Sprawozdanie z listy 4

Eksploracja danych

Marta Stankiewicz (282244)
Paweł Nowak (282223)

2025-06-14

Spis treści

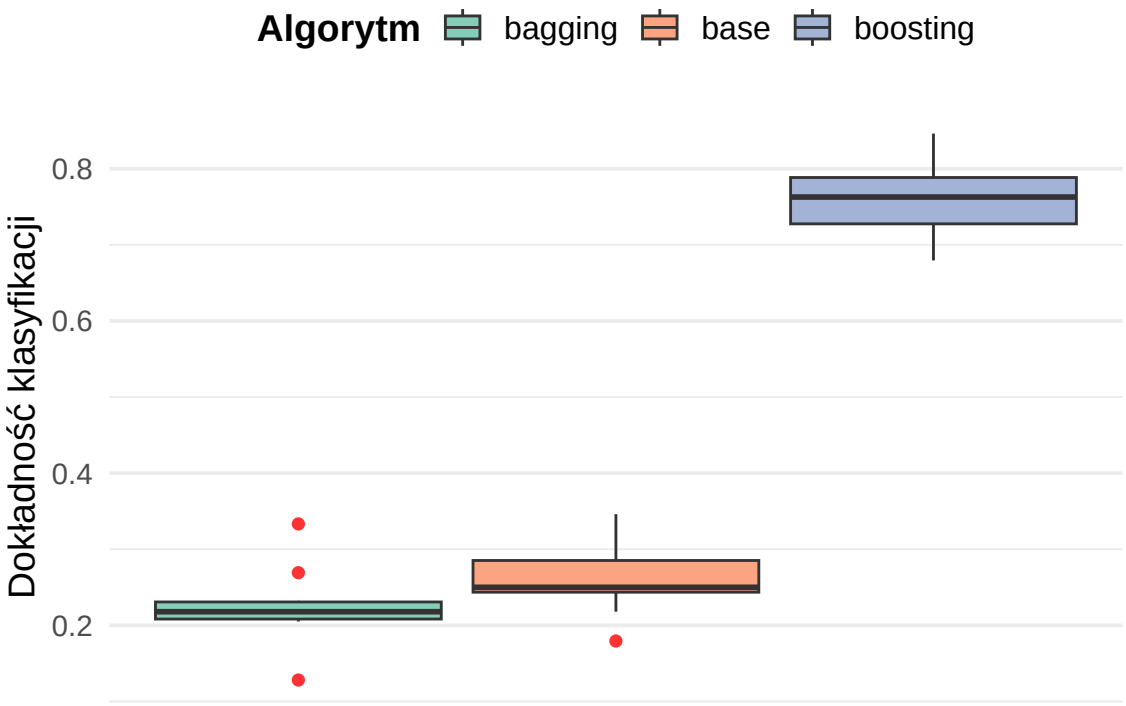
Spis rysunków

Spis tabel

1 Zaawansowane metody klasyfikacji

1.1 Rodziny klasyfikatorów/uczenie zespołowe

Celem niniejszej analizy jest zbadanie wpływu metod zespołowych (ensemble learning) na jakość i stabilność klasyfikacji na przykładzie zbioru danych PimaIndiansDiabetes2. W szczególności skonstruujemy modele oparte na klasyfikatorze bazowym — drzewie decyzyjnym — oraz rozszerzymy je za pomocą trzech popularnych technik: baggingu, boostingu oraz random forest. Naszą hipotezą jest to, że zastosowanie tych metod zespołowych pozwoli na istotne zmniejszenie wariancji estymatora, co przełoży się na poprawę stabilności modelu oraz redukcję błędu klasyfikacji w porównaniu do pojedynczego drzewa decyzyjnego. Spodziewamy się, że agregacja wielu modeli (bagging i random forest) oraz sekwencyjne poprawianie błędów (boosting) przyczynią się do zwiększenia dokładności predykcji stanu zdrowia pacjentów. W dalszej części pracy przedstawimy wyniki eksperymentów, porównamy skuteczność poszczególnych metod oraz ocenimy, czy obserwowany spadek wariancji przekłada się na realną poprawę jakości klasyfikacji. Aby uzyskać rzetelną i stabilną estymatę błędu klasyfikacji dla każdej z wymienionych metod, proces uczenia oraz testowania modeli został powtórzony 10 razy. W każdej iteracji modele były trenowane na tym samym zbiorze treningowym, co pozwoliło na sprawiedliwe i porównywalne ocenienie skuteczności poszczególnych algorytmów. Dzięki temu możliwe było także zbadanie stabilności wyników oraz ocena wariancji błędu klasyfikacji dla każdej techniki.



