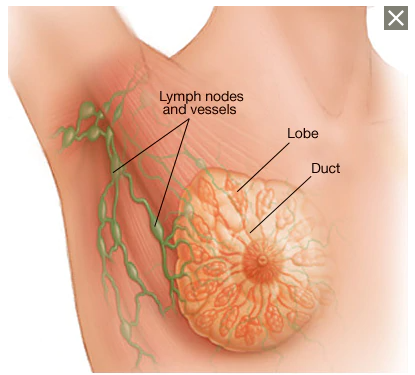
Sprawozdanie do projektu nr 2

Porównanie klasyfikatorów na przykładzie bazy

Breast Cancer Winconsin

(Rak Piersi Winconsin)



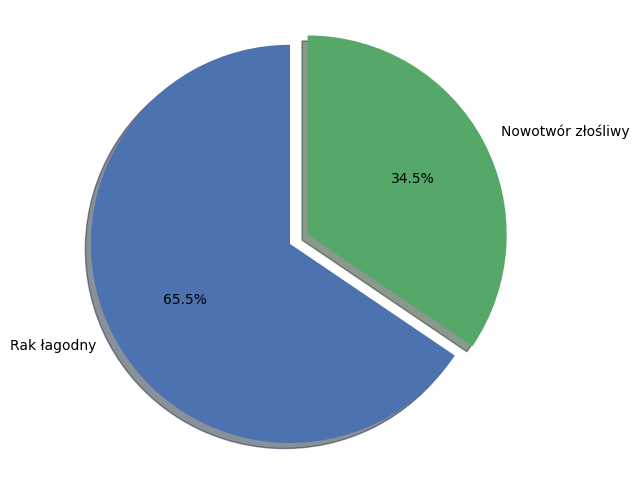
Baza danych: <https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)>

Kod projektu: github.com/Saafine/breast-cancer-data-analysis

1. **Wstęp**
   1. Podstawowe informacje o kolumnach

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Kolumna** | **Min** | **Max** | **Średnia** | **Mediana** | **% brakujących danych** |
| Grubość guza (Clump Thickness) | 1 | 10 | 4.42 | 4.42 | 0 |
| Jednorodność wielkości komórek (Uniformity of Cell Size) | 1 | 10 | 3.13 | 3.13 | 0 |
| Jednorodność kształtu komórek (Uniformity of Cell Shape) | 1 | 10 | 3.21 | 3.21 | 0 |
| Adhezja (Marginal Adhesion) | 1 | 10 | 2.81 | 2.81 | 0 |
| Rozmiar pojedynczej komórki nabłonka (Single Epithelial Cell Size) | 1 | 10 | 3.22 | 3.22 | 0 |
| Jądro - nagie (Bare Nuclei) | 1 | 10 | 3.54 | 3.54 | 2.28 |
| Chromatyna (Bland Chromatin) | 1 | 10 | 3.44 | 3.44 | 0 |
| Jądro - normalne (Normal Nuclei) | 1 | 10 | 2.87 | 2.87 | 0 |
| Mitozy (Mitoses) | 1 | 10 | 1.59 | 1.59 | 0 |
| Klasyfikacja (Class):  2 - rak łagodny,  4 – nowotwór złośliwy |  |  |  |  |  |

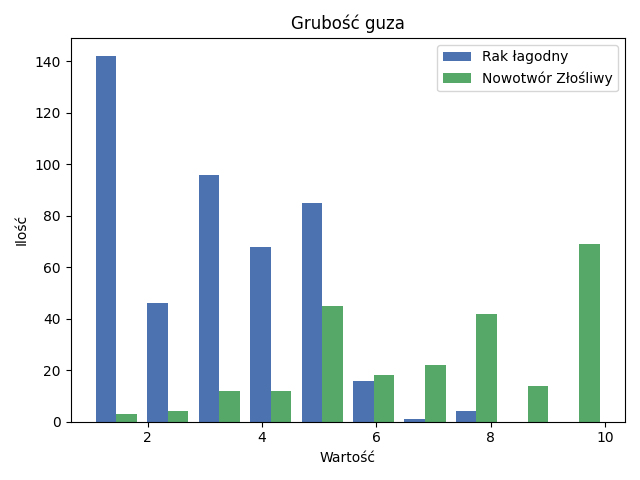
* 1. Częstość występowania poszczególnych klasyfikacji (diagnoz)

