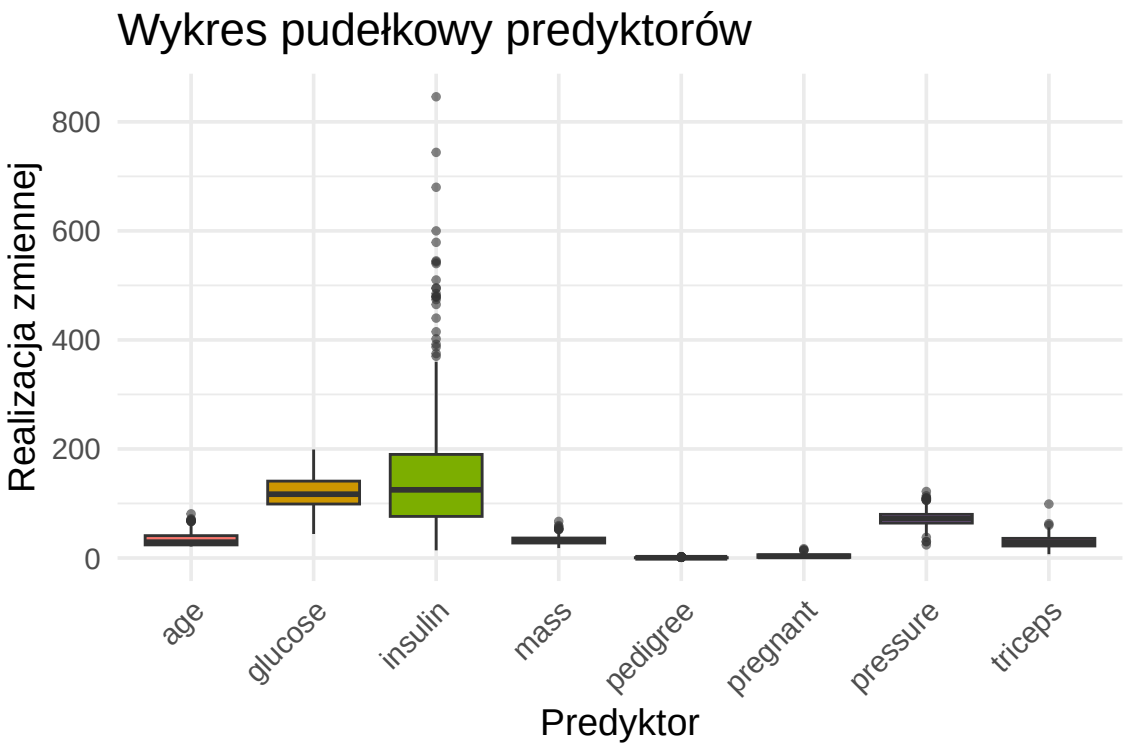
3.1 Wstępna analiza danych.

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

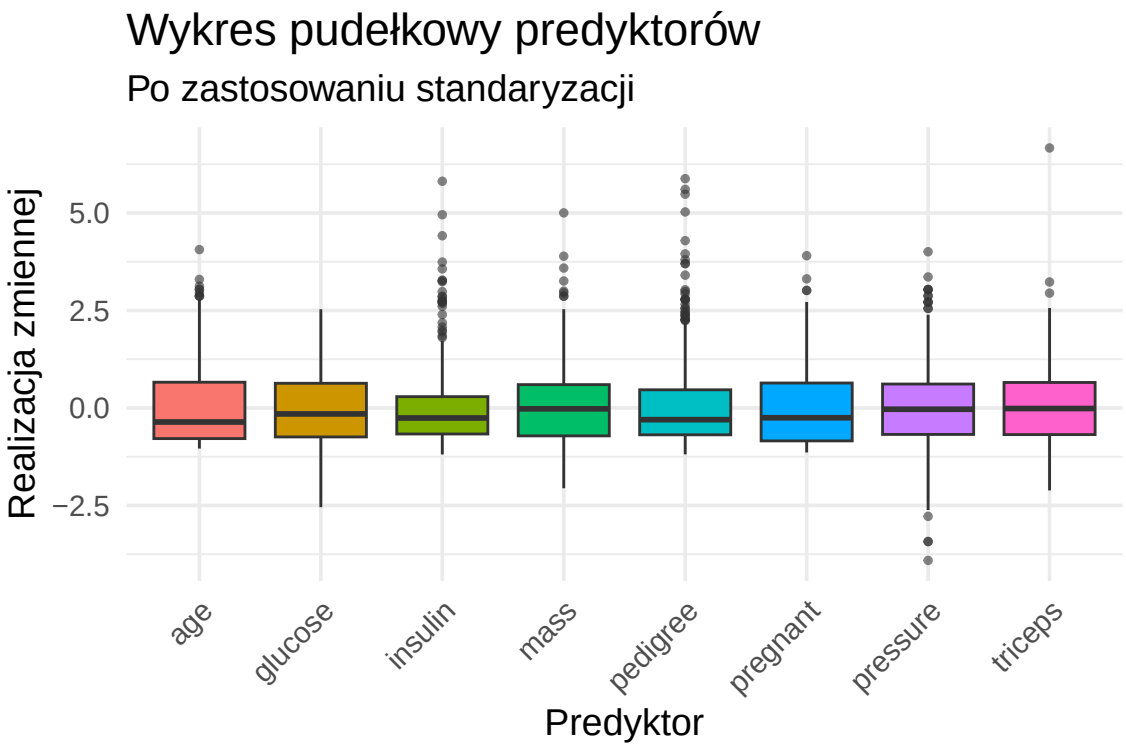
Z rysunku 6 wyraźnie wynika, że mamy do czynienia z problemem niezbalansowanych danych — liczba osób niechorujących na cukrzycę typu 2 jest niemal dwukrotnie większa niż liczba osób chorych. Teoretycznie, gdybyśmy zastosowali prosty, „naiwny" klasyfikator przypisujący każdą obserwację do klasy dominującej, 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.