Rysunek 1: Porównanie skuteczności metod uczenia zespołowego

Na podstawie wykresu 1 można sformułować kilka istotnych wniosków dotyczących porównywanych klasyfikatorów. Najwyższą średnią dokładnością charakteryzuje się metoda boosting, która wyraźnie przewyższa zarówno bagging, jak i klasyfikator bazowy. Co więcej, dla boostingu nie zaobserwowano wartości odstających w estymacjach dokładności, co dodatkowo świadczy o jego wysokiej stabilności.

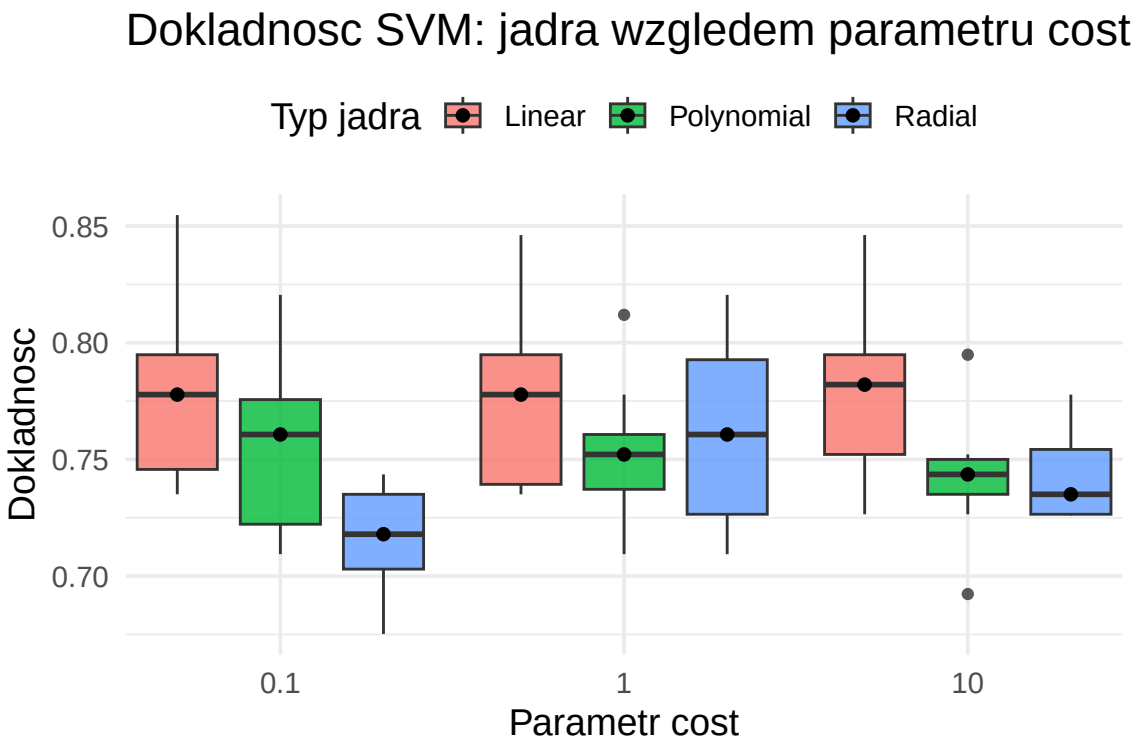
Metoda bagging okazała się nieznacznie mniej skuteczna niż klasyfikator bazowy pod względem średniej dokładności, co może być zaskakujące w kontekście ogólnej skuteczności metod zespołowych. Warto jednak zwrócić uwagę, że wariancje wszystkich trzech podejść są do siebie bardzo zbliżone, co potwierdzają dane zawarte w tabeli 1.