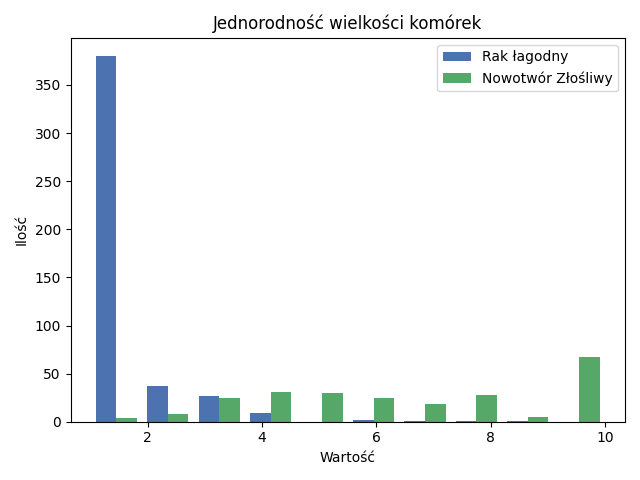


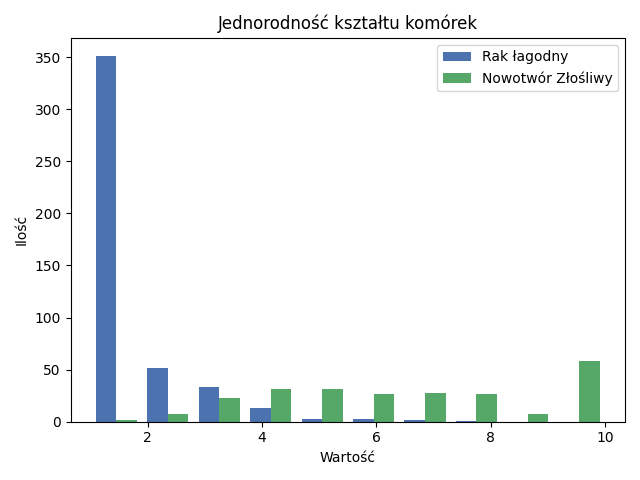
- Rak łagodny: 458 (65.5%)

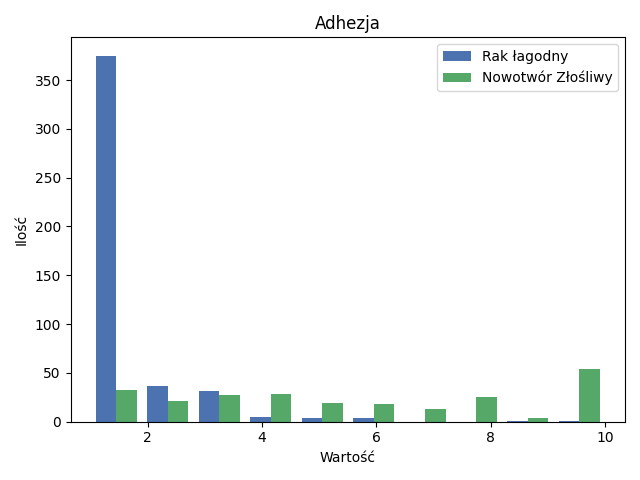
- Nowotwór złośliwy: 241 (34.5%)

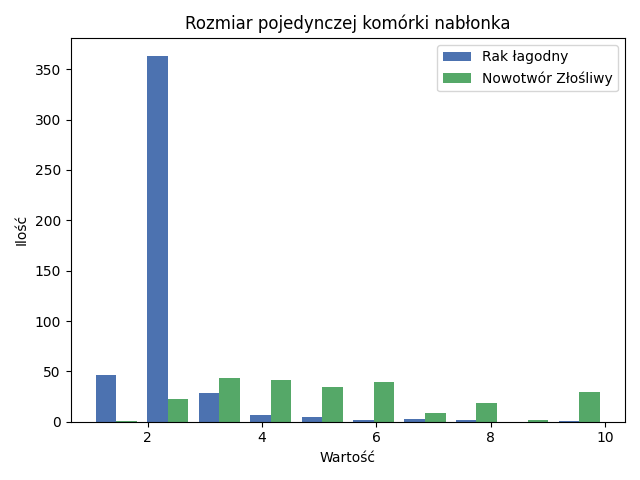
* 1. Częstość występowania poszczególnych odpowiedzi w kolumnach:

