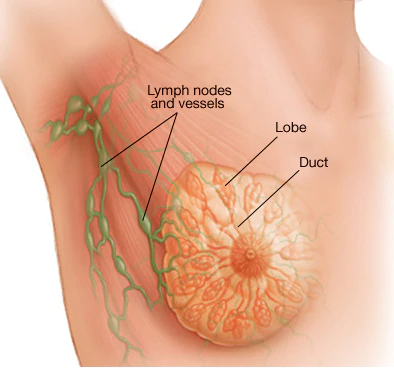
Sprawozdanie do projektu nr 2

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

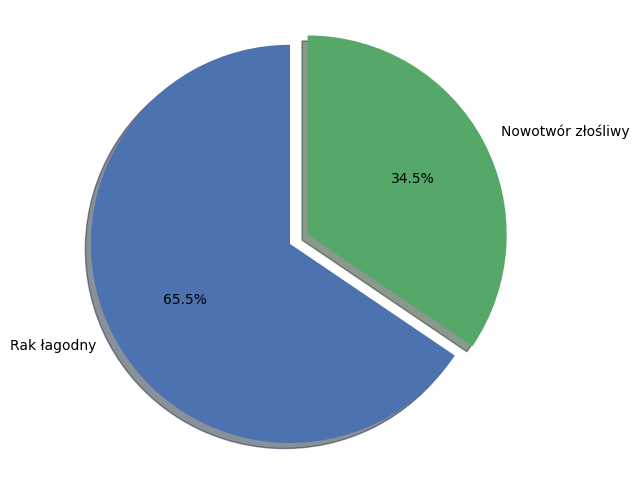
1. **Wstęp**
   1. Podstawowe informacje o kolumnach

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Kolumna** | **Min** | **Max** | **Średnia** | **Mediana** | **% brakujących danych** |
| Grubość guza (Clump Thickness) | 1 | 10 | 4.42 | 4.0 | 0 |
| Jednorodność wielkości komórek (Uniformity of Cell Size) | 1 | 10 | 3.13 | 1.0 | 0 |
| Jednorodność kształtu komórek (Uniformity of Cell Shape) | 1 | 10 | 3.21 | 1.0 | 0 |
| Adhezja (Marginal Adhesion) | 1 | 10 | 2.81 | 1.0 | 0 |
| Rozmiar pojedynczej komórki nabłonka (Single Epithelial Cell Size) | 1 | 10 | 3.22 | 2.0 | 0 |
| Jądro - nagie (Bare Nuclei) | 1 | 10 | 3.54 | 1.0 | 2.28 |
| Chromatyna (Bland Chromatin) | 1 | 10 | 3.44 | 3.0 | 0 |
| Jądro - normalne (Normal Nuclei) | 1 | 10 | 2.87 | 1.0 | 0 |
| Mitozy (Mitoses) | 1 | 10 | 1.59 | 1.0 | 0 |
| Klasyfikacja (Class):  2 - rak łagodny,  4 – nowotwór złośliwy |  |  |  |  |  |
| Ilość brakujących wartości | 16 | | | | |
| Ilość wierszy | 699 | | | | |

Każdy z atrybutów jest oceniany w skali od 1 do 10, gdzie 1 oznacza wartość najbardziej odpowiadającą łagodnemu nowotworowi, a 10 złośliwemu.

Baza danych nie zawiera błędnych danych. Jedyną kolumną brakującymi danymi jest Bare Nuclei.

