Prognoza raka piersi z użyciem klasyfikatora Bayesowskiego

Fabian Zbrański

PRO1D 2023/2024

Cel badania i opis zbioru danych

Celem przeprowadzonego badania jest opracowanie modelu klasyfikacyjnego do prognozowania wystąpienia raka piersi na podstawie danych pochodzących z zestawu “Breast Cancer Wisconsin (Prognosis)”. Poniżej przedstawiam kluczowe informaczje dotyczące zbioru danych:

- Liczba Atrybutów: 35 (Łącznie z ID)
- Liczba Rekordów: 198
- Rozkład Klas Decyzyjnych:
    - Klasa 0 (nonrecur): 151
    - Klasa 1 (recur): 47

W zestawie danych znajduję się 35 kolumn, z których 33 to cechy, a pozostałe dwie to identyfikator (ID) oraz wynik klasyfikacji (Outcome). Aby zrealizować cel badania, dokonałem konwersji kolumny Outcome na postać binarną gdzie R (recur) zostało zamienione na wartość 1, a N (nonrecur) warość 0

Metodologia i Rozwiązanie

Eksperyment został przeprowadzony przy użyciu klasyfikatora Bayesowskiego w celu prognozowania obecności raka piersi na podstawie danych ze zbioru “Breast Cancer Wisconsin (Prognosis)”. Poniżej opisuje kroki podjęte w procesie analizy:

Podział Danych na Zbiór Treningowy i Testowy

Dane zostały podzielone na zbiór treningowy (75%) i testowy (25%) w celu oceny skuteczności modelu.

Inicjalizacja Klasyfikatora Bayesowskiego

Wykorzystałem klasyfikator GaussianNB z modułu scikit-learn.

Trenowanie Klasyfikatora

Klasyfikator został wytrenowany na zbiorze treningowym, korzystając z wcześniej znormalizowanych danych.

Ewaluacja Modelu

- Dokonałem predykcji na zbiorze testowym.
- Obliczyłem dokładność (accuracy) klasyfikacji.

Generowanie Raportu Klasyfikacji

Wykorzystałem funkcję classification_report z modułu scikit-learn do uzyskania szczegółowego raportu, zawierającego precision, recall, f1-score i support dla obu klas

Wstępne przetwarzanie danych

W etapie wstępnego przetwarzania danych skoncentrowałem się na przygotowaniu zbioru danych do analizy oraz trenowania modelu klasyfikacyjnego. Przeprowadzone kroki miały na celu usunięcie potencjalnych zakłóceń i dostosowanie danych do wymagań klasyfikatora. W efekcie uzyskaliśmy przygotowany zbiór danych gotowy do trenowania modelu klasyfikacyjnego w kontekście prognozowania raka piersi. Oto główne kroki podjęte w procesie.

Konwersja Kolumny "Outcome"

Kolumna "Outcome", reprezentująca wynik klasyfikacji (recur lub nonrecur), została przekształcona na postać binarną:

- R (recur) = 1
- N (nonrecur) = 0

Zastąpienie Brakujących Wartości

W kolumnie "Lymph_Node_Status" stwierdzono obecność brakujących wartości oznaczonych jako "?". Brakujące wartości zastąpiłem medianą tej kolumny, aby nie wpływały one negatywnie na analizę.

Normalizacja Danych

W celu uzyskania jednolitej skali, znormalizowałem dane przy użyciu StandardScaler z modułu scikit-learn. Normalizacja pomaga w poprawie wydajności klasyfikacyjnego, eliminując wpływ różnic w skali wartości atrybutów.

Metody Oceny Jakości Modelu

Do oceny jakości modelu klasyfikacyjnego wykorzystałem kilka kluczowych metryk, które umożliwiają holistyczną analizę jego sutecznośći. Metodyki oceny zostały przeze mnie dobrane w taki sposób, aby dostarczyć pełny obraz wydajności modelu w kontekście zadania klasyfikacji raka piersi. Oto główne elementy metodyki oceny.

Dokładność (Accuracy)

Obliczyłem dokładność klasyfikatora, która określa stosunek liczby poprawnie sklasyfikowanych przypadków dla ogólnej liczby przypadków w zbiorze testowym. Dokładność jest szczególnie istotna w przypadku równomiernego rozkładu klas.