Tabela 1: Średnia dokładność i wariancja dla każdej z metod klasyfikacyjnych

	base	**bagging**	**boosting**
Accuracy	*0.2590*	*0.2256*	*0.7577*
Variance	*0.0022*	*0.0027*	*0.0025*

1.2 Metoda wektorów nośnych (SVM)

W niniejszym podrozdziale przeanalizowana zostanie skuteczność jednego z bardziej zaawansowanych algorytmów klasyfikacyjnych — metody wektorów nośnych (ang. Support Vector Machine, SVM). W celu zapewnienia wszechstronności analizy, uwzględniono wpływ różnych funkcji jądra (liniowego, wielomianowego oraz radialnego) na jakość klasyfikacji. Dodatkowo rozważono, w jakim stopniu zmiana wartości parametru kosztu (ang. cost) przekłada się na dokładność działania algorytmu. Takie podejście pozwala lepiej zrozumieć, jak poszczególne konfiguracje SVM wpływają na efektywność klasyfikatora w kontekście analizowanego zbioru danych.



Rysunek 2: Porównanie wyników klasyfikacji SVM dla różnych jąder i parametrów kosztu

Na podstawie wykresu 2 można zauważyć, że dla każdej rozważanej wartości parametru kosztu najwyższą medianę dokładności osiąga klasyfikator SVM z jądrem liniowym. Wskazuje to na jego dobrą skuteczność niezależnie od przyjętej wartości parametru regularyzacji.

Największą niestabilnością, rozumianą jako duży rozrzut wyników, cechuje się natomiast metoda wykorzystująca jądro wielomianowe stopnia drugiego. Co istotne, dokładność tej konfiguracji systematycznie spada wraz ze wzrostem wartości parametru kosztu, co może świadczyć o podatności na przeuczenie w przypadku zbyt wysokiej penalizacji błędów.

Z kolei najbardziej stabilne wyniki — przy stosunkowo niskiej zmienności — uzyskuje klasyfikator oparty na jądrze radialnym. Biorąc pod uwagę właśnie ten aspekt, SVM z jądrem RBF (radialnym) można uznać za najpewniejsze rozwiązanie pod względem powtarzalności działania.

1.3 Strojenie hiperparametrów modelu SVM

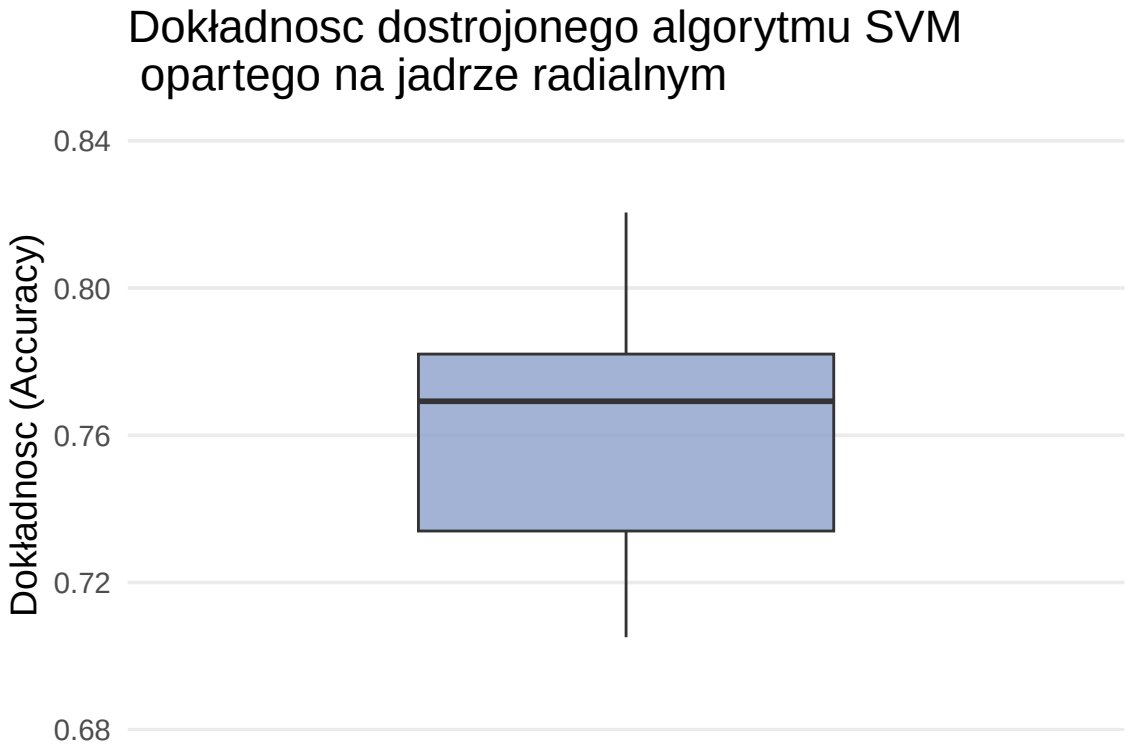
Dla klasyfikatora SVM z jądrem radialnym przeprowadzono proces strojenia parametrów: kosztu (C) oraz gamma, w celu oceny, czy dostrajanie prowadzi do istotnego wzrostu dokładności modelu oraz redukcji jego wariancji. Wykres 3 przedstawia wykres pudełkowy, który umożliwia wizualną ocenę zarówno stabilności, jak i skuteczności klasyfikatora po strojeniu.

