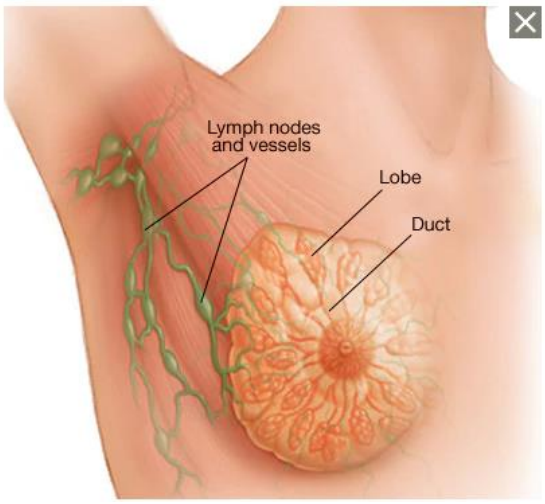
Sprawozdanie do projektu nr 2

Porównanie klasyfikatorów na przykładzie bazy Breast Cancer Winconsin

(Rak Piersi Winconsin)



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

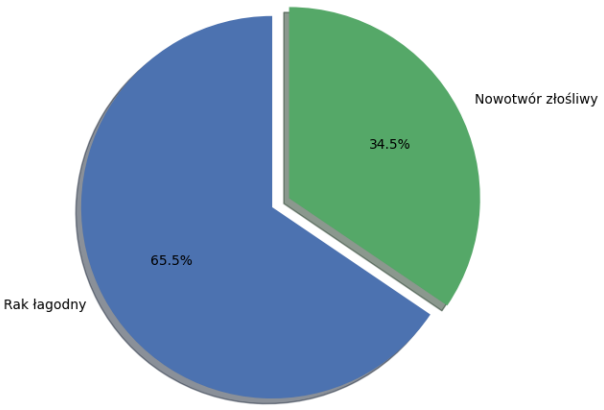
Kod projektu: github.com/Saafine/breast-cancer-data-analysis

1. Wstęp

a. Podstawowe informacje o kolumnach

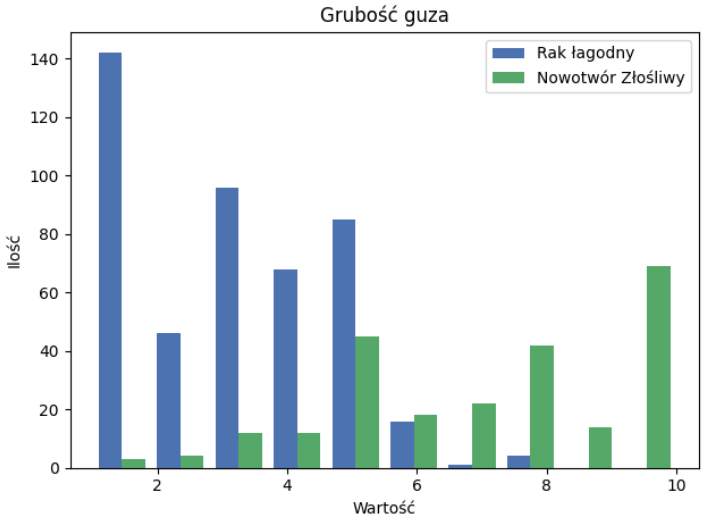
Kolumna	Min	Max	Średnia	Mediana	% brakujących danych
Grubość guza (Clump Thickness)	1	10	4.42	4.42	0
Jednorodność wielkości komórek (Uniformity of Cell Size)	1	10	3.13	3.13	0
Jednorodność kształtu komórek (Uniformity of Cell Shape)	1	10	3.21	3.21	0
Adhezja (Marginal Adhesion)	1	10	2.81	2.81	0
Rozmiar 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					