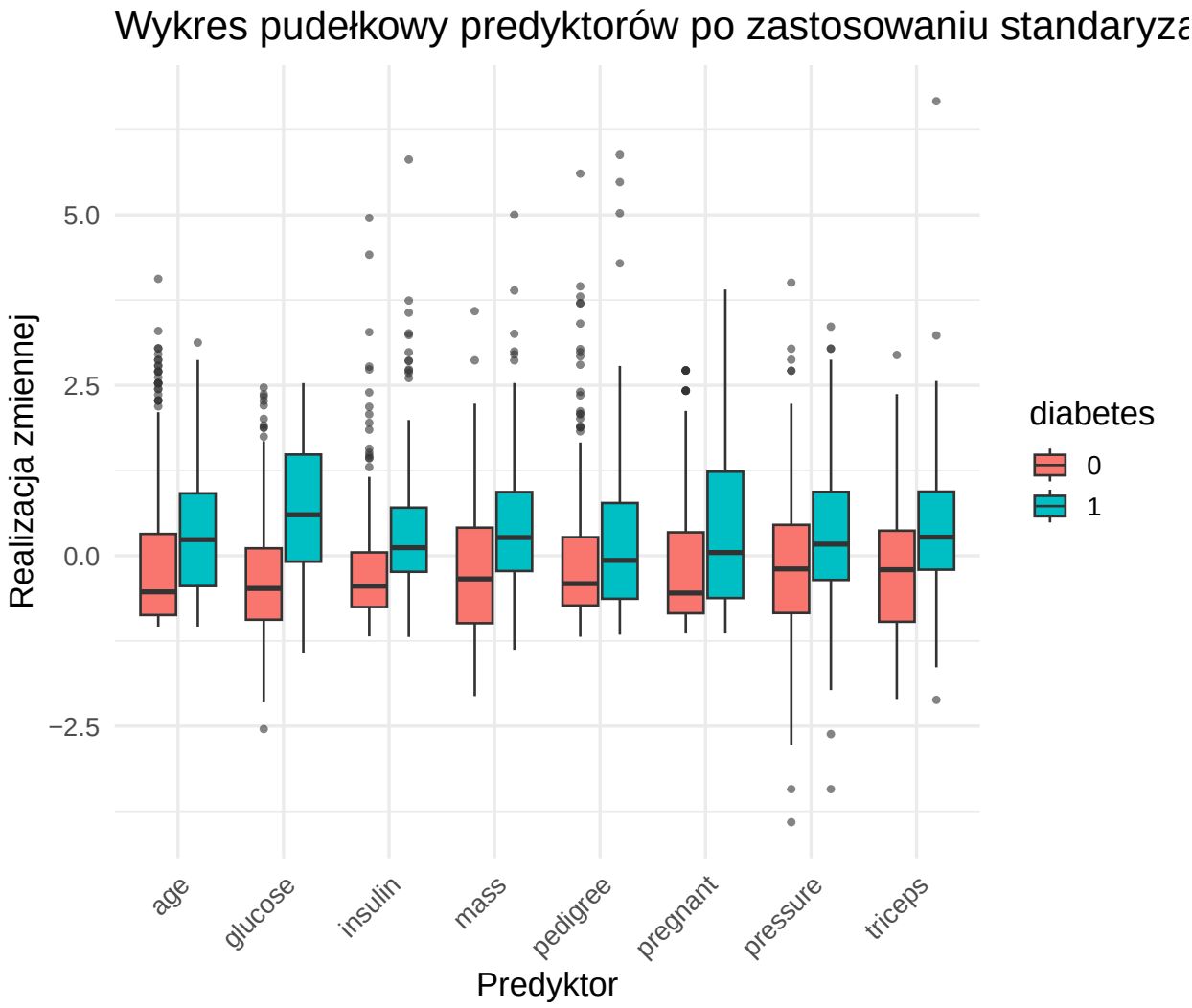
Rysunek 7: Porównanie wariancji predyktorów

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.