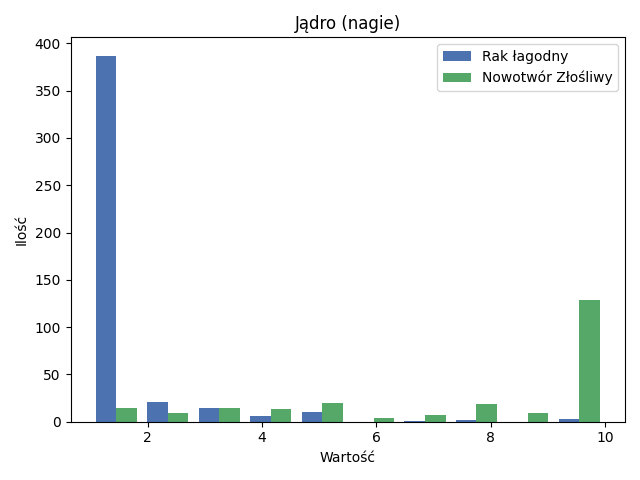


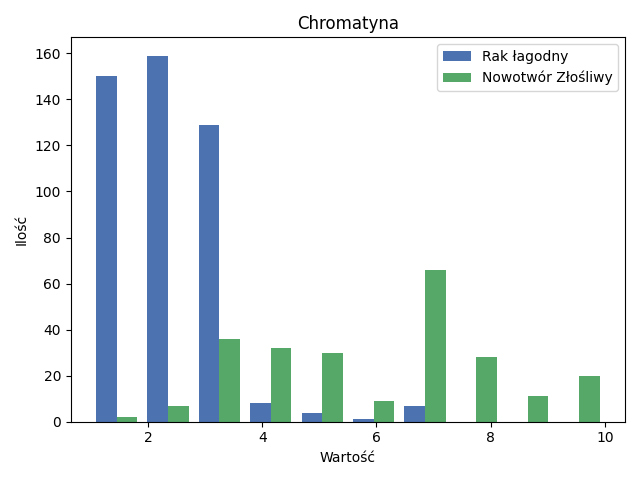


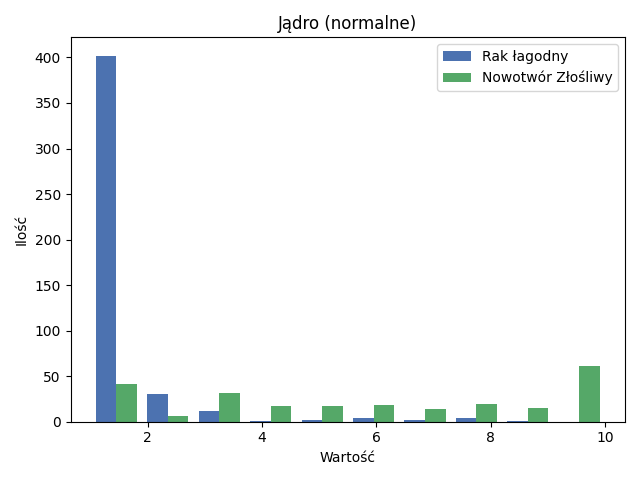


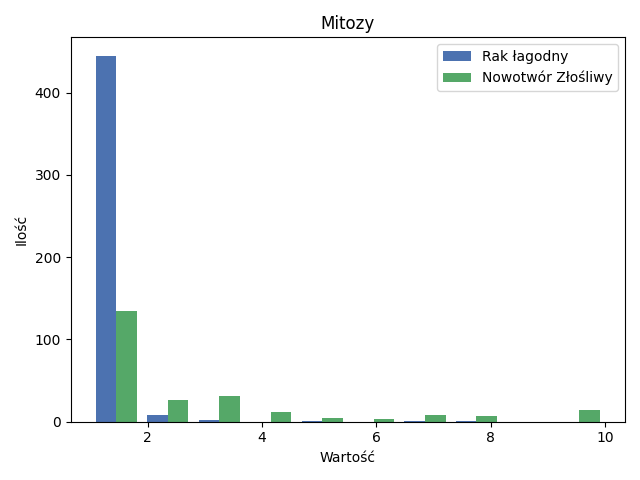




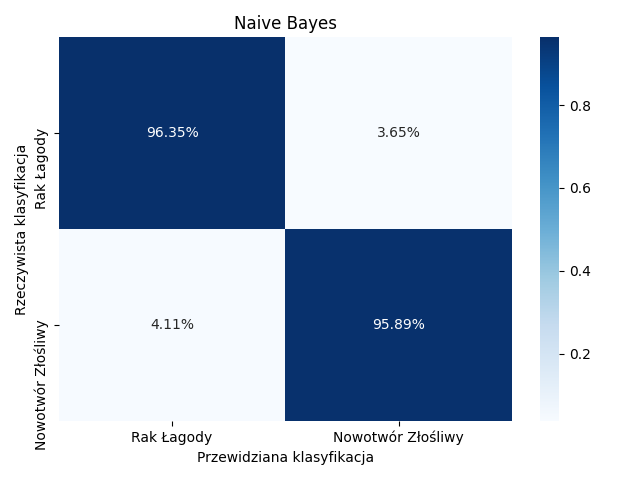