Raport Klasyfikacji (classification_report)

Skorzystałem z funkcji 'classification_report' z modułu 'sklearn.metrics', aby uzyskać szczegółowy raport klasyfikacyjny. Raport zawiera precision, recall, f1-score i support dla obu klas, co pozwala na lepsze zrozumienie wydajności klasyfikatora w kontekście każdej klasy.

Macierz Pomyłek (Confusion Matrix)

Analizowałem macierz pomyłek, która przedstawia liczbę przypadków prawdziwie pozytywnych, prawdziwie negatywnych, fałszywie pozytywnych i fałszywie negatywnych. Macierz pomyłek dostarcza informacji o rodzajach błędów popełnianych przez klasyfikator.

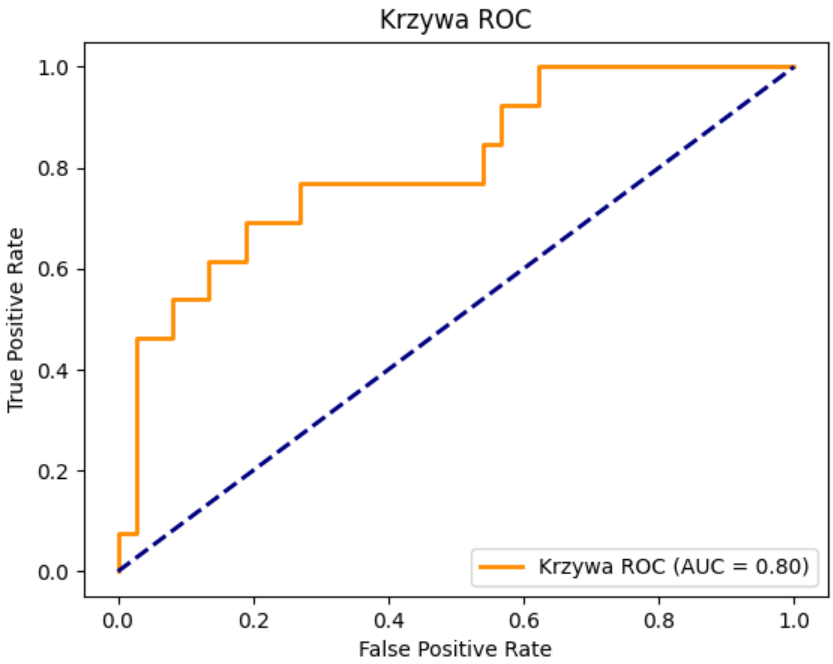
Krzywa ROC (Receiver Operating Characteristic)

W celu oceny zdolności klasyfikatora do rozróżniania między klasami, zwłaszcza w przypadku niezrównoważonych zbiorów danych, analizowałem krzywą ROC. Powierzchnia pod krzywą ROC (AUC-ROC) stanowi miarę skuteczności klasyfikatora.

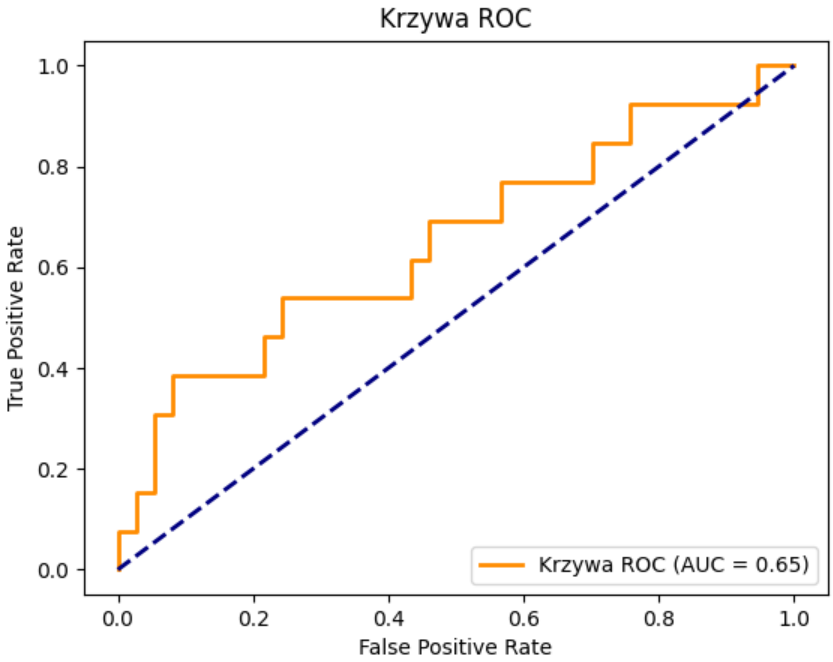
Wyniki Eksperymentalne + Wykresy w Razie Potrzeby