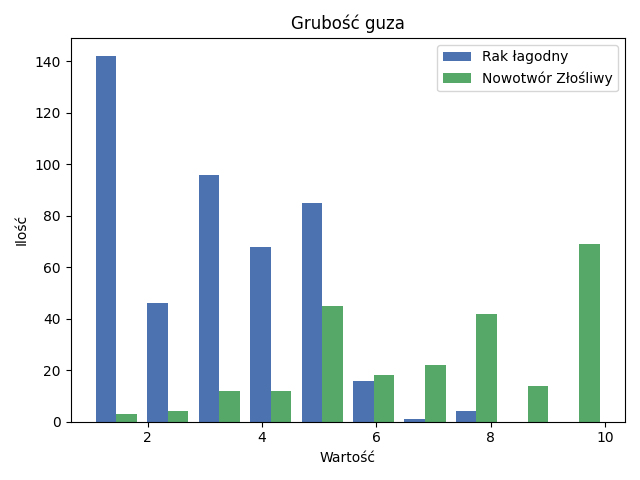
* 1. Częstość występowania poszczególnych klasyfikacji (diagnoz)

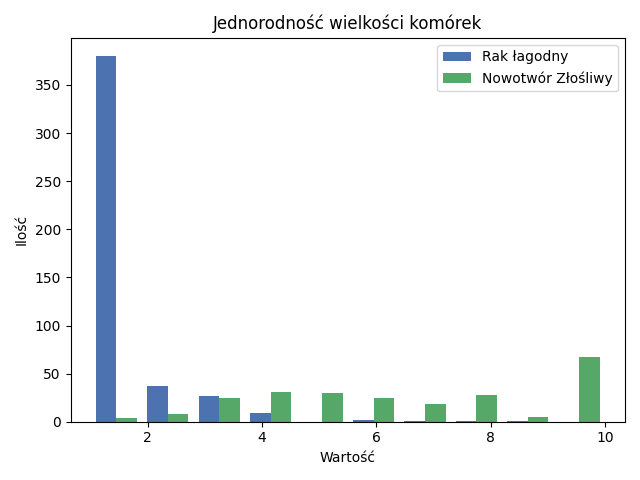


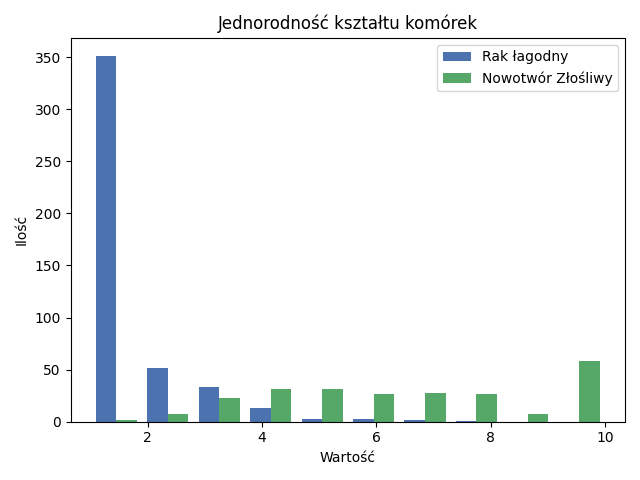
- Rak łagodny: 458 (65.5%)

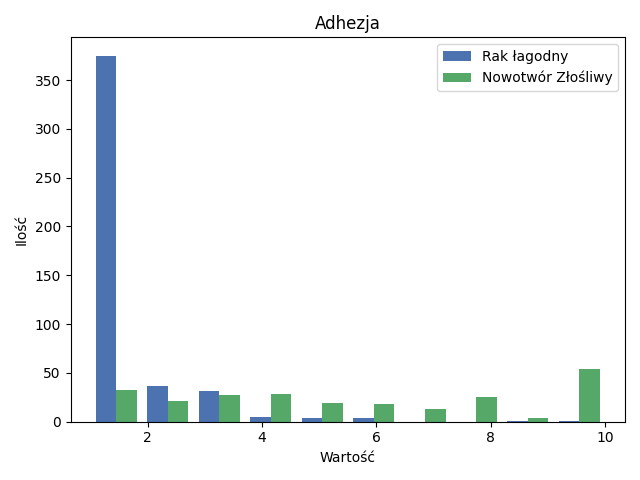
- Nowotwór złośliwy: 241 (34.5%)