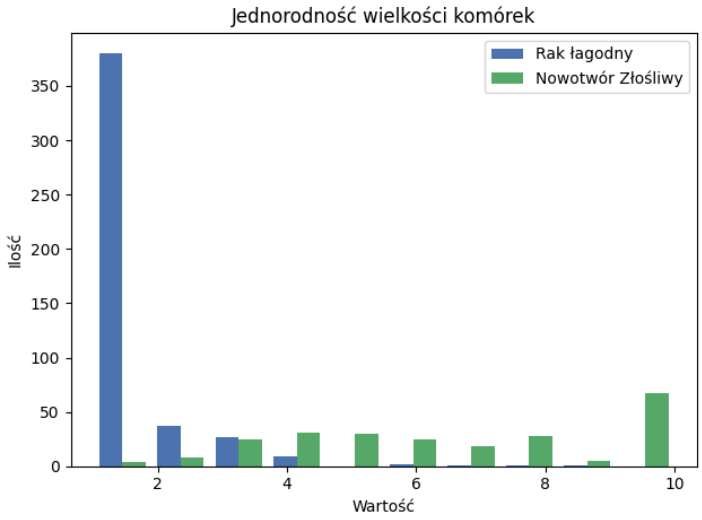
b. Częstość występowania poszczególnych klasyfikacji (diagnoz)

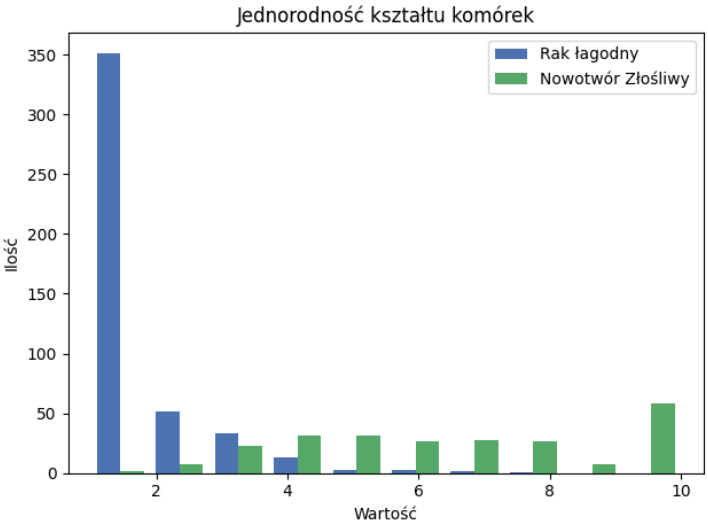


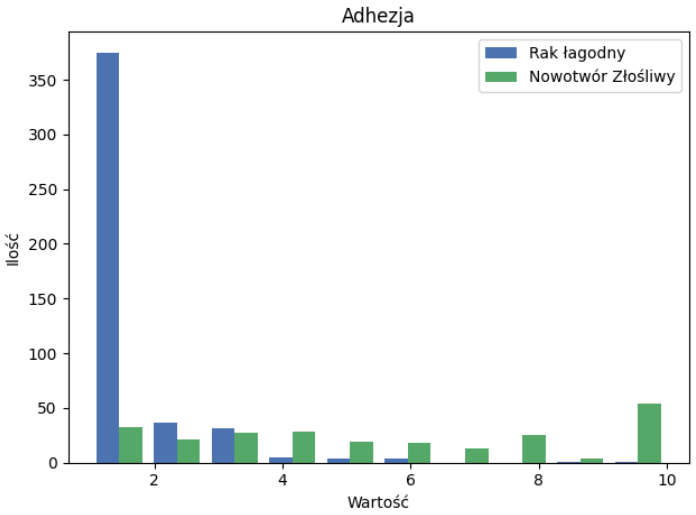
- Rak łagodny: 458 (65.5%)
- Nowotwór złośliwy: 241 (34.5%)

