Naiwny Klasyfikator Bayesowski

Mikołaj Szawerda 318731

Opis polecenia

Zadanie polega na zaimplementowaniu nawinego klasyfikatora Bayesowskiego oraz zbadaniu jego działania w zastosowaniu do zbioru danych Iris Data Set. Działanie klasyfikatora polega na przypoprządkowaniu prawdopodobieństwa przynależności do klas dla danego zestawu wartości cech i wybraniu tego o największej wartości.Prawdopodobieństwo jest wyliczane z założeniem warunkowej niezależności cech - jest więc iloczynem prawdopodobieństw dla każdej cechy. Ponieważ cechy są typu ciągłego, oraz po przeprowadzeniu analizy rozkładu wartości, do wyliczenia potrzebnych wartości użyję gęstości rozkładu normalnego.

Algorytm realizuje następujący wzór:

$$\hat{y} = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}}\ p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k).$$

gdzie

$$p(x = v \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}}\, e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}},$$

$p(C_k)$ - prawdopodobieństwo klasy, zostało wyznaczone na podstawie liczności w zbiorze treningowym

Trening polega więc na wyznaczeniu μ_k i σ_k^2 dla każdej klasy i cechy z zbioru uczącego, a predykcja polega na wyliczeniu wartości dla każdej z możliwych klas i wybraniu tej najbardziej prawdopodobnej.

Planowane eksperymenty numeryczne

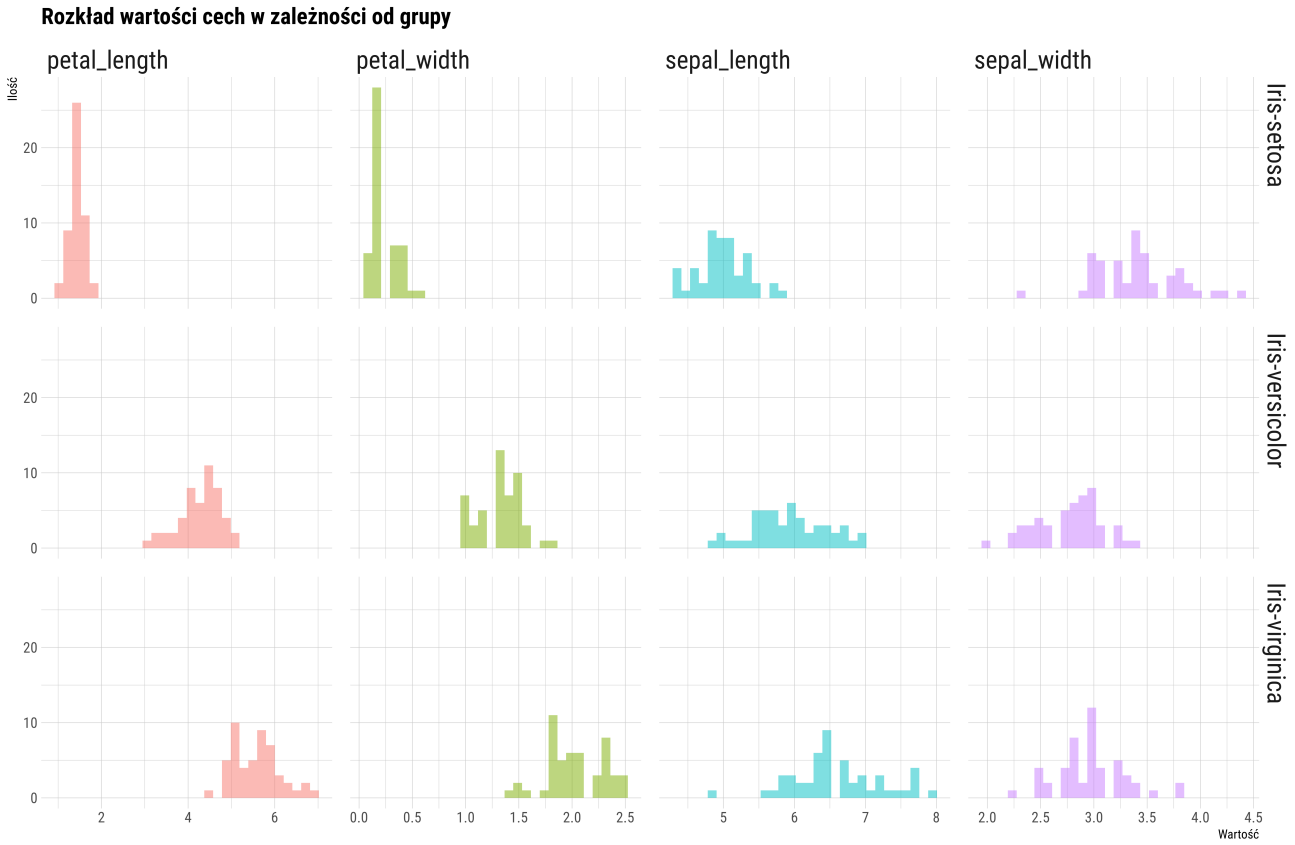
Przeprowadzę klasyfikację dla zbioru testowego, dla wytrenowanego klasyfikatora odpowiednio dla 10%,...,90% dostępnych danych jako zbiór uczący, oraz zbadam osiągniętą dokładność.