Krzywa ROC

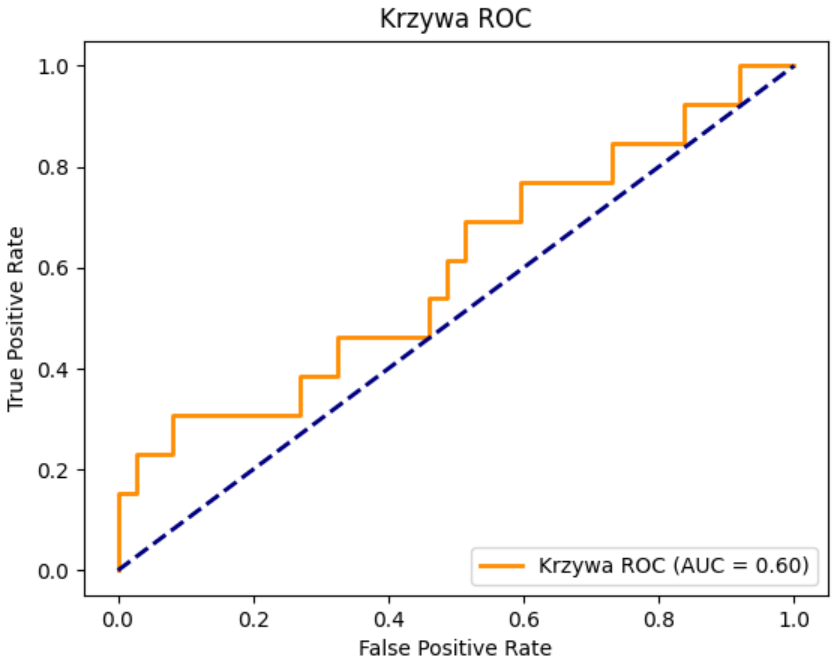
- Krzywa ROC dla random_state równego 75



- Krzywa ROC dla random_state równego 50

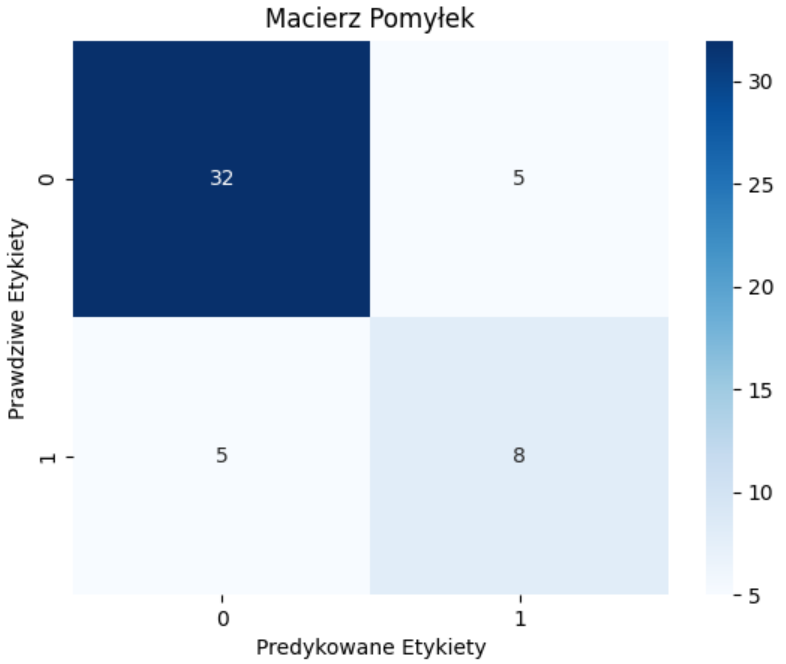


- Krzywa ROC dla random_state równego 85

