

Sprawozdanie z listy 4

Eksploracja danych

Marta Stankiewicz (282244)

Paweł Nowak (282223)

2025-06-09

Spis treści

1	Zaawansowane metody klasyfikacji	1
1.1	Rodziny klasyfikatorów/uczenie zespołowe	1

Spis rysunków

1	Porównanie skuteczności metod uczenia zespołowego	2
---	---	---

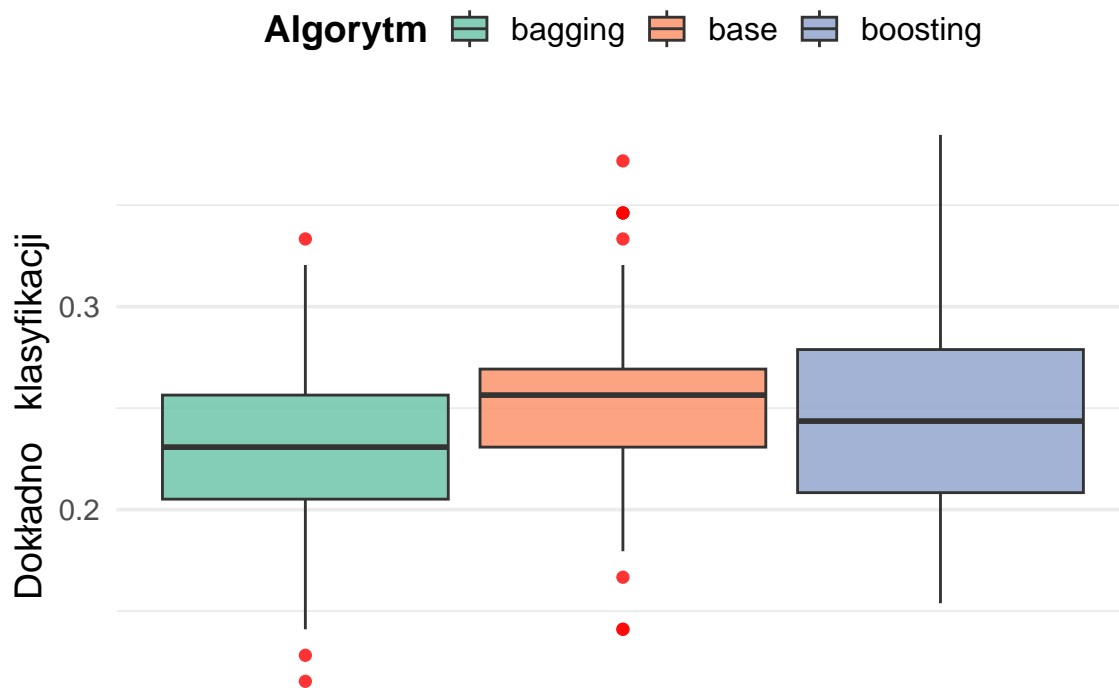
Spis tabel

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 r

n_iter 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

Z wykresu 1 możemy wysunąć następujące wnioski.