Wyniki

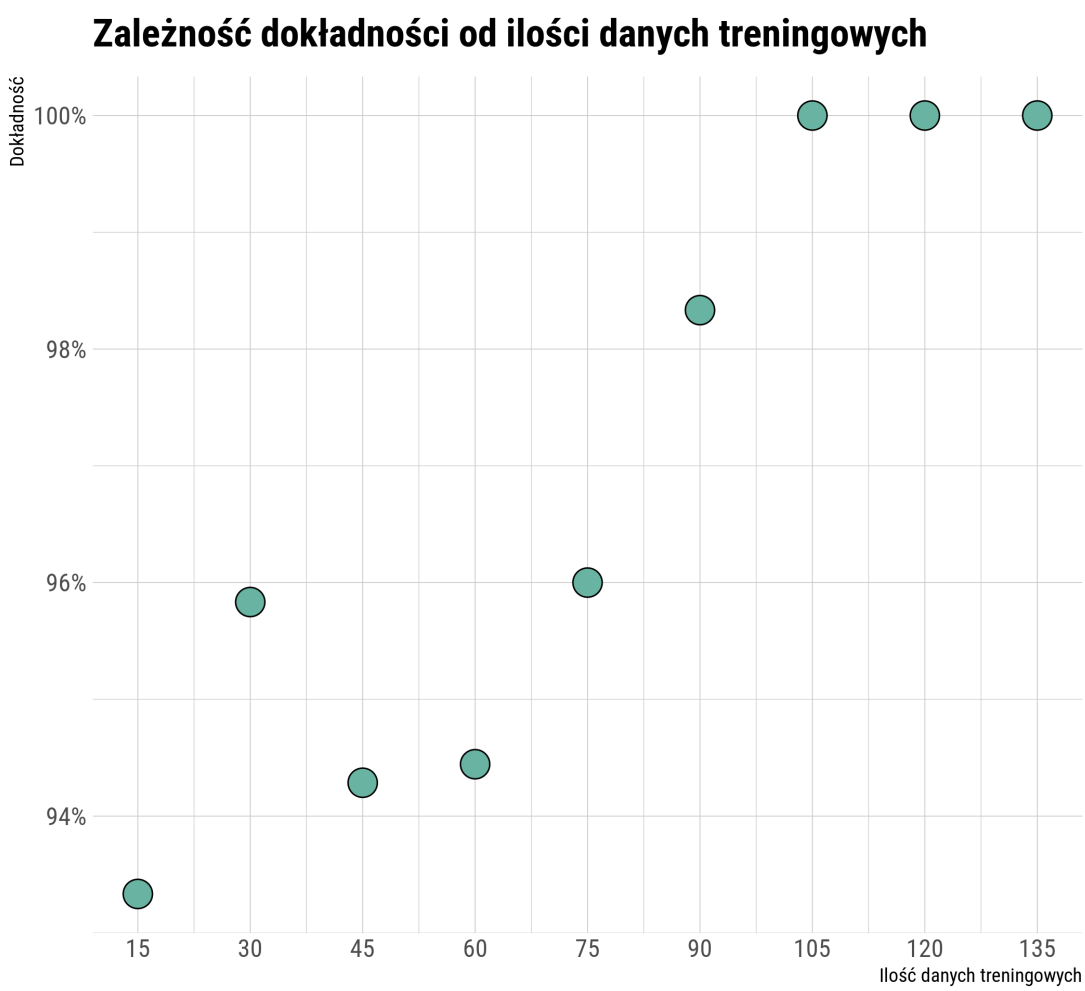
Średnia i wariancja cech w zależności od klasy