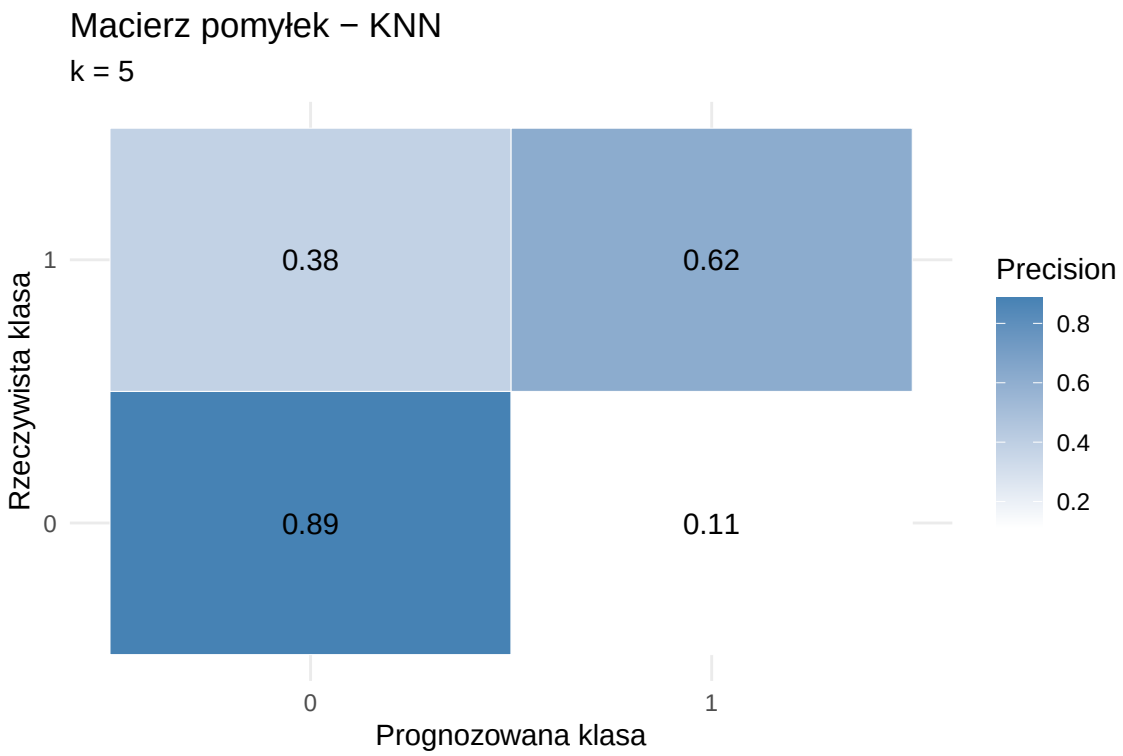


Rysunek 9: Porównanie zdolności dyskryminacyjnych predyktorów

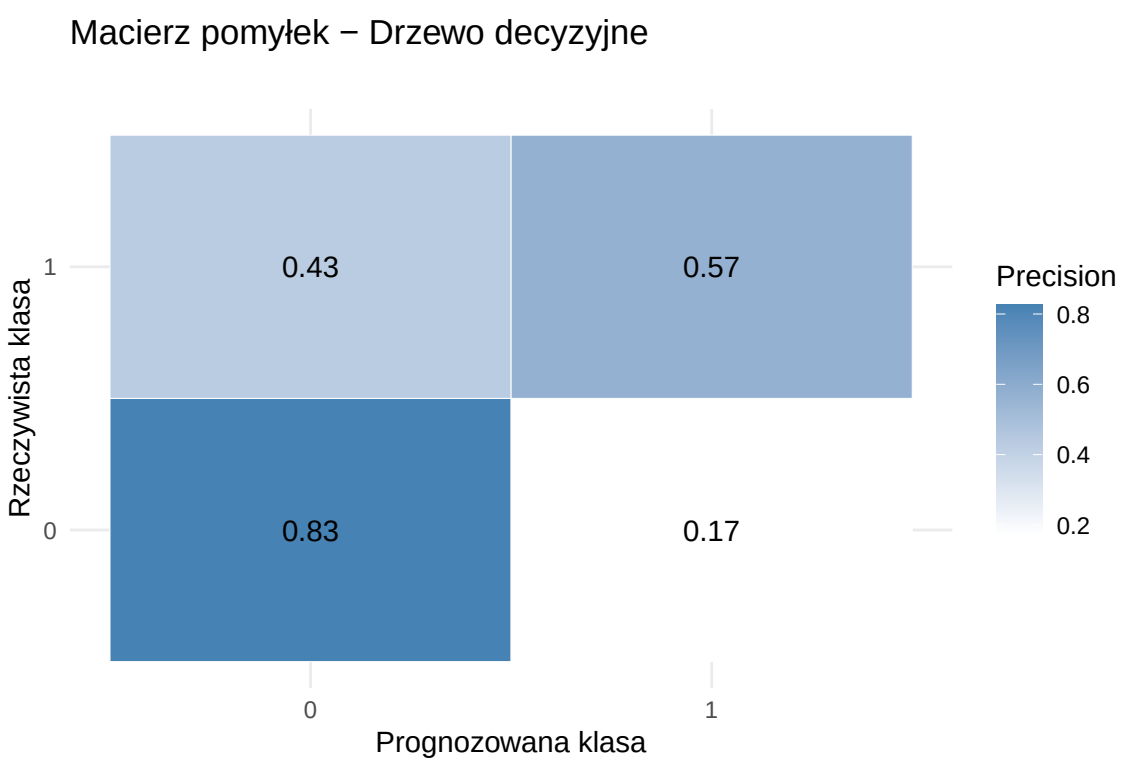
Jak pokazano na rysunku 9, żadna ze zmiennych nie rozdziela klas docelowych w sposób idealny. Mimo to, można wyróżnić predyktory o wyraźnie silniejszych zdolnościach dyskryminacyjnych. Do takich atrybutów należą zmienne: ‘glucose’, ‘insulin’, ‘triceps’ oraz ‘mass’. Natomiast najsłabszą zdolność separacji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

4 Ocena dokładności klasyfikacji i porównanie metod

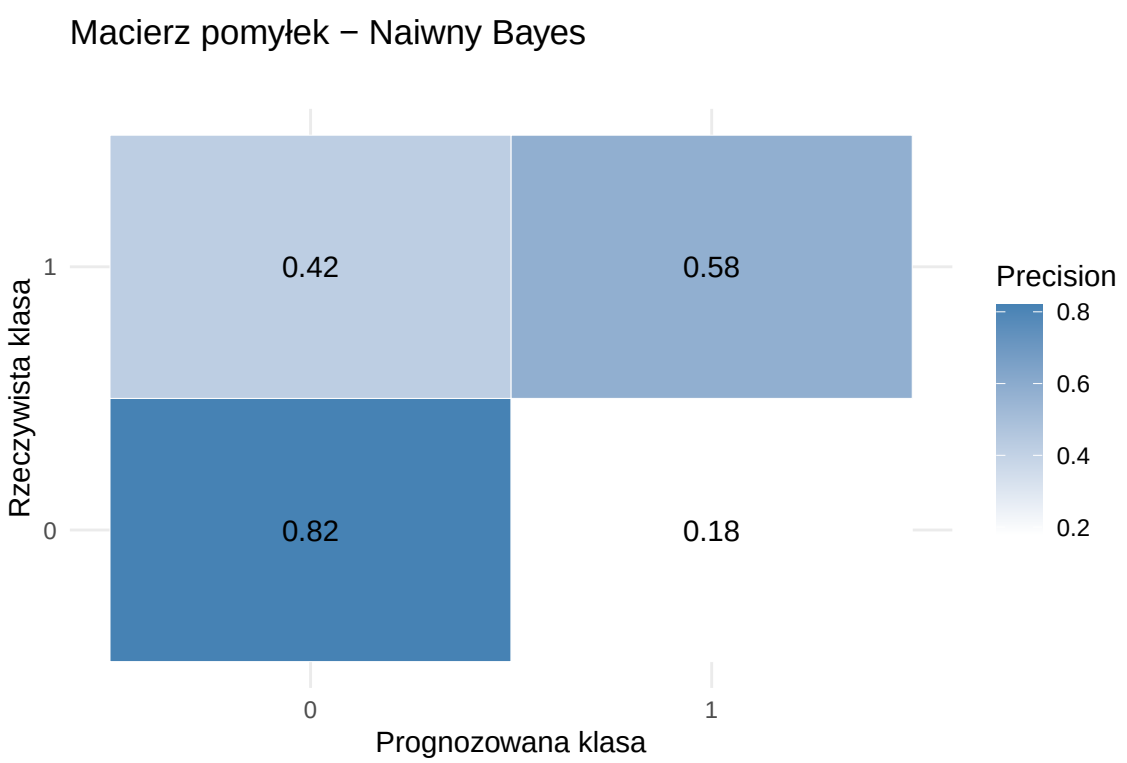
Celem niniejszej części analizy jest szczegółowe porównanie trzech wybranych algorytmów klasyfikacyjnych: K-najbliższych sąsiadów (KNN), naiwnego klasyfikatora Bayesowskiego oraz drzewa decyzyjnego. Aby zapewnić obiektywną i rzetelną ocenę skuteczności poszczególnych metod, wszystkie modele zostaną wytrenowane 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