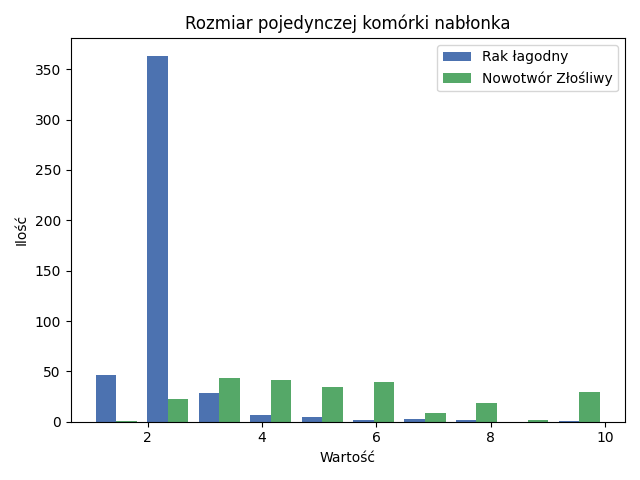
* 1. Częstość występowania poszczególnych odpowiedzi w kolumnach:

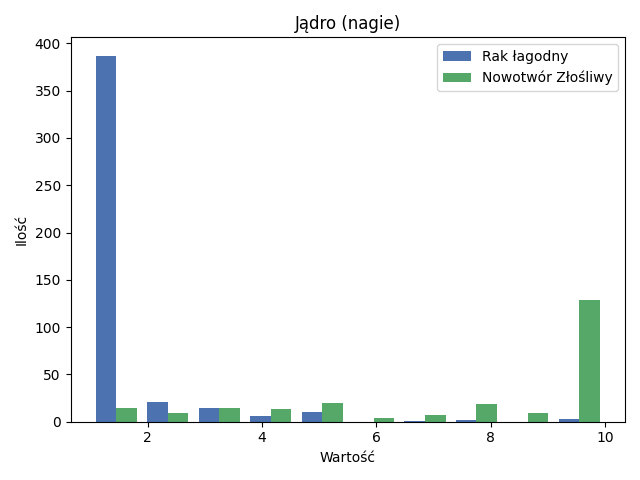


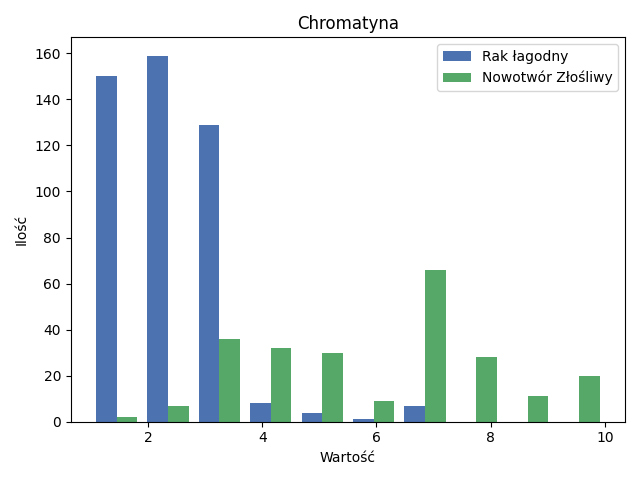


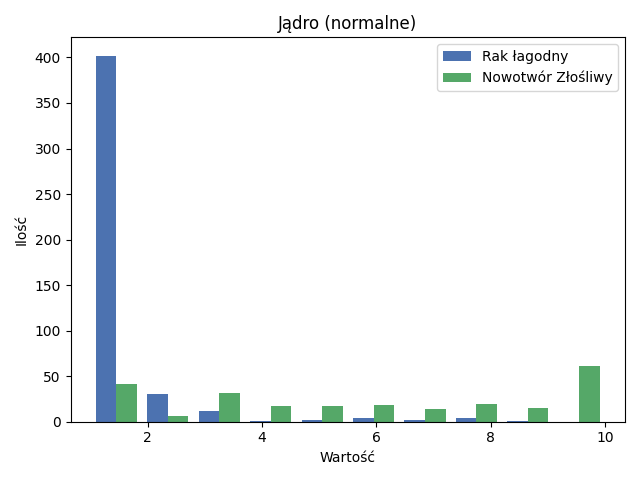