Characteristic	**Iris-setosa**, N = 50[1]	**Iris-versicolor**, N = 50[1]	**Iris-virginica**, N = 50[1]
sepal_length	5.01 (0.35)	5.94 (0.52)	6.59 (0.64)
sepal_width	3.42 (0.38)	2.77 (0.31)	2.97 (0.32)
petal_length	1.46 (0.17)	4.26 (0.47)	5.55 (0.55)
petal_width	0.24 (0.11)	1.33 (0.20)	2.03 (0.27)

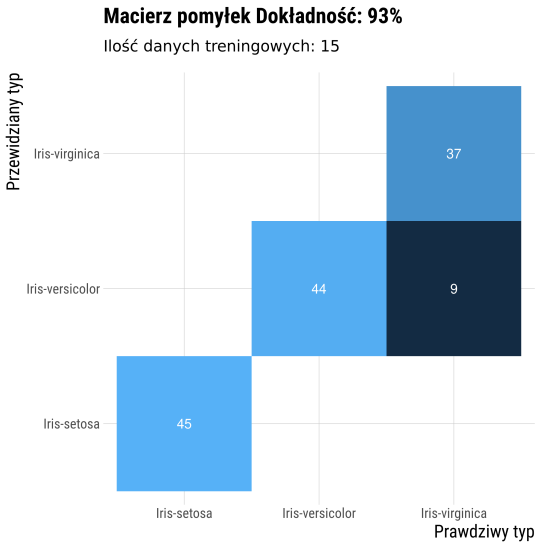
[1] Mean (SD)

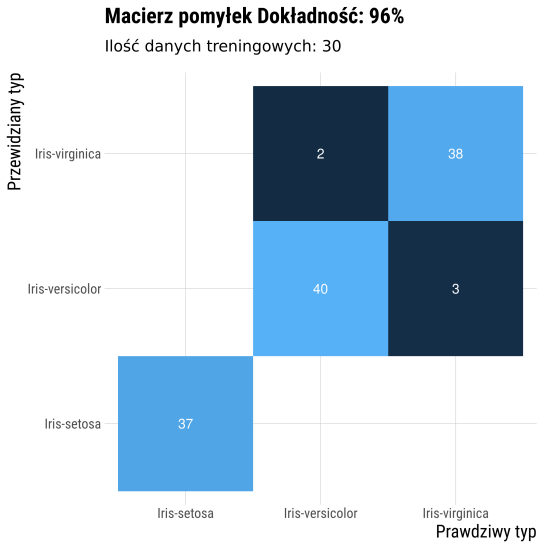


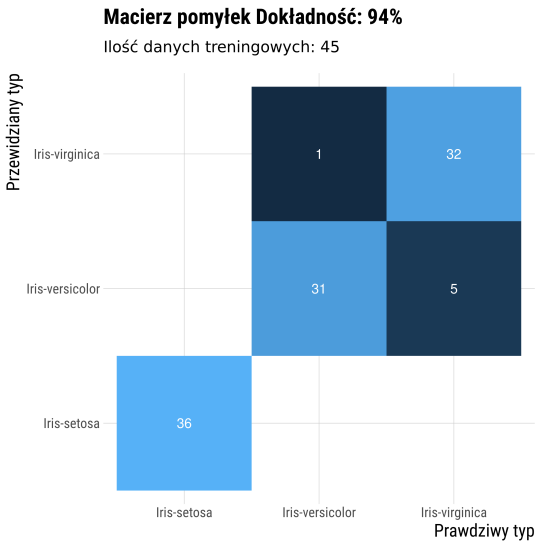
Można zauważyć, że praktycznie każda z cech ma w przybliżeniu rozkład normalny - przyjęcie gestości rozkładu normalnego ma więc swoje uzasadnienie. Można również zauważyć, że klasa "Iris-setosa" znacząco różni się od pozostałych dwóch, co może sugerować lepsze osiągi klasyfikacji dla tej klasy.