Rysunek 11: Macierz pomyłek dla drzewa klasyfikacyjnego



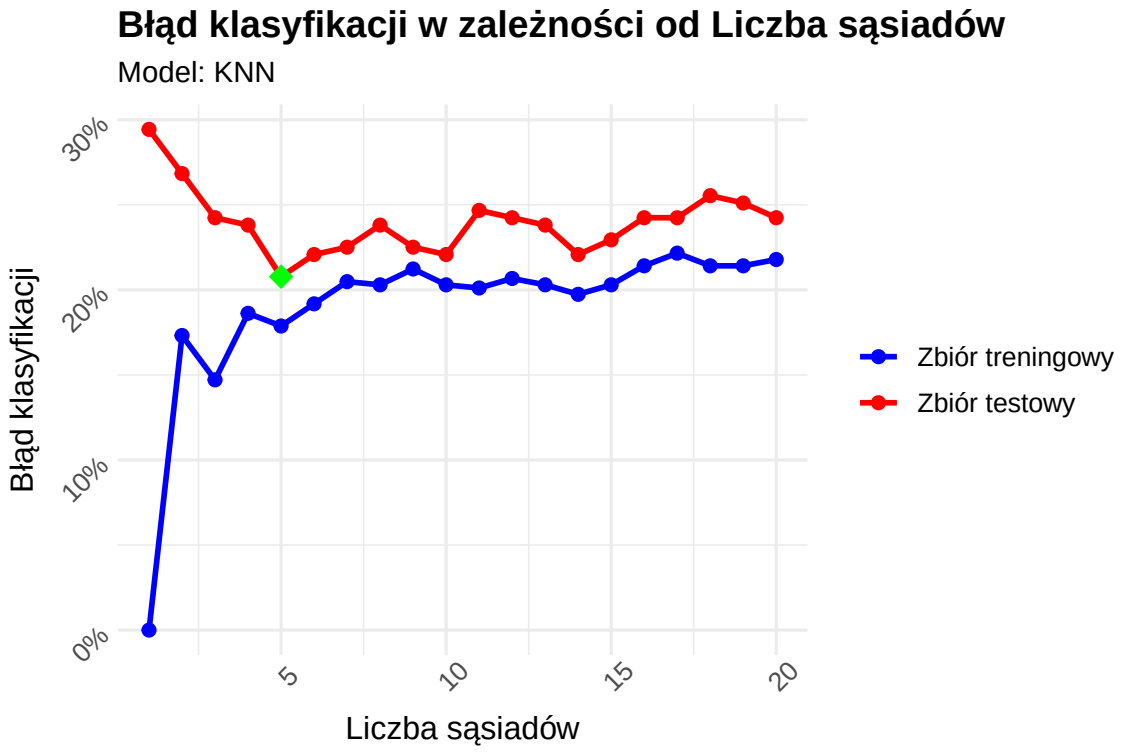
Rysunek 12: Macierz pomyłek dla naiwnego klasyfikatora bayesa

Analizując macierze pomyłek dla wybranych algorytmów można zauważyć, że najlepsze wyniki uzyskał algorytm KNN przy liczbie sąsiadów 'k=5'. Wartość precyzji tego modelu dla klasy 0 wynosi 0.89 natomiast dla klasy 1 - 0.62. Wskazuje to na jego wysoką skuteczność w rozpoznawaniu zarówno dominującej, jak i mniejszościowej klasy. W porównaniu do pozostałych metod KNN popełnił najmniej błędów klasyfikacyjnych, co czyni go najbardziej trafnym algorytmem dla analizowanego przypadku.

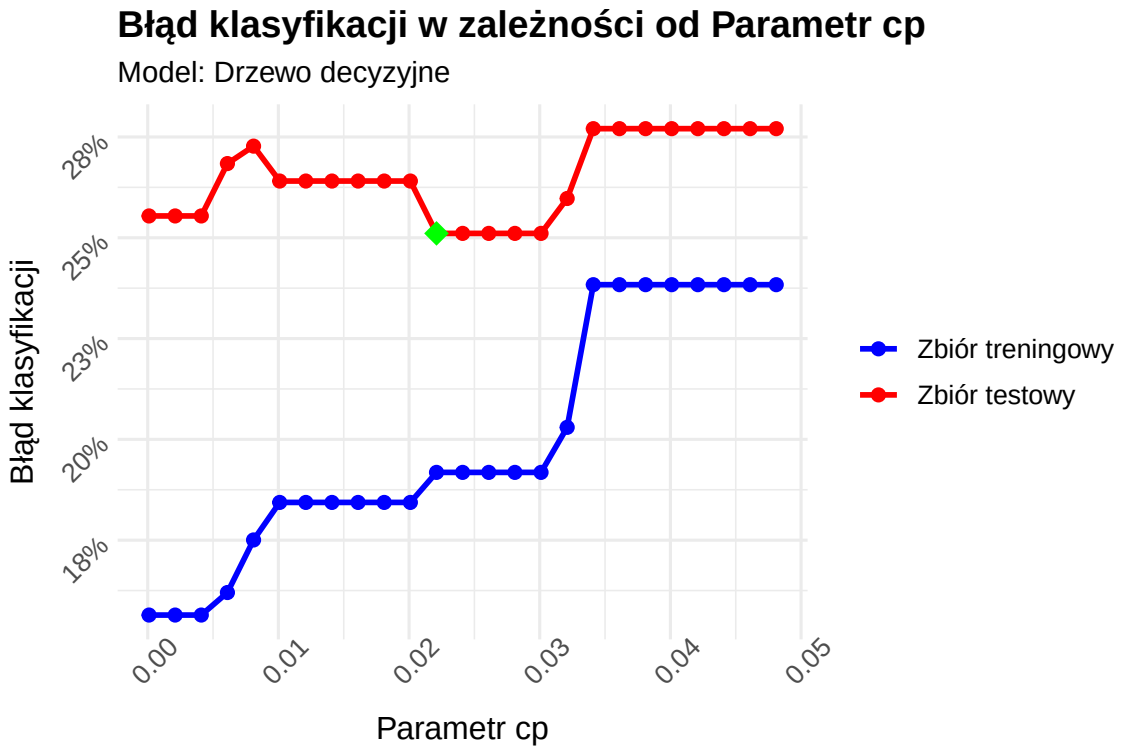
Drzewo decyzyjne osiągnęło nieco gorsze wyniki. Precyzja w tym przypadku wynosi 0.83 dla klasy 0 oraz 0.57 dla klasy 1. Choć jego skuteczność była niższa niż w przypadku KNN, to nadal pozostaje ona na akceptowalnym poziomie, szczególnie biorąc pod uwagę prostotę interpretacji struktury drzewa decyzyjnego.

Najniższą skuteczność wykazał naiwny klasyfikator Bayesa, który poprawnie rozpoznał 0.82 przypadków klasy 0 oraz 0.58 przypadków klasy 1. Chociaż jego wyniki są zbliżone do drzewa decyzyjnego, nie oferuje on wyraźnych przewag nad pozostałymi modelami.