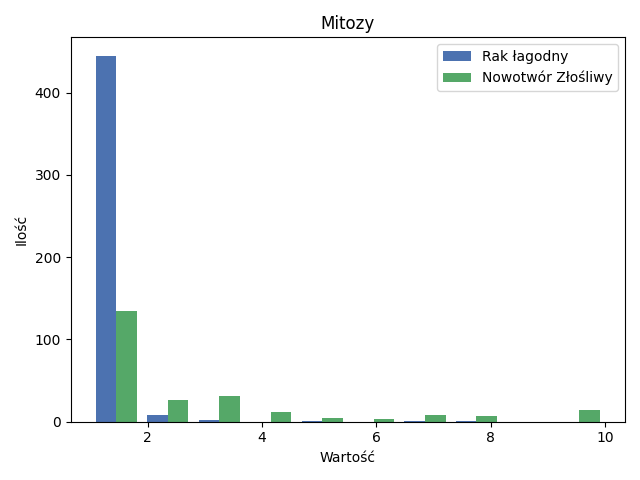




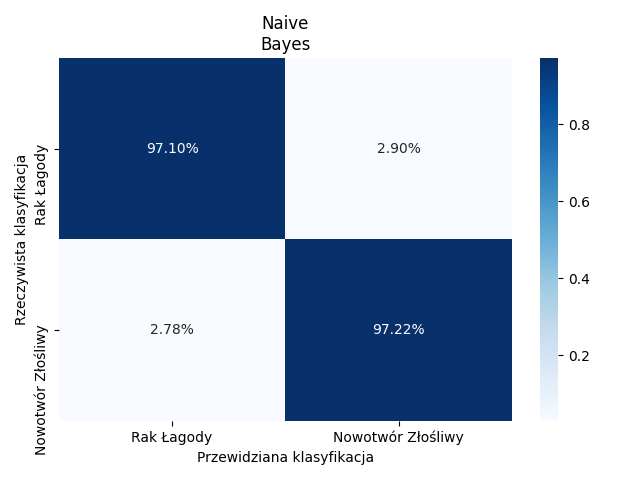




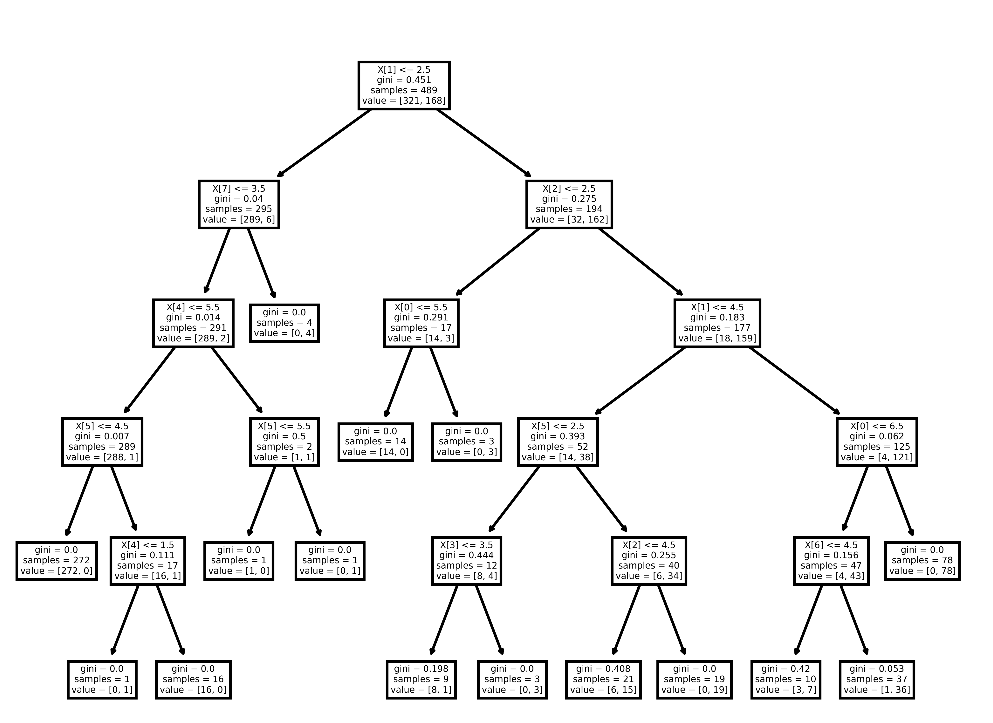


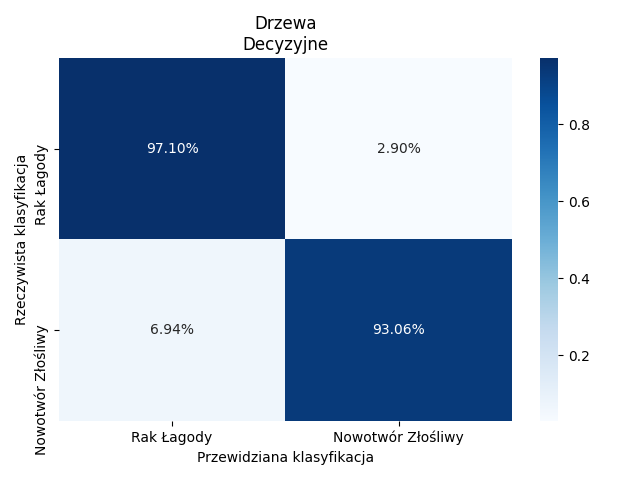


1. **Skuteczność klasyfikatorów**
   1. **Naive Bayes – 97,14 %**