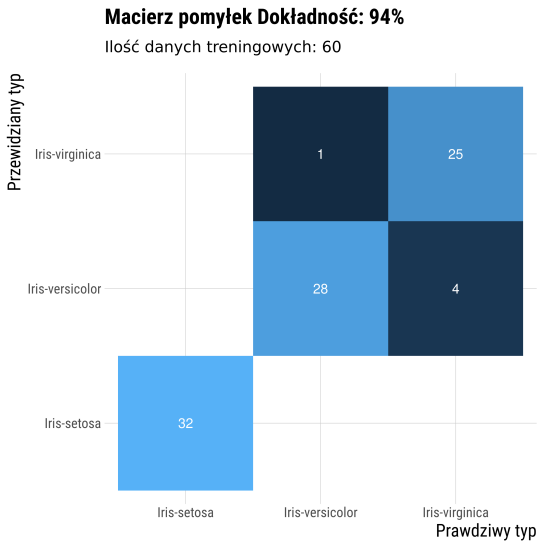


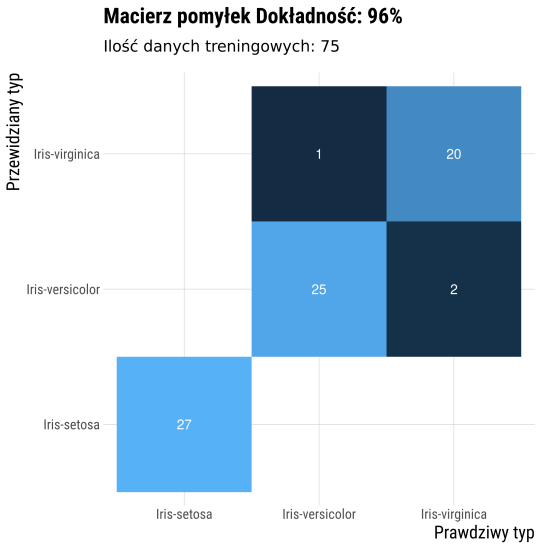
Wraz z ilością danych treningowych rośnie dokładność klasyfikacji - od stosunku $\frac{2}{3}$ danych treninogowych do testowych klasyfikator osiąga nieomylność