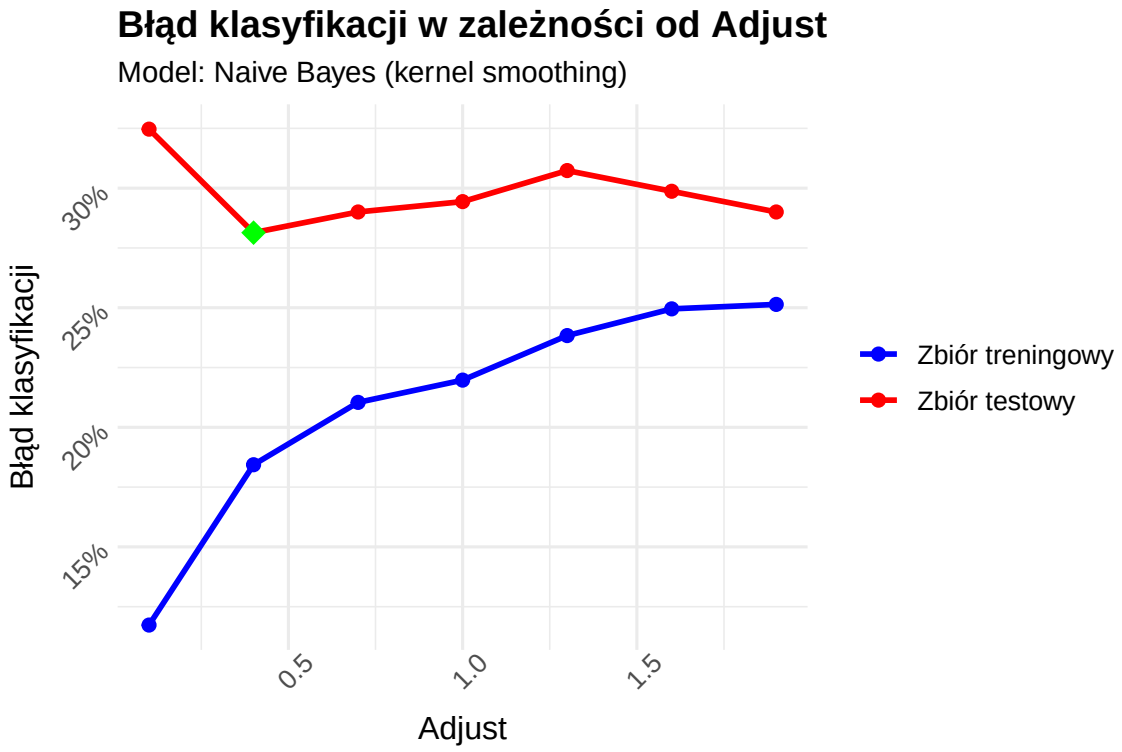
Warto jednak zwrócić uwagę, że zaprezentowane wyniki odnoszą się do konkretnej konfiguracji parametrów modeli. Można zatem sprawdzić, czy możliwe jest uzyskanie większej skuteczności po odpowiednim dobraniu parametrów w powyższych algorytmach. W tym celu porównamy błędy klasyfikacji algorytmu dla różnych wartości parametrów, najpierw stosując jednokrotny podział danych na zbiór treningowy i testowy, a następnie wykorzystując walidację krzyżową. Podejście to pozwoli na bardziej obiektywną ocenę efektywności modeli.



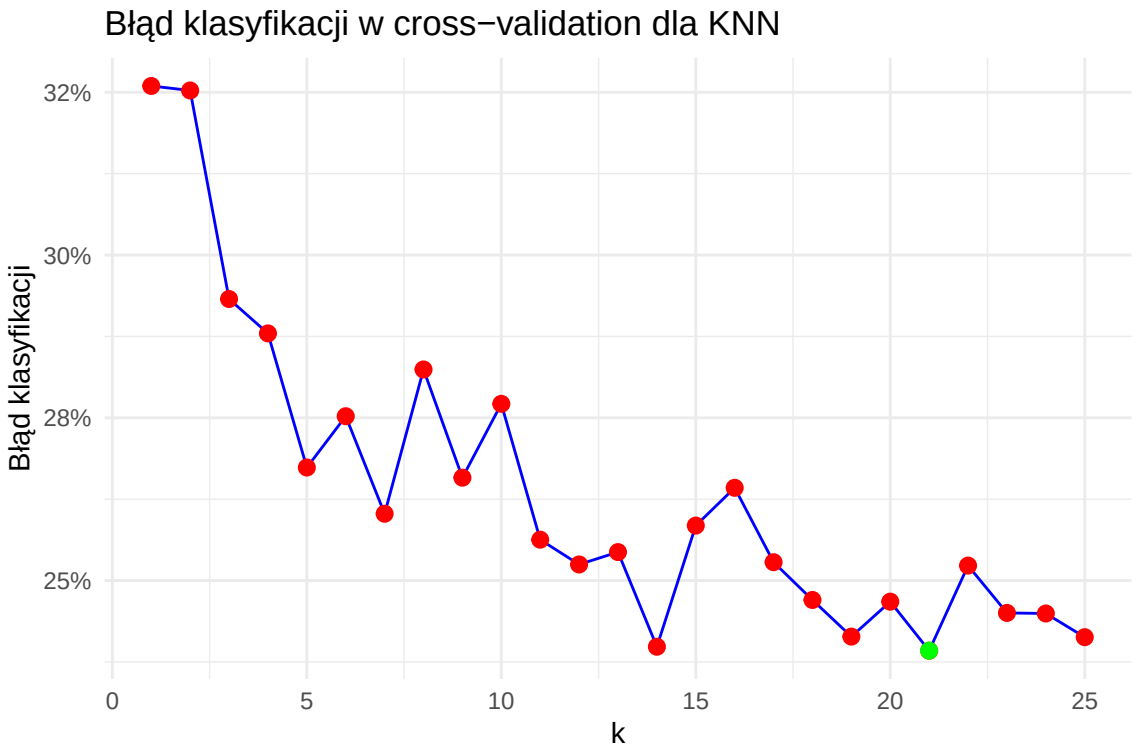
Rysunek 13: Porównanie błędu klasyfikacji KNN dla różnych wartości hiperparametru k