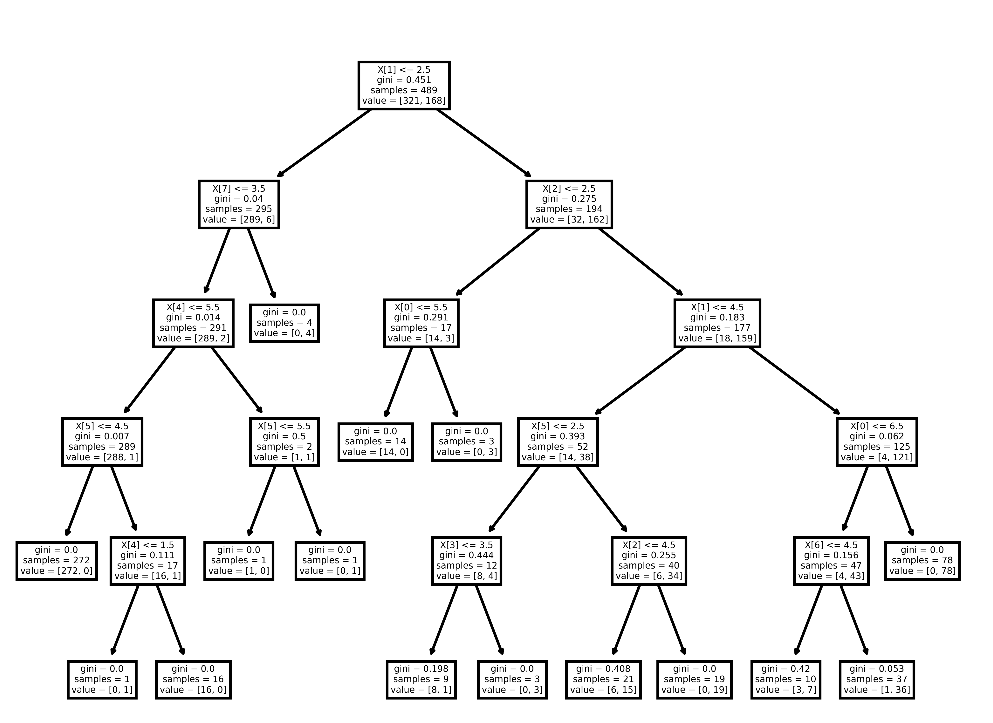


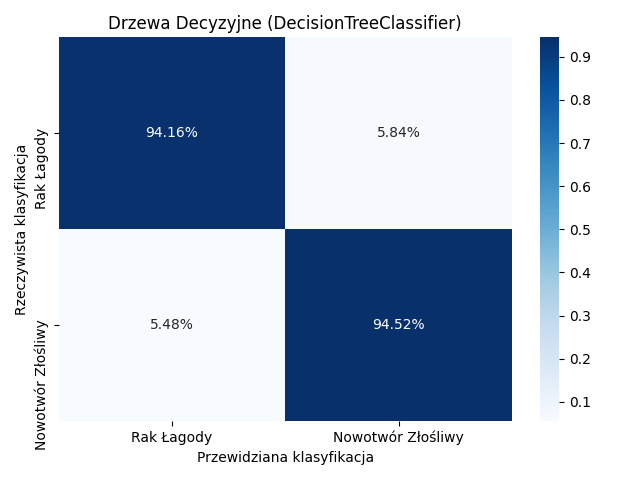


1. **Skuteczność klasyfikatorów**
   1. **Naive Bayes – 96,19 %**

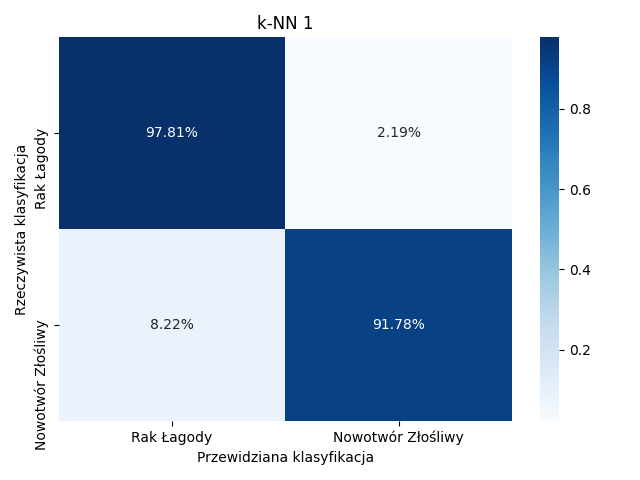