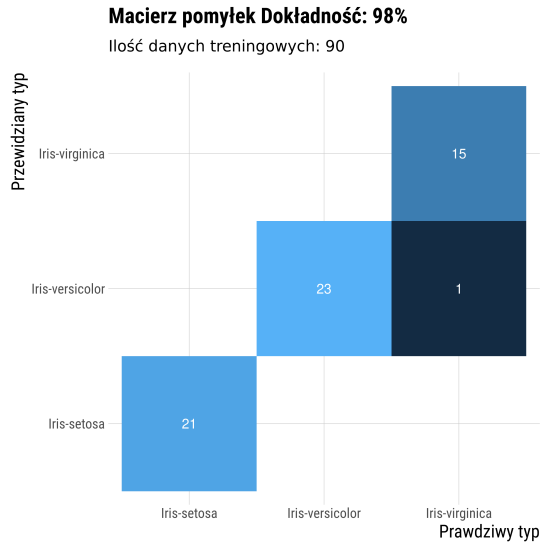


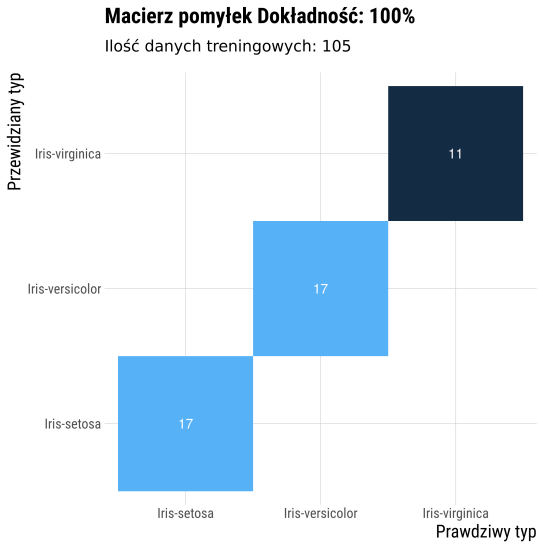


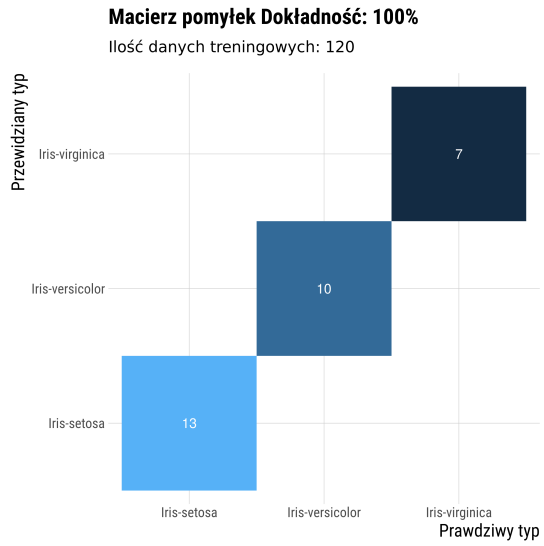


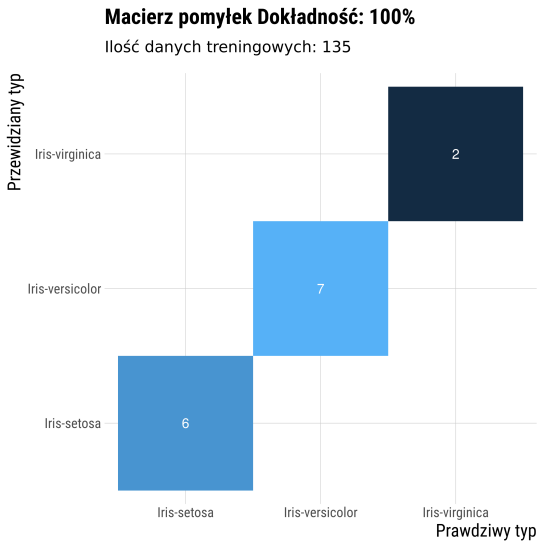


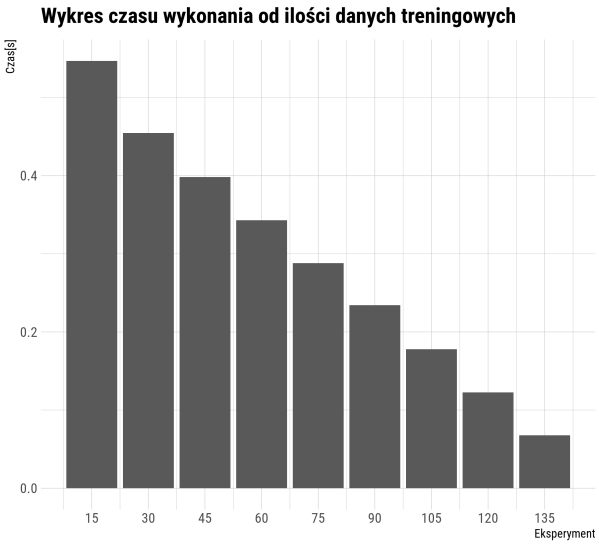
Pomyłki występują tylko pomiędzy klasami "Iris-virginica" i "Iris-versicolor" - co można było przewidzieć na podstawie wartości średniej i wariancji cech











Spadek czasu wykonania wraz z ilością danych treningowych wynika z wzrostu ilości danych testowych. Można zauważyć liniowy charakter algorytmu.

Wnioski

Naiwny klasyfikator Bayesowski dla zadanego zbioru danych osiągnął prawidłowe rezultaty. Należy zwrócić uwagę na potrzebną ilość danych do rozpoczęcia zwracania przez algorytm pożądanych rezultatów - już dla zbioru 15 przykładów algorytm osiągnął 93% dokładność. Fakt ten można wytłumaczyć normalnym rozkładem wartości cech, przez co wyliczone przybliżone prawdopodobieństwa przynależności były bliskie wartościom teoretycznym.

Przypadki w których algorytm dokonywał błędnej klasyfikacji są związane z podobnym rozkładem cech pomiędzy klasami - jednakże w raz z odpowiednią ilością danych błąd ten w najlepszych próbach osiągnął 0.

Złożoność czasowa algorytmu jest liniowa, natomiast w czasie wykonania algorytm potrzebuje tylko tablicy wartości średniej i wariancji dla każdej kombinacji klasa-cecha