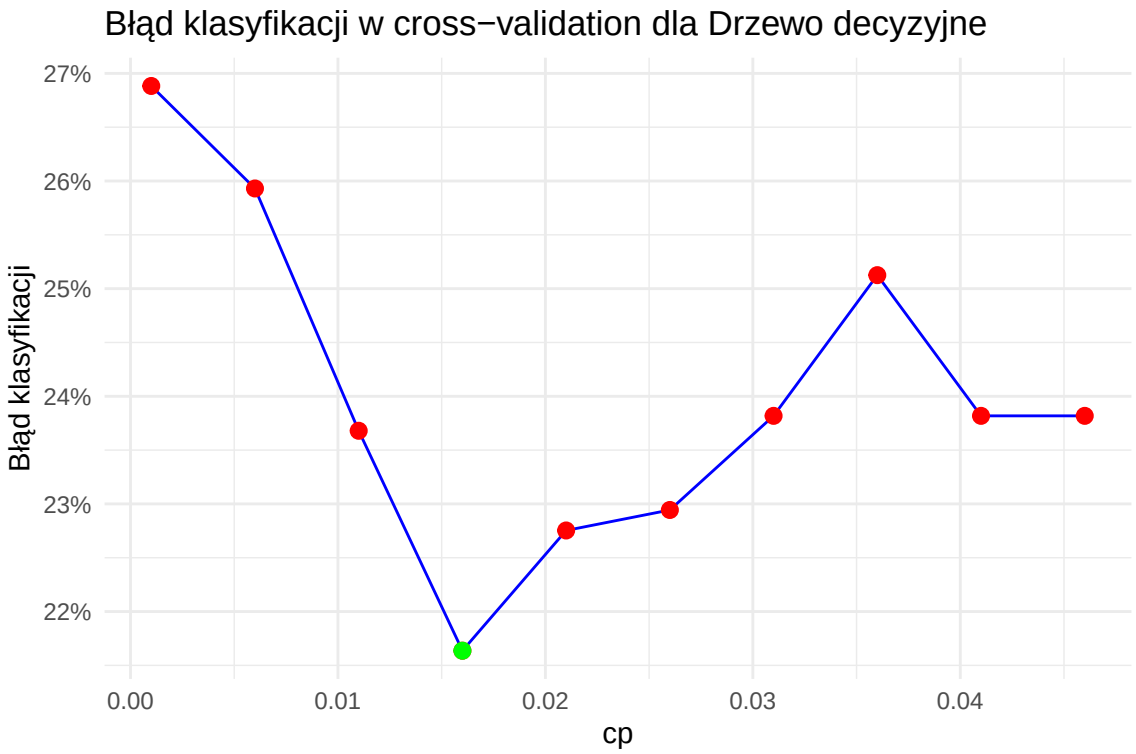
Rysunek 14: Porównanie błędu klasyfikacji drzewa decyzyjnego dla różnych wartości parametru cp



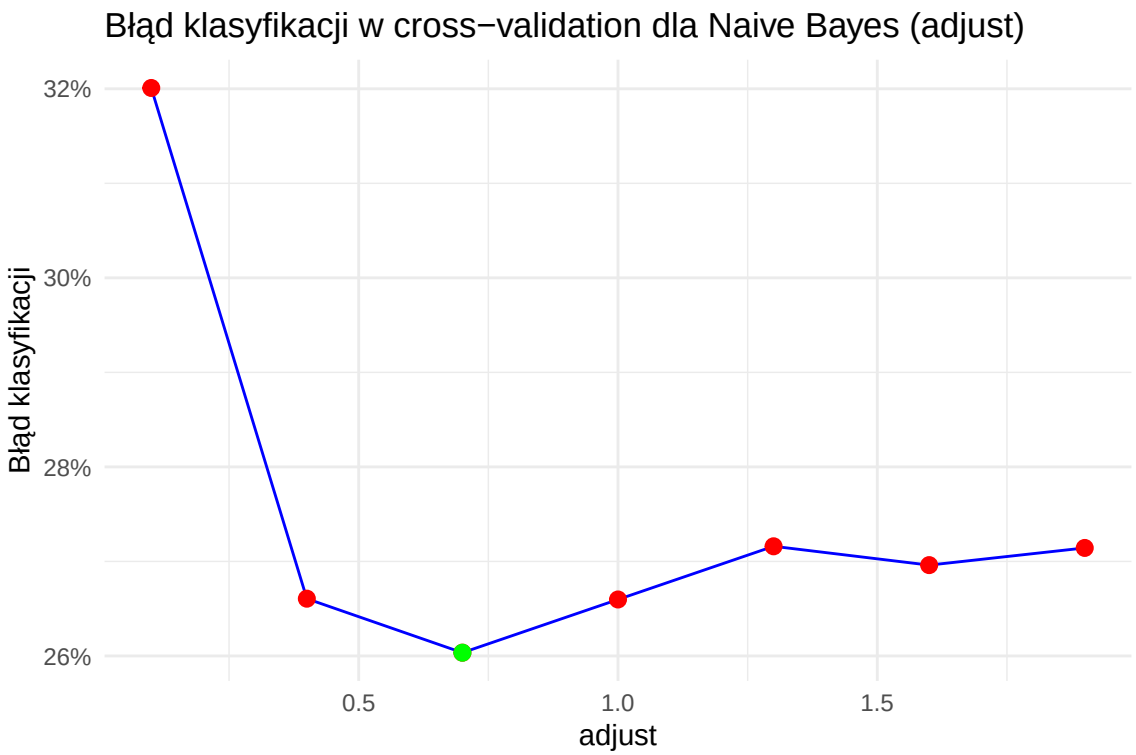
Rysunek 15: Porównanie błędu klasyfikacji naiwnego klasyfikatora bayesowskiego dla różnych wartości parametru laplace



Rysunek 16: Porównanie błędu klasyfikacji KNN dla różnych wartości hiperparametru k z uwzględnieniem walidacji krzyżowej



Rysunek 17: Porównanie błędu klasyfikacji drzewa klasyfikacyjnego dla różnych wartości parametru cp z uwzględnieniem walidacji krzyżowej



Rysunek 18: Porównanie błędu klasyfikacji naiwnego klasyfikatora bayesa dla różnych wartości parametru laplace z uwzględnieniem walidacji krzyżowej

Na wykresie 13 przedstawiono zależność błędu klasyfikacji od wartości hiperparametru k w algorytmie KNN. Najmniejszy błąd klasyfikacji dla zbioru testowego uzyskano dla k = 5, a jego wartość wynosiła 0.21.

Warto jednak zauważyć, że pojedynczy podział danych na zbiór treningowy i testowy może prowadzić do niestabilnych wyników – jest wrażliwy na sposób losowego podziału. Właśnie dlatego stosuje się walidację krzyżową, która uśrednia wyniki z wielu podziałów, co prowadzi do bardziej wiarygodnej oceny skuteczności modelu.