Aby umożliwić bardziej obiektywną ocenę wpływu strojenia, porównano wyniki uzyskane przez model dostrojony z rezultatami modelu niedostrojonego — dla tych samych wartości

Tabela 2: Porównanie dokładności SVM z jądrem radialnym

Parametr C	Średnia Dokładność	Odchylenie Standardowe	Typ strojenia
0.1	0.7145	0.0246	untuned
1.0	0.7624	0.0390	untuned
10.0	0.7419	0.0197	untuned
100.0	**0.7603**	**0.0358**	**tuned**

parametru C, przy domyślnej wartości gamma. Wyniki tej analizy zostały zestawione w tabeli 2.



Rysunek 3: Porównanie dokładności i stabilności dostrojonego SVM opartego na jądrze radialnym

Jak pokazują wyniki zawarte w tabeli 2, średnia dokładność dostrojonego modelu SVM jest nieco niższa niż maksymalna dokładność osiągnięta wśród wszystkich konfiguracji przedstawionych w tabeli. Warto jednak zaznaczyć, że dostrojony model charakteryzuje się stosunkowo wysoką wariancją, co potwierdza konieczność podjęcia kompromisu między obciążonością modelu, a jego wariancją.

Ostateczny wybór modelu zależy więc od priorytetów — czy zależy nam bardziej na maksymalizacji średniej dokładności, czy może na uzyskaniu większej stabilności i przewidywalności wyników. Jest to klasyczny przykład kompromisu (ang. trade-off) pomiędzy skutecznością a

niezawodnością działania modelu.

1.4 Wybór najskuteczniejszego modelu

Wybór najskuteczniejszego modelu klasyfikacyjnego jest w dużej mierze kwestią subiektywną i zależy od przyjętego kryterium oceny. Można bowiem kierować się różnymi aspektami — takimi jak wysoka średnia dokładność klasyfikacji, bądź niska wariancja, która świadczy o stabilności modelu.

W niniejszej analizie jako główne kryterium przyjęto wysoką skuteczność klasyfikacyjną. W oparciu o to założenie, najlepszym rozwiązaniem okazała się metoda boosting, zastosowana w połączeniu z drzewem decyzyjnym jako klasyfikatorem bazowym. Zarówno jej skuteczność, jak i stabilność zostały przedstawione na wykresie 1.

2 Analiza skupień - algorytmy grupujące i hierarchiczne

2.1 Wybór i przygotowanie danych

Czy standaryzacja jest konieczna? Zaraz się pewnie okaże.

2.2 Wizualizacja wyników grupowania