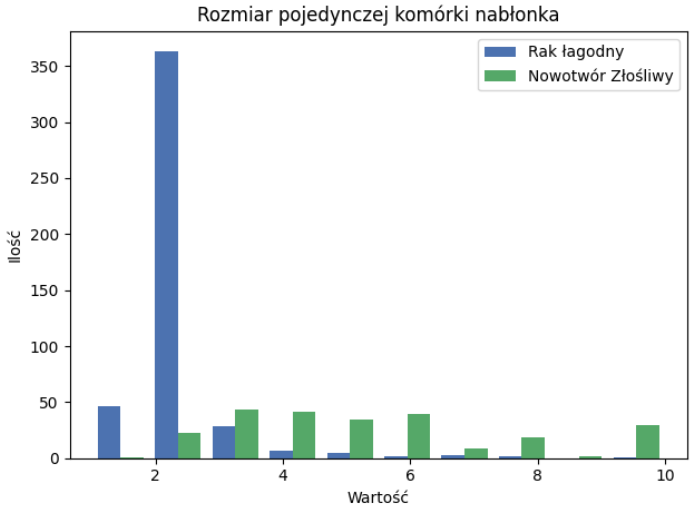
c. Częstość występowania poszczególnych odpowiedzi w kolumnach:

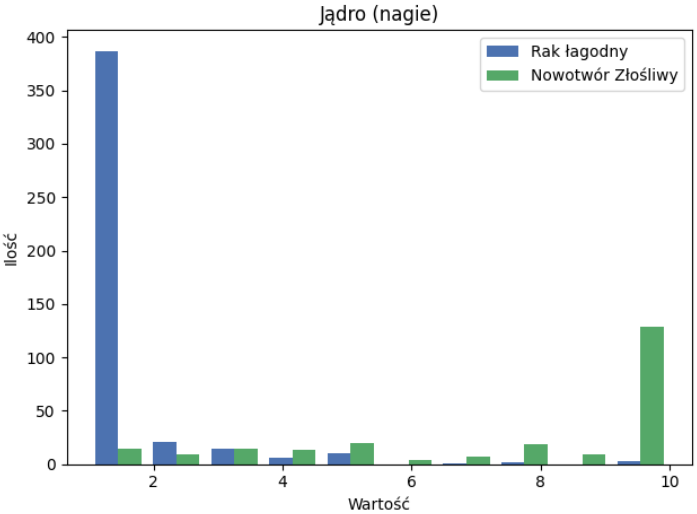


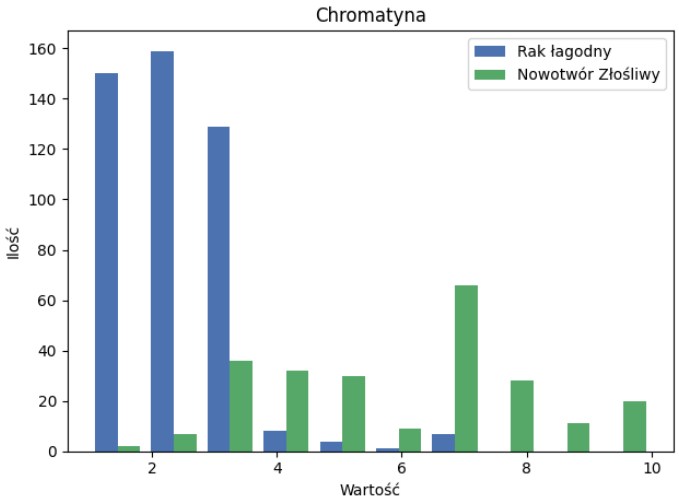


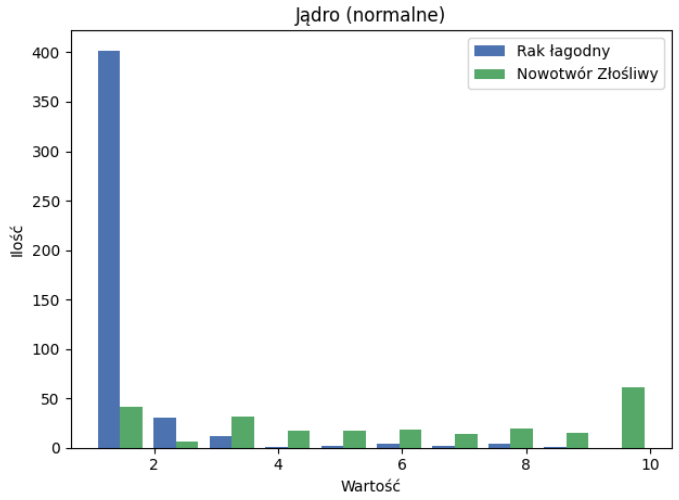


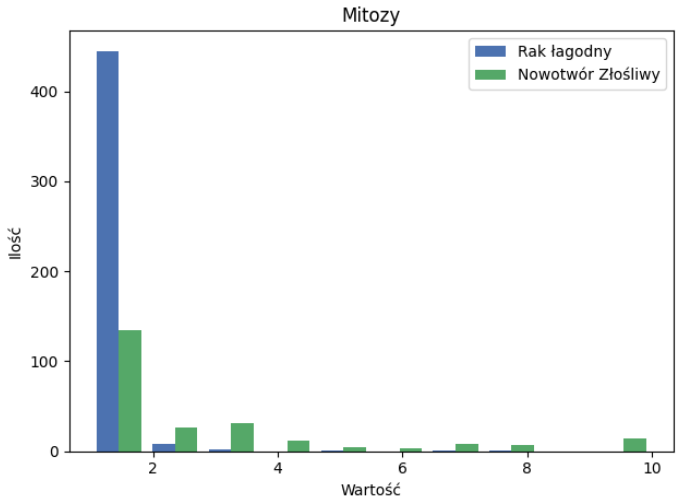

Adhezja
Rak łagodny
Nowotwór Złośliwy
Ilość
Wartość
0
50
100
150
200
250
300
350
2
4
6
8
10


Rozmiar pojedynczej komórki nabłonka
Rak łagodny