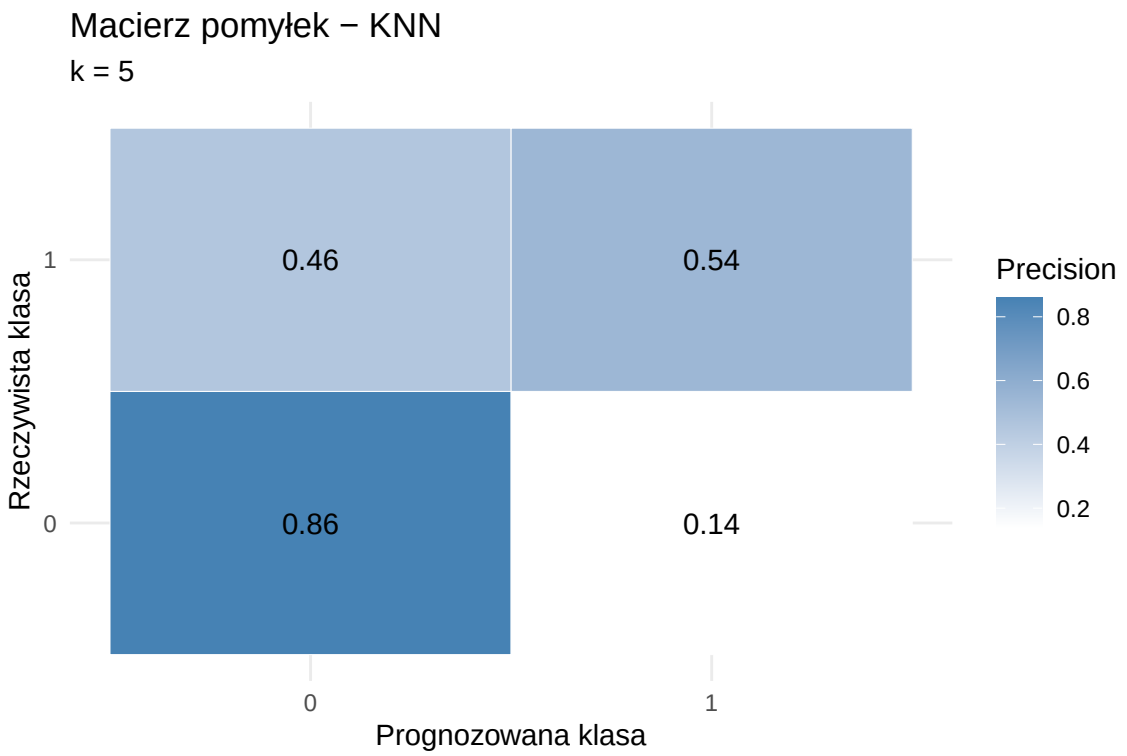
W przypadku walidacji krzyżowej, najmniejszy błąd klasyfikacji osiągnięto dla k = 21, a jego wartość wyniosła 0.24. Choć wynik z walidacji krzyżowej może być nieco gorszy w konkretnym przypadku, jego średnia efektywność w różnych scenariuszach jest wyższa, co czyni go bardziej reprezentatywnym.

Na wykresie 14 zaprezentowano wpływ wartości parametru cp (complexity parameter) na błąd klasyfikacji w drzewie decyzyjnym. Parametr ten określa minimalną wartość poprawy jakości podziału, wymaganą do wykonania kolejnego rozgałęzienia drzewa. Najniższy błąd klasyfikacji na zbiorze testowym zaobserwowano dla cp = 0.02, przy wartości błędu równej 0.25. Z kolei wyniki walidacji krzyżowej wskazują, że optymalna średnia wartość parametru to 0.02, dla której błąd klasyfikacji wyniósł 0.22.

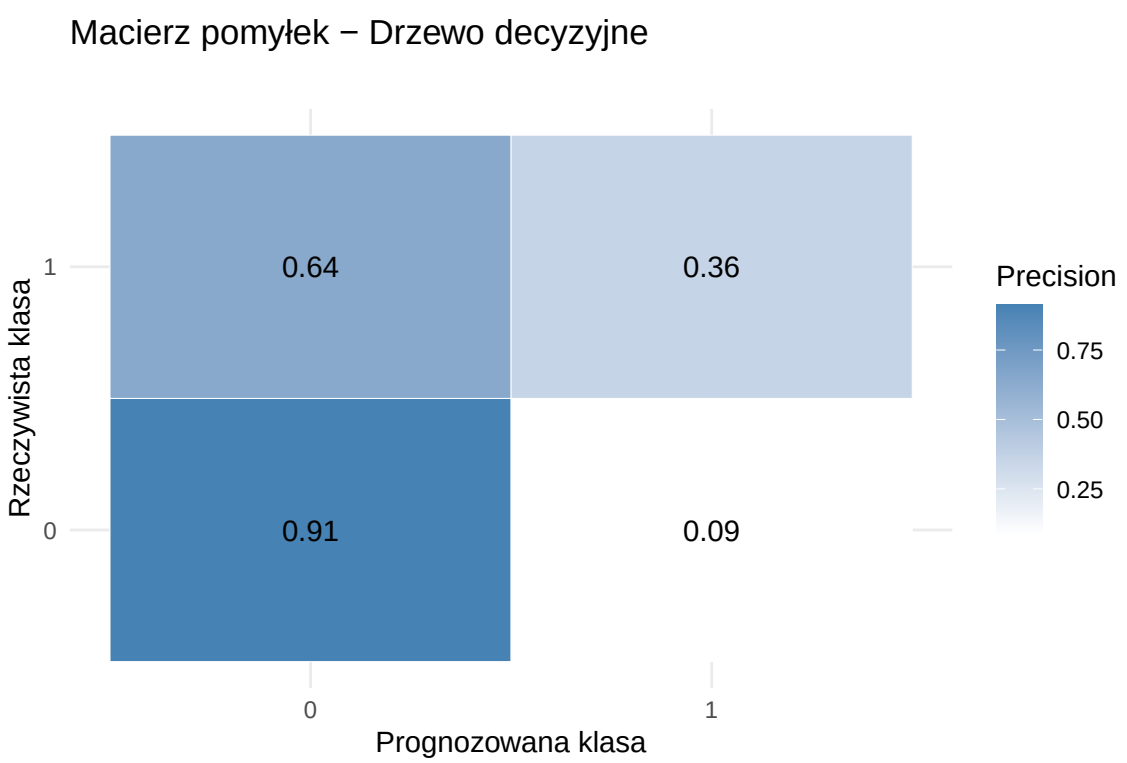
Dla naiwnego klasyfikatora Bayesa analizowano wpływ parametru adjust, regulującego stopień wygładzania przy estymacji rozkładów jądrowych. Najlepszy wynik na zbiorze testowym osiągnięto przy adjust = 0.4, a odpowiadający mu błąd klasyfikacji wyniósł 0.28. Z kolei walidacja krzyżowa wykazała, że najniższy błąd klasyfikacji (0.26) uzyskano przy adjust = 0.7.