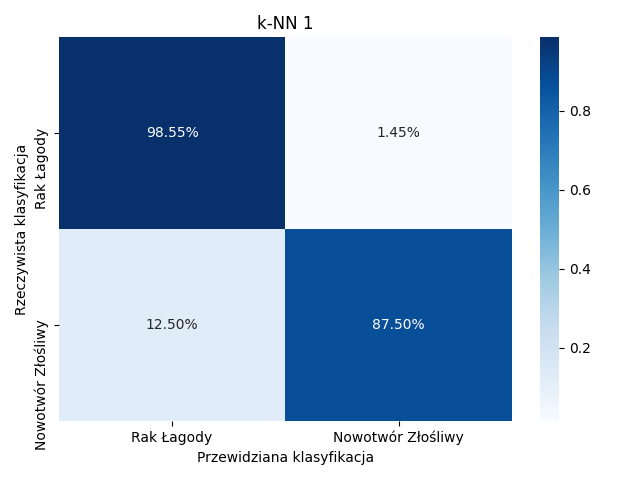


* 1. **Drzewa decyzyjne – 95.71%**

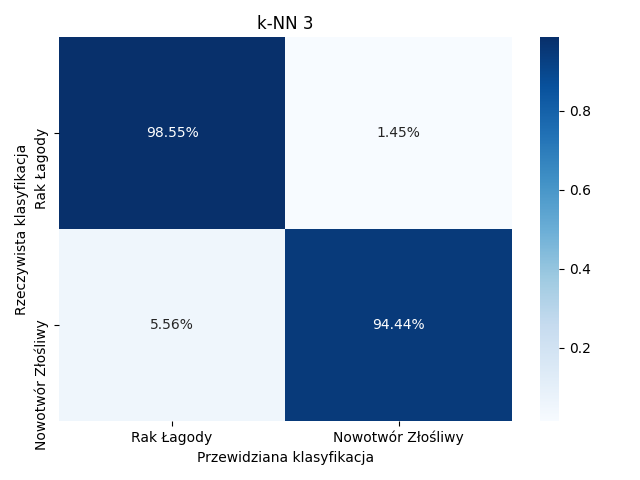




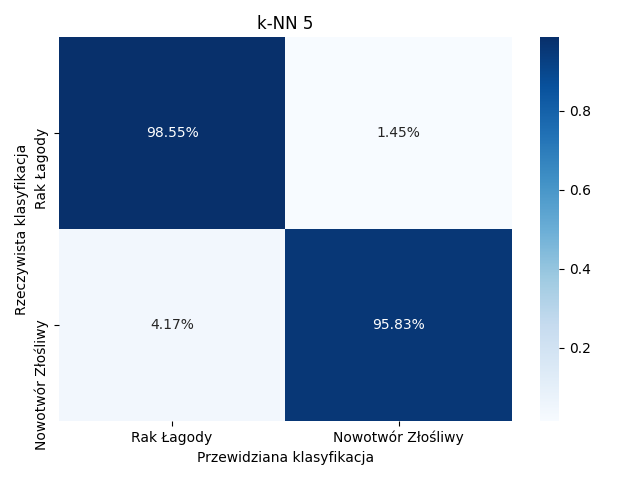
* 1. **k-Najbliższych sąsiadów**
     + k-NN-1 – **94.76%**