* 1. **Drzewa decyzyjne – 95.71%**

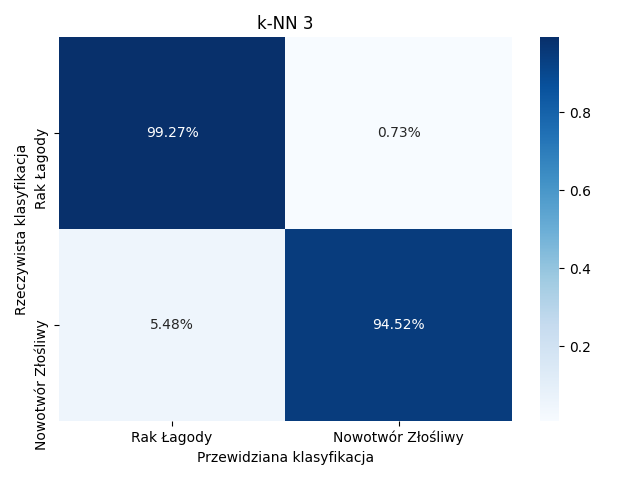




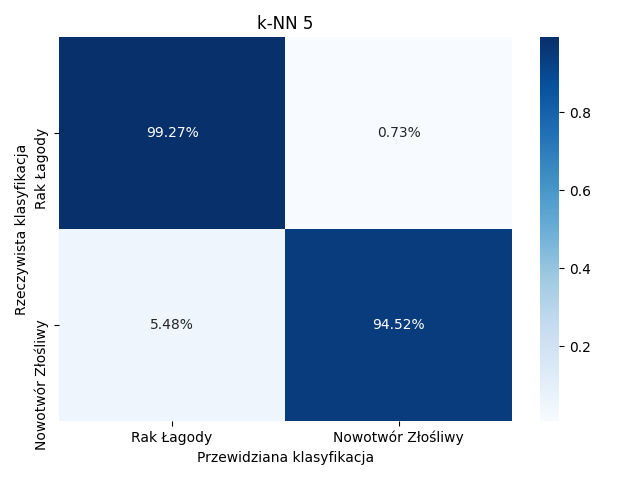
* 1. **k-Najbliższych sąsiadów**
     + k-NN-1 – **95.71%**



