Predykcja stadium choroby serca

Zuzanna Paluch

Spis treści

[1. Cel badania i opis zbioru danych 3](#_Toc188189231)

[2. Metodologia i rozwiązanie 3](#_Toc188189232)

[3. Wstępne przetwarzanie danych 4](#_Toc188189233)

[4. Metoda oceniania jakości modelu 4](#_Toc188189234)

[5. Wyniki eksperymentalne 5](#_Toc188189235)

[6. Podsumowanie 5](#_Toc188189236)

# Cel badania i opis zbioru danych

**Cel badania:**

Celem analizy jest opracowanie modelu klasyfikującego zdolnego przewidzieć, czy pacjent jest zdrowy, czy cierpi na chorobę serca. W przypadku pacjentów chorych model ma określić stopień zaawansowania chorob

**Opis zbioru danych:**

* **Źródło danych:** [UCI Machine Learning Repository - Heart Disease Dataset](https://archive.ics.uci.edu/ml/datasets/Heart+Disease)
* **Liczba rekordów:** 303 pacjentów
* **Liczba atrybutów:** 13 wybranych cech spośród 75 dostępnych (opisujących dane demograficzne, wyniki badań klinicznych oraz inne istotne informacje).
* **Rozkład klas decyzyjnych:**
  + Klasa 0: Pacjenci zdrowi
  + Klasa 1–4: Pacjenci chorzy z różnym stopniem zaawansowania choroby

# 2.Metodologia i rozwiązanie

**Selekcja cech:** Przeprowadzono analizę w celu wyboru najistotniejszych atrybutów wpływających na wynik klasyfikacji.

**Normalizacja danych:** W przypadku wykrycia znacznych różnic w skali wartości cech zastosowano normalizację, aby poprawić działanie modeli uczących się.

**Budowa modelu:** Rozważono kilka algorytmów klasyfikacji:

* + Sieci neuronowe
  + Klasyfikator Bayesowski
  + Support Vector Machine (SVM)
  + Drzewa decyzyjne
  + Reguły decyzyjne
  + Bagging
  + Boosting

**Podział danych:** Zbiór danych podzielono na:

* Zbiór treningowy (75%)
* Zbiór testowy (25%)

# 3.Wstępne przetwarzanie danych

**Selekcja cech:** Wybrano 5 kluczowych cech o największej istotności ['cp', 'thalach', 'oldpeak', 'ca', 'thal'].

**Usuwanie brakujących danych:** Usunięto rekordy z brakującymi wartościami.

**Normalizacja:** Wszystkie cechy ciągłe zostały przeskalowane do zakresu [0, 1] przy użyciu metody Min-Max scaling.

**Kodowanie zmiennych kategorycznych:** Zastosowano technikę one-hot encoding, aby umożliwić efektywne wykorzystanie tych zmiennych przez algorytmy klasyfikacyjne.

# 4.Metoda oceniania jakości modelu

**Metryki oceny:**

* Dokładność (Accuracy)
* Czułość (Recall)
* Precyzja (Precision)
* F1-score

# Obraz zawierający tekst, zrzut ekranu, diagram, Wykres Opis wygenerowany automatycznie5.Wyniki eksperymentalne

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Recall | Precision | F1-score |
| Sieć neuronowa | 0.67 | 0.93 | 0.87 | 0.90 |
| Klasyfikator Bayesowski | 0.48 | 0.81 | 0.89 | 0.85 |
| SVM | 0.61 | 0.79 | 0.94 | 0.86 |
| Drzewa decyzyjne | 0.51 | 0.74 | 0.84 | 0.78 |
| Reguły decyzyjne | 0.51 | 0.76 | 0.76 | 0.76 |
| Bagging | 0.49 | 0.79 | 0.72 | 0.75 |
| Boosting | 0.53 | 0.83 | 0.83 | 0.83 |

# 6.Podsumowanie

Na podstawie wyników przedstawionych w tabeli można zauważyć, że najlepsze wyniki pod względem ogólnej jakości klasyfikacji (mierzonych metryką F1-score) osiągnął model sieci neuronowej (F1-score = 0.90). Model ten charakteryzował się również wysoką precyzją (0.87) i znakomitą czułością (Recall = 0.93), co czyni go najbardziej efektywnym spośród testowanych algorytmów.

Kolejnym modelem, który uzyskał dobre wyniki, był klasyfikator Bayesowski z F1-score równym 0.85. Mimo niższej dokładności (Accuracy = 0.48) model ten dobrze zbalansował precyzję (0.89) i czułość (0.81), co sugeruje, że radził sobie lepiej z niektórymi klasami.

SVM osiągnął umiarkowaną jakość klasyfikacji z F1-score na poziomie 0.86, dzięki wysokiej precyzji (0.94), ale miał niższy Recall (0.79). Drzewo decyzyjne i reguły decyzyjne uzyskały podobne wyniki z F1-score około 0.76–0.78, co wskazuje na ich porównywalną skuteczność, choć nadal niższą w porównaniu do bardziej zaawansowanych metod.

Bagging i Boosting uzyskały najniższe wyniki (F1-score odpowiednio 0.75 i 0.83). Boosting wykazał się nieco lepszą ogólną skutecznością niż Bagging, jednak oba te podejścia nie przewyższyły wyników sieci neuronowych czy SVM.

**Wnioski:**

1. Sieć neuronowa jest najbardziej obiecującym modelem w kontekście analizowanego problemu, szczególnie ze względu na wysoką czułość (ważne w problemach zdrowotnych).
2. Modele oparte na prostszych zasadach (drzewa decyzyjne i reguły decyzyjne) mogą być stosowane w sytuacjach, gdy interpretowalność ma kluczowe znaczenie, mimo niższej skuteczności.
3. Boosting oferuje lepsze wyniki niż Bagging, ale nadal pozostaje mniej skuteczny niż sieci neuronowe i SVM.

Przyszłe prace mogą skupić się na dalszej optymalizacji parametrów modeli, eksploracji dodatkowych cech oraz wykorzystaniu bardziej złożonych algorytmów sieci neuronowych w celu poprawy wyników klasyfikacji.