Sprawozdanie z listy 2

Eksploracja danych

Marta Stankiewicz (282244)
Paweł Nowak (282223)

2025-04-04

Spis treści

1 Ocena zdolności separacyjnych zmiennych, dyskretyzacja zmiennych ciągłych. 1
1.1 Ocena zdolności dyskryminacyjnych zmiennych ciągłych. 1
1.2 Porównanie róznych metod dyskretyzacji nienadzorowanej. 3

Spis rysunków

1 Wykresy skrzypcowo-pudełkowe dla zmiennych ciągłych 2

Spis tabel

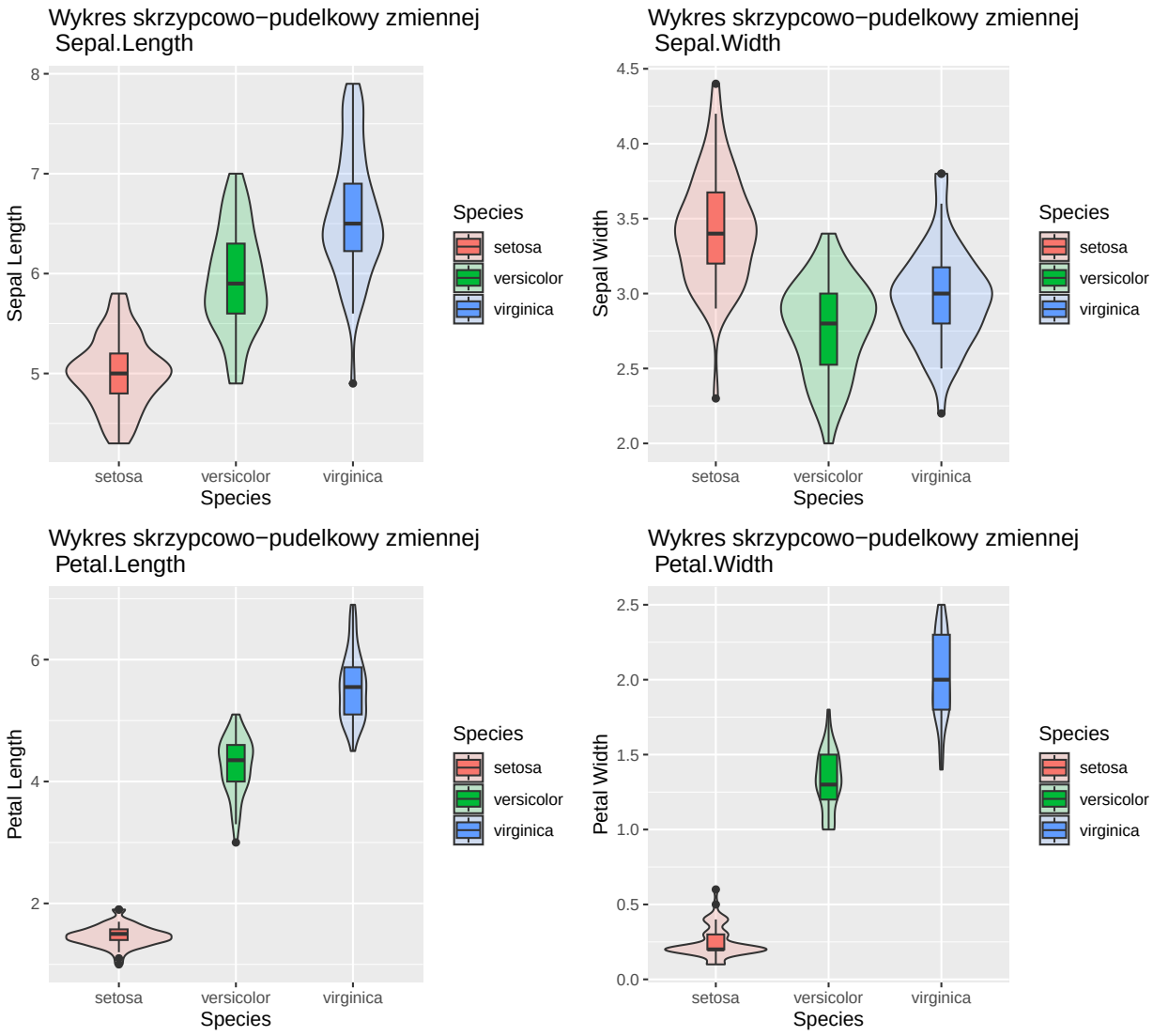
1 Skuteczność wybranych metod dyskretyzacji dla zmiennej Sepal Width . . . 3
2 Skuteczność wybranych metod dyskretyzacji dla zmiennej Petal Width . . . 3