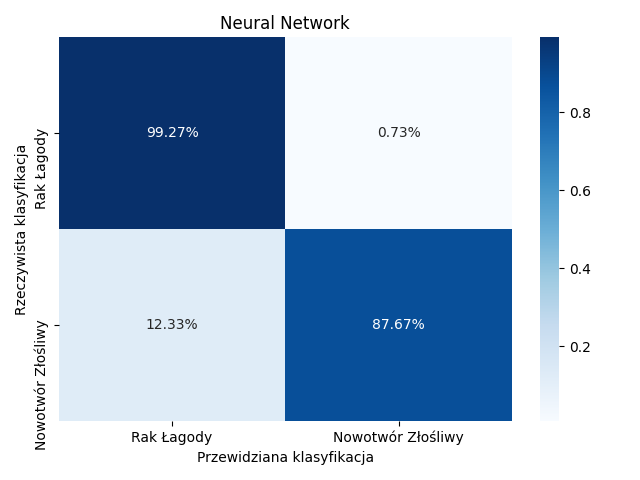
* + - k-NN-3 – **97.62%**



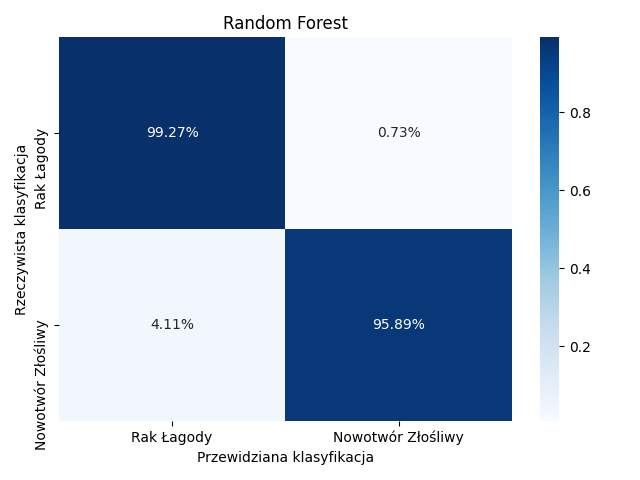
* + - k-NN-5 – **97.62%**



* 1. **Sieci neuronowe – 99,27%** (dla wybranej topologii lub kilku na wyższą ocenę)

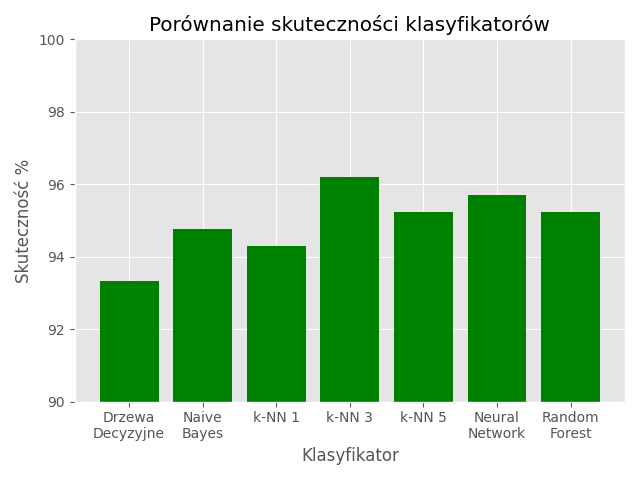


* 1. **Random Forest – 98.1% opis**

****

* 1. + 2-3 inne klasyfikatory (Support Vector Machines, Metody typu Ensemble lub inne). Dla innych zrób krótki wstęp teoretyczny, wyjasnij na jakiej zasadzie działają

1. **Porównanie skuteczności klasyfikatorów**



1. **Wnioski**

Najlepiej działający klasyfikator

Nie usunięcie kolumny „id” zmniejsza skuteczność klasyfikatorów do rzutu monetą / lub klasy

Podmienienie brakujących danych na -999, 0, lub medianę wartości kolumny nie wpływa na skuteczność klasyfikatorów

- dla wybranej bazy danych testowane klasyfikatory osiągały zbliżoną skuteczność

• Rozszerzona ewaluacja klasyfikatorów. Jak inaczej można oceniać

klasyfikatory? Które miary będą miały sens w Twoich badaniach? Na

początek można rzucić okiem na:

https://en.wikipedia.org/wiki/Sensitivity\_and\_specificity i poszukać innych

źródeł rozwijających temat.

• Szukanie reguł asocjacyjnych. Czy ma sens? Jakich najlepiej szukać? Podaj

te dla nas szczególnie interesujące.

• Porównanie skuteczności klasyfikatorów na jakichś wykresach. Słupkowy?

ROC?

• Czy w naszej bazie danych jakieś rodzaje błędów są ważniejsze /

poważniejsze niż inne? Dlaczego?