* + - k-NN-3 – **97.14%**

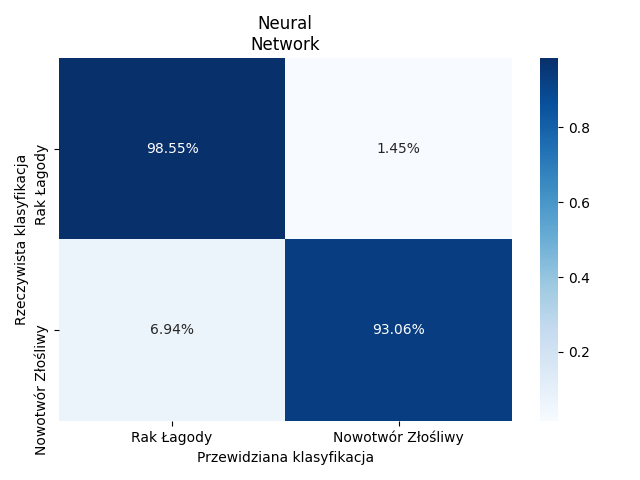


* + - k-NN-5 – **97.62%**



* 1. **Sieci neuronowe**

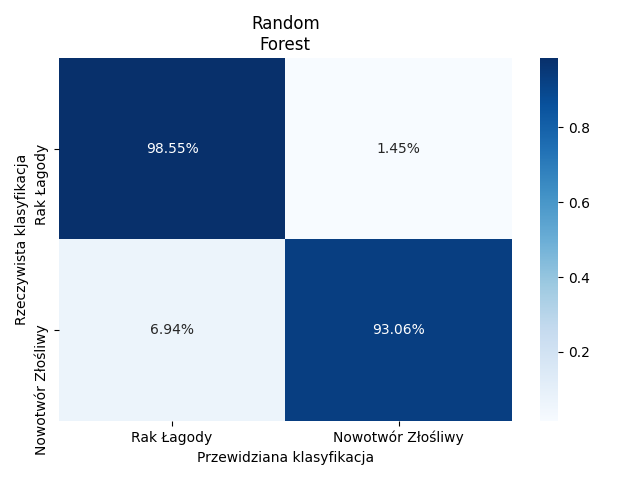
**Sklearn - MLPClassifier - 96,67%**



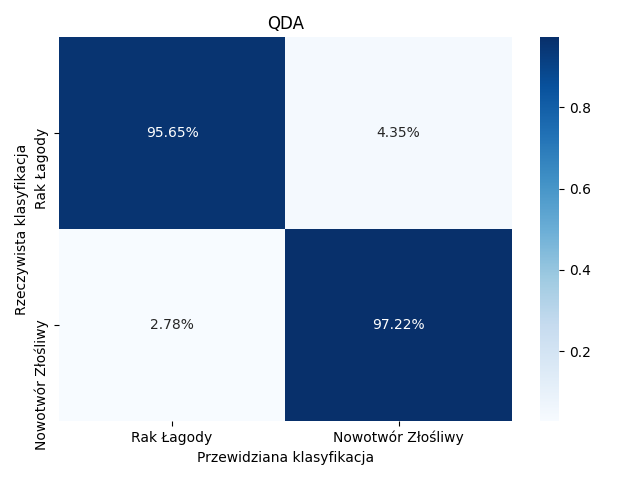
**TensorFlow - Keras Sequential** – **95,7%**

* 1. **Random Forest – 96.67%**

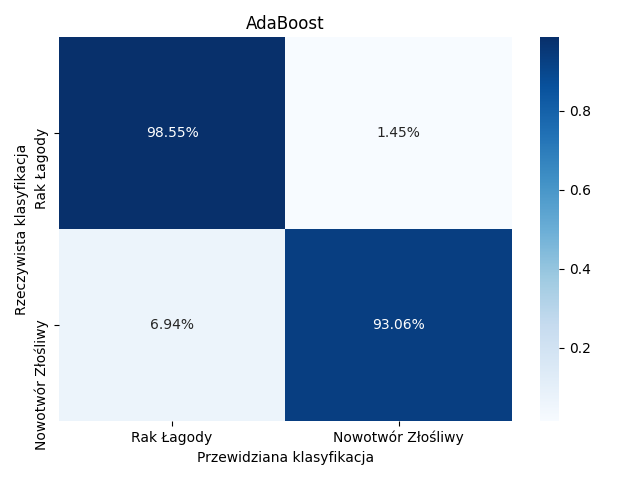
metoda zespołowa uczenia maszynowego dla klasyfikacji, regresji i innych zadań, która polega na konstruowaniu wielu drzew decyzyjnych w czasie uczenia i generowaniu klasy, która jest