---

1 Ocena zdolności separacyjnych zmiennych, dyskretyzacja zmiennych ciągłych.

1.1 Ocena zdolności dyskryminacyjnych zmiennych ciągłych.

W celu zbadania zdolności dyskryminacyjnej cech, posłużymy się wykresem skrzypcowo-pudełkowym (tj. wykresem skrzypcowym wraz z wykresem pudełkowym).



Rysunek 1: Wykresy skrzypcowo-pudełkowe dla zmiennych ciągłych

Z wykresów 1 wnioskujemy, że największe zdolności dyskryminacyjne wykazuje zmienna *Petal. Width*. Z kolei najmniejsze zdolności do separacji gatunków obserwujemy u zmiennej *Sepal. Width*.

1.2 Porównanie róznych metod dyskretyzacji nienadzorowanej.

Dla wymienionych wyżej zmiennych (tj. *Petal.Width* oraz *Sepal.Width*) zastosujemy teraz różne techniki przedziałowania (dyskretyzacji) według, odpowiednio, **stałej szerokości** przedziału, **równej częstości**, **algorytmu K-średnich**, **stałych granicach** przedziałów ustalonych przez użytkownika.

1.2.1 Metodologia oceny skuteczności dyskretyzacji.

Aby ocenić skuteczność każdej ze wspomnianych metod, przyjęliśmy następującą metodologię. Najpierw dokonaliśmy przedziałowania każdego przypadku, korzystając ze wszystkich metod, a następnie wybraliśmy tę klasę, która występuje najczęściej (w przypadku tzw. "remisu" wybierana jest dowolna klasa). Następnie sprawdzaliśmy, w ilu przypadkach wynik przedziałowania każdej metody zgadzał się ze zagregowaną klasą. Tę liczbę podzieliliśmy przez liczbę wszystkich przypadków, aby uzyskać skuteczność metody dyskretyzacji wyrażoną w procentach. Porównanie róznych metod przedziałowania zostały przedstawione poniżej

Tabela 1: Skuteczność wybranych metod dyskretyzacji dla zmiennej Sepal Width

Równomierna częstotliwość	Równomierne przedziały	Dyskretyzacja oparta na K-średnich	Stałe granice przedziału
79.33	93.33	72	84.67

Tabela 2: Skuteczność wybranych metod dyskretyzacji dla zmiennej Petal Width

Równomierna częstotliwość	Równomierne przedziały	Dyskretyzacja oparta na K-średnich	Stałe granice przedziału
97.33	100	98.67	86

1.2.2 Wnioski dotyczące skuteczności przedziałowania.

Z tabel **??** oraz **??** możemy wywnioskować, że w obu przypadkach największą skutecznością charakteryzuje się metoda dyskretyzacji oparta na **algorytmie K-średnich**. Z kolei najgorszą skuteczność przedziałowania obserwujemy dla metody opartej na **stałych granicach** przedziału. Wyniki dyskretyzacji zastosowanej dla zmiennej *Petal.Width* znaczącą rożnią się od wyników przedziałowania zastosowanego dla atrybutu *Sepal.Width*. Jest to zgodne z intuicją — jak wykazaliśmy wcześniej, najgorsze zdolności separacyjne klas wykazuje właśnie zmienna **Sepal.Width**, co znacząco wpływa na niską skuteczność metod przedziałowania. Analogiczna zależność występuje w przypadku cechy **Petal.Width**, która z kolei charakteryzowała się wysokimi zdolnościami dyskryminacyjnymi, co przełożyło się na wysoką dokładność podejść dyskretyzacji.