Nowotwór Złośliwy
Ilość
Wartość
0
50
100
150
200
250
300
350
2
4
6
8
10


Jądro (nagie)
Rak łagodny
Nowotwór Złośliwy
Ilość
Wartość
0
50
100
150
200
250
300
350
400
2
4
6
8
10

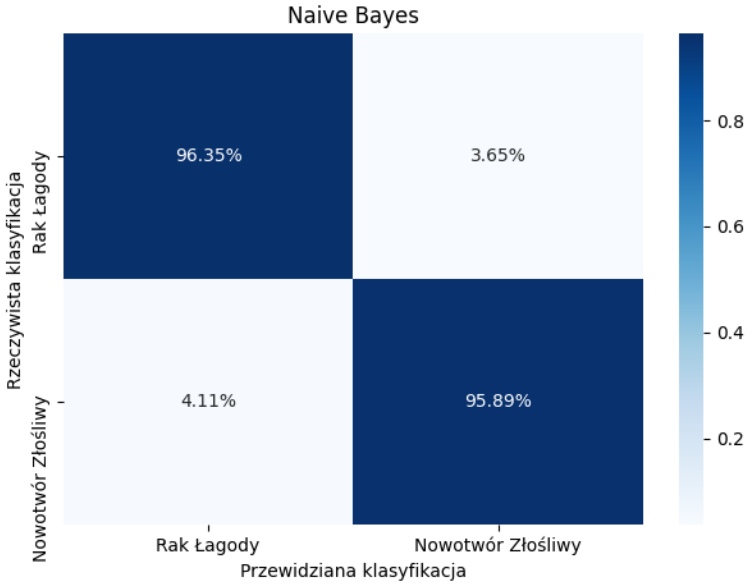

Chromatyna
Rak łagodny
Nowotwór Złośliwy
Ilość
Wartość
0
20
40
60
80
100
120
140
160
2
4
6
8
10


Jądro (normalne)
Rak łagodny
Nowotwór Złośliwy
Ilość
Wartość
0
50
100
150
200
250
300
350
400
2
4
6
8
10

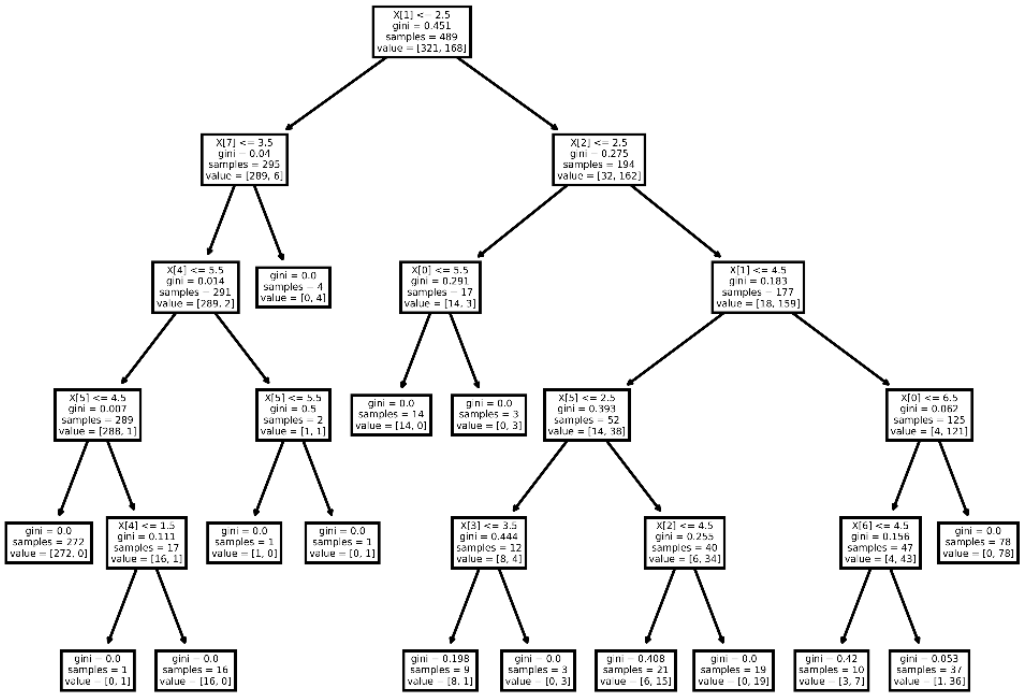

Mitozy
Rak łagodny
Nowotwór Złośliwy
Ilość
Wartość
0
100
200
300
400
2
4
6
8
10

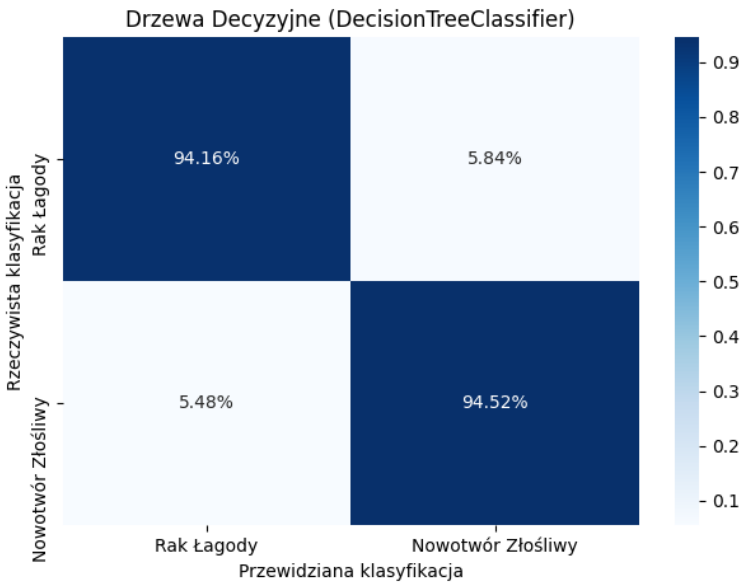
2. Skuteczność klasyfikatorów

a. Naive Bayes – 96,19 %



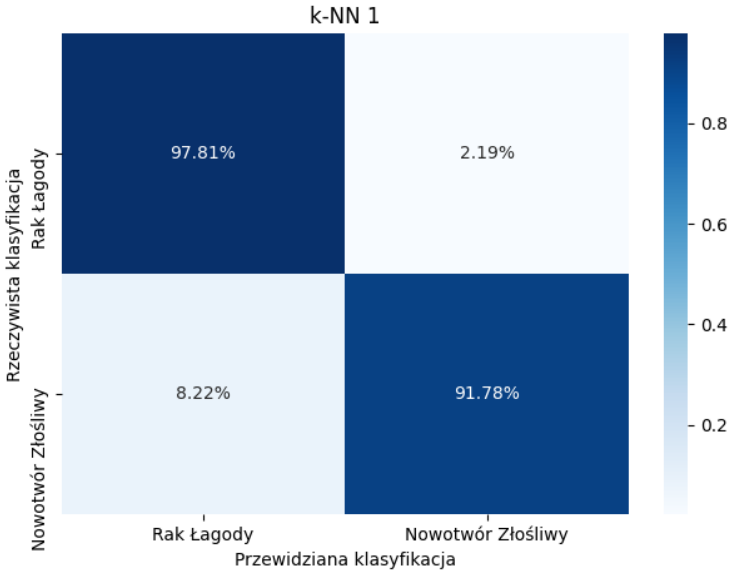
b. Drzewa decyzyjne – 95.71%



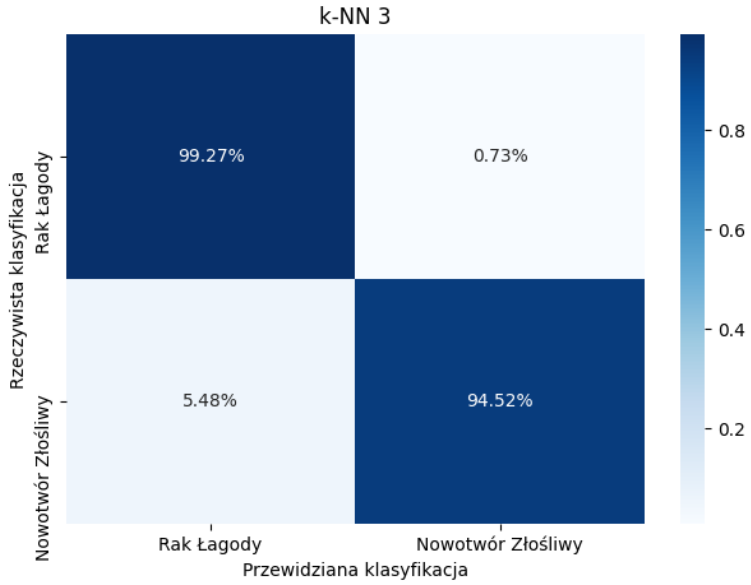