dominantą klas (klasyfikacja) lub przewidywaną średnią (regresja) poszczególnych drzew.

****

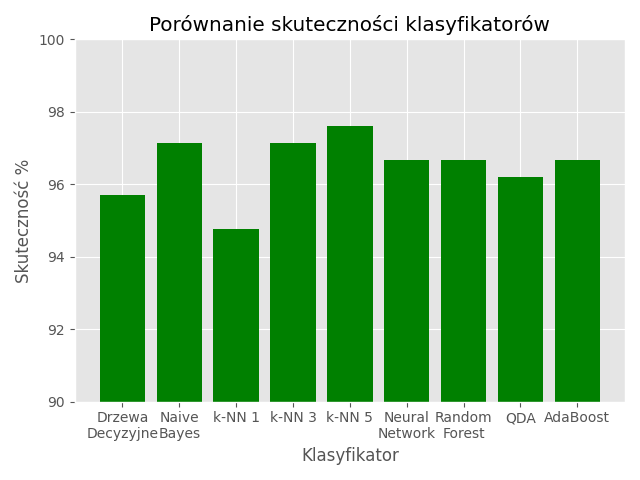
* 1. **Kwadratowa analiza dyskryminacyjna (QDA) – 96.19%**

****

* 1. **AdaBoost – 96.67%**



1. **Porównanie skuteczności klasyfikatorów**



1. **Wnioski**

- wszystkie klasyfikatory uzyskały wyniki większe niż 94%

- skuteczność była różna w zależności od podziału danych testowych

- wielkość guza, jednorodność kształtu i wielkości miały największe znaczenie przy klasyfikacji nowotworu jako łagodnego lub złośliwego