4.1 Nowy podzbiór zmiennych

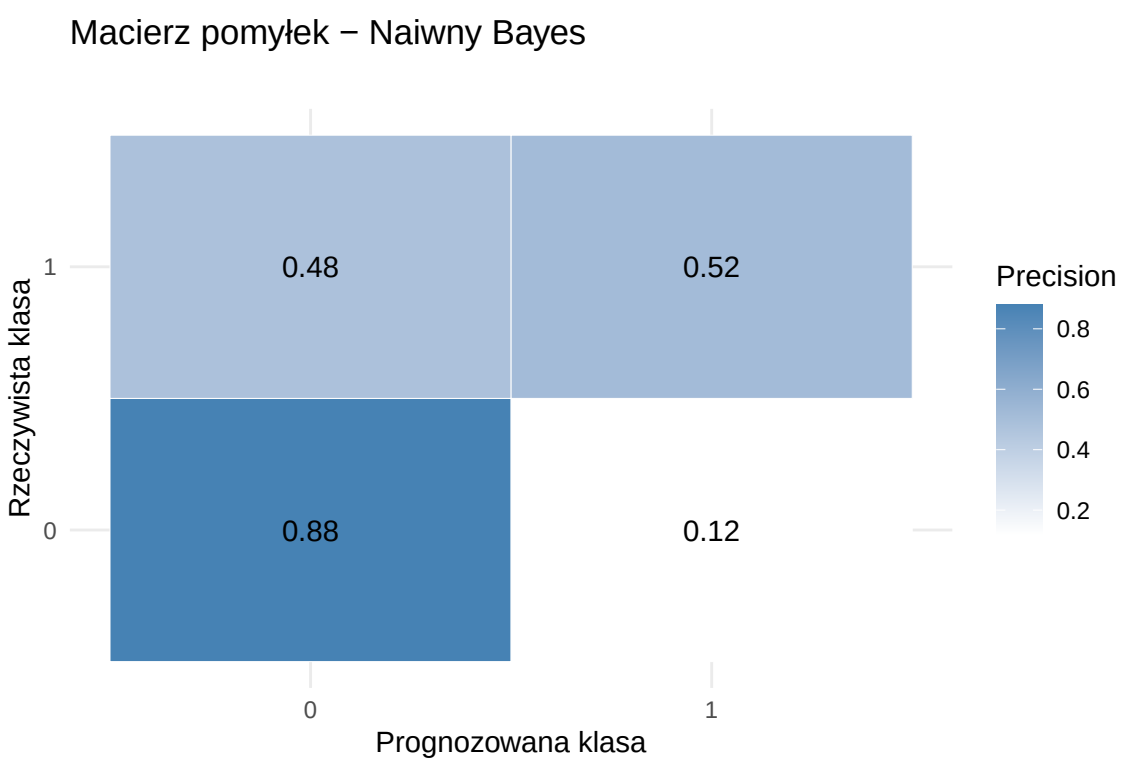
Na podstawie wstępnej analizy zdolności dyskryminacyjnych zmiennych ze zbioru danych PimaIndiansDiabetes2, można wyodrębnić podzbiór cech najlepiej różnicujących obserwacje - w naszym przypadku będą to zmienne: “glucose”, “mass” oraz “insulin”. Taki podzbiór następnie posłuży do kontynuacji porównania skuteczności trzech wybranych algorytmów klasyfikacyjnych: k-najbliższych sąsiadów (k-NN), drzewa decyzyjnego oraz naiwnego klasyfikatora Bayesa.



Rysunek 19: Macierz pomyłek dla algorytmu KNN, k = 5, dla nowego podzbioru danych



Rysunek 20: Macierz pomyłek dla drzewa klasyfikacyjnego dla nowego podzbioru danych



Rysunek 21: Macierz pomyłek dla naiwnego klasyfikatora bayesa dla nowego podzbioru danych

Po wybraniu nowego podzbioru, mimo że składa się on ze zmiennych o najlepszych własnościach dyskryminacyjnych, może dojść do pogorszenia jakości klasyfikacji w porównaniu do modelu trenowanego na pełnym zestawie cech.

Redukcja liczby zmiennych oznacza utratę części informacji, która — choć może wydawać się mniej istotna pojedynczo — w połączeniu z innymi cechami pomaga modelowi lepiej rozróżniać klasy. Czasem dodatkowe zmienne zawierają wzorce lub korelacje, które poprawiają zdolność predykcyjną modelu. Dodatkowo zmniejszenie wymiarowości danych może wpływać na stabilność i uogólnialność modelu, zwłaszcza w przypadku metod takich jak k-najbliższych sąsiadów czy drzew decyzyjnych, które korzystają z pełnego obrazu przestrzeni cech.

Warto jednak zauważyć, że mimo tych ograniczeń, modele: drzewa decyzyjnego oraz naiwny klasyfikator bayesa wyraźnie lepiej radzą sobie z klasyfikacją klasy 0 na ograniczonym podzbiorze cech. Szczególnie drzewo klasyfikacyjne osiąga wysoką dokładność — ‘około 91%’ — co sugeruje, że wybrane cechy istotnie wspierają rozpoznawanie tej klasy. Taki wynik wskazuje, że selekcja cech, choć powoduje pewną utratę ogólnej efektywności, może poprawić wydajność w kontekście konkretnych zadań lub klas.

5 Końcowe wnioski - podsumowanie

Najlepsze wyniki uzyskano korzystając z pełnego zestawu zmiennych predykcyjnych, co pozwoliło modelom na wykorzystanie wszystkich dostępnych informacji. W przypadku algorytmu k-najbliższych sąsiadów (KNN) optymalna liczba sąsiadów wyniosła 21 (walidacja krzyżowa), co przełożyło się na najniższy błąd klasyfikacji 0.24. Dla drzewa decyzyjnego najlepszą wartością parametru złożoności (cp) była 0.02, a dla naiwnego klasyfikatora Bayesa optymalny stopień wygładzania (adjust) wyniósł 0.7.

Selekcja podzbioru cech ograniczona do "glucose", "mass" oraz "insulin" poprawiła skuteczność klasyfikacji klasy 0, zwłaszcza w drzewie decyzyjnym, gdzie dokładność klasyfikacji tej klasy osiągnęła około 91%. Jednakże w przypadku pełnego zestawu cech modele osiągały lepsze ogólne wyniki, co wskazuje na wartość zachowania wszystkich dostępnych informacji.

Najlepsze rezultaty w analizie uzyskał algorytm KNN, który charakteryzuje się największą precyzją oraz najniższym błędem klasyfikacji zarówno dla klasy dominującej (0), jak i mniejszościowej (1). Drzewo decyzyjne zapewnia wyniki na poziomie akceptowalnym, szczególnie dobrze radząc sobie z klasyfikacją klasy 0, co jest istotne w kontekście interpretowalności modelu. Naiwny klasyfikator Bayesa wykazał najniższą skuteczność, choć jego wyniki nadal pozostają zbliżone do wyników drzewa, co czyni go potencjalnie użytecznym w niektórych zastosowaniach.

Wybór schematu oceny miał znaczący wpływ na wnioski dotyczące skuteczności poszczególnych metod. Wyniki uzyskane na podstawie jednokrotnego podziału danych były bardziej niestabilne i wrażliwe na losowość podziału, co mogło prowadzić do błędnych lub mylących wniosków. Wykorzystanie walidacji krzyżowej pozwoliło uśrednić wyniki z wielu podziałów, dzięki czemu ocena skuteczności modeli stała się bardziej rzetelna i reprezentatywna. W praktyce oznacza to, że parametry wybrane na podstawie walidacji krzyżowej lepiej generalizują na nowe, nieznane dane, a decyzje o wyborze najlepszych modeli są bardziej wiarygodne.