c. k-Najbliższych sąsiadów

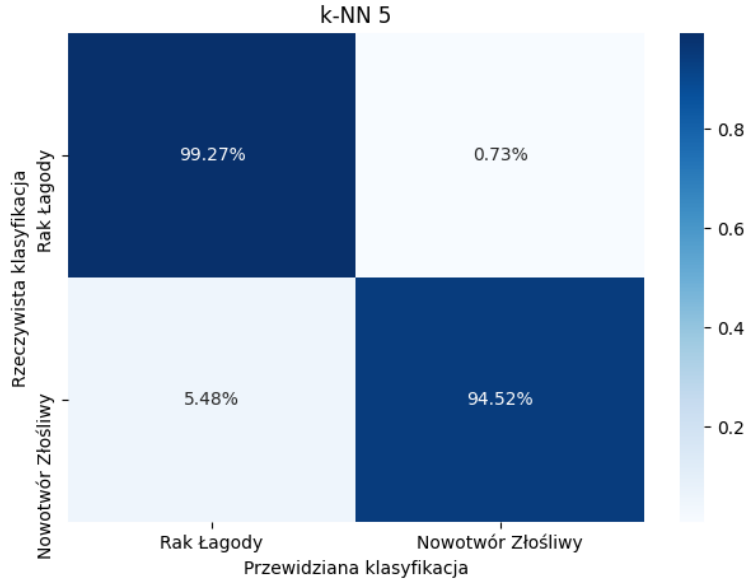
- k-NN-1 – **95.71%**



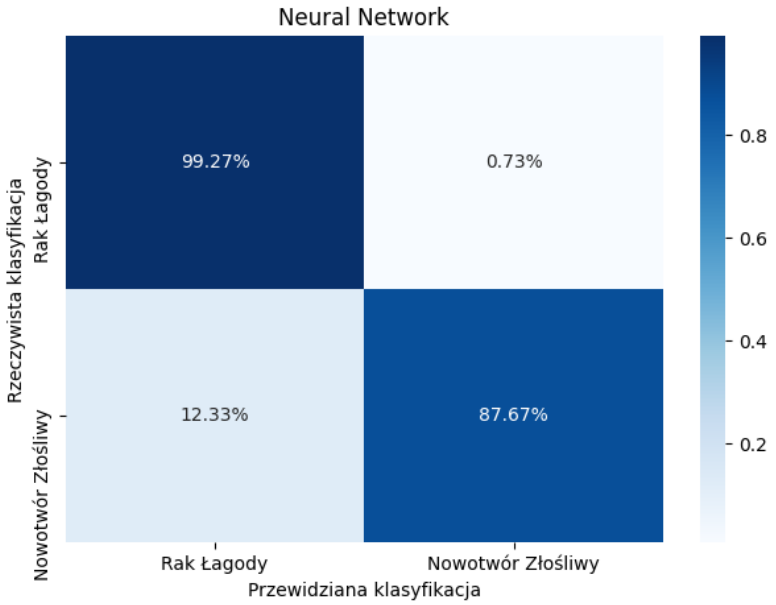
- k-NN-3 – **97.62%**



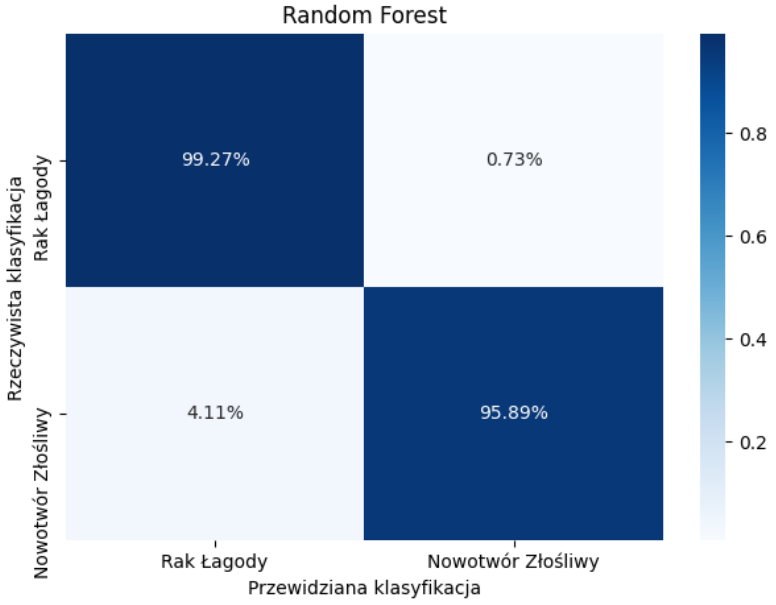
- k-NN-5 – **97.62%**



d. **Sieci neuronowe – 99,27%**



e. **Random Forest – 98.1%**
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.



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

g. **AdaBoost – 94.29%**
podstawowy algorytm do boostingu, metoda dzięki której z dużej liczby słabych klasyfikatorów można otrzymać jeden lepszy

3. Porównanie skuteczności klasyfikatorów

4. Wnioski

- brak możliwości określenia najlepiej działającego klasyfikatora na podanym zbiorze danych