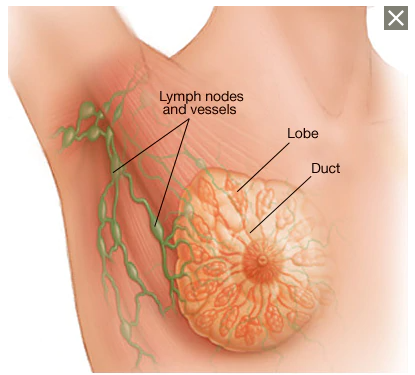
Sprawozdanie do projektu nr 2

Porównanie klasyfikatorów na przykładzie bazy

Breast Cancer Winconsin

(Rak Piersi Winconsin)



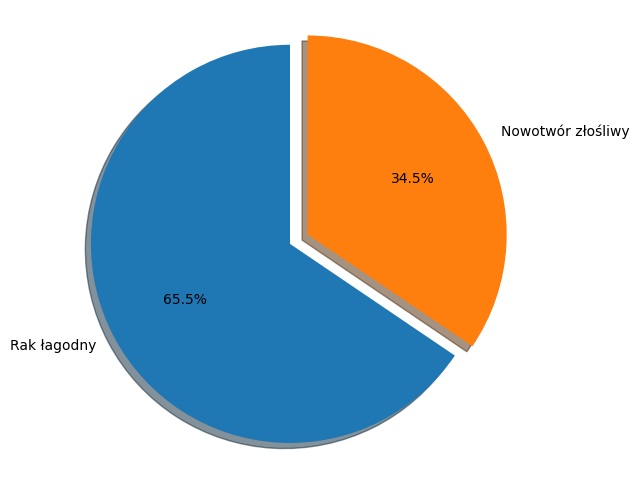
Baza danych: <https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)>

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

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Kolumna** | **Min** | **Max** | **Średnia** | **% brakujących danych** |
| Grubość zgrubenia (Clump Thickness) | 1 | 10 | 4.42 | 0 |
| Jednolitość rozmiaru komórki (Uniformity of Cell Size) | 1 | 10 | 3.13 | 0 |
| Jednorodność kształtu komórki (Uniformity of Cell Shape) | 1 | 10 | 3.21 | 0 |
| Marginalna przyczepność (Marginal Adhesion) | 1 | 10 | 2.81 | 0 |
| Rozmiar pojedynczej komórki nabłonka (Single Epithelial Cell Size) | 1 | 10 | 3.22 | 0 |
| Nagie jądra (Bare Nuclei) | 1 | 10 | 3.54 | 2.28 |
| Mdła Chromatyna (Bland Chromatin) | 1 | 10 | 3.44 | 0 |
| Normalne Jądra (Normal Nucleoli) | 1 | 10 | 2.87 | 0 |
| Mitozy (Mitoses) | 1 | 10 | 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:

wykres kołowy lub słupkowy

1. Porównianie skuteczności klasyfikatorów
   1. Naive Bayes

Ewaluacja powinna zawierać dokładność klasyfikatora i macierz błędu.

* 1. Drzewa decyzyjne visualize decision tree – latwo można zobrazować decyzje podejmowane przez algorytm do klasyfikacji
  2. k Najbliższych sąsiadów (1-k, 3-k, 5-k,)
  3. Sieci neuronowe (dla wybranej topologii lub kilku na wyższą ocenę)
  4. + 2-3 inne klasyfikatory (Support Vector Machines, Random Forest, Metody typu Ensemble lub inne). Dla innych zrób krótki wstęp teoretyczny, wyjasnij na jakiej zasadzie działają

1. Wnioski

Najlepiej działający klasyfikator

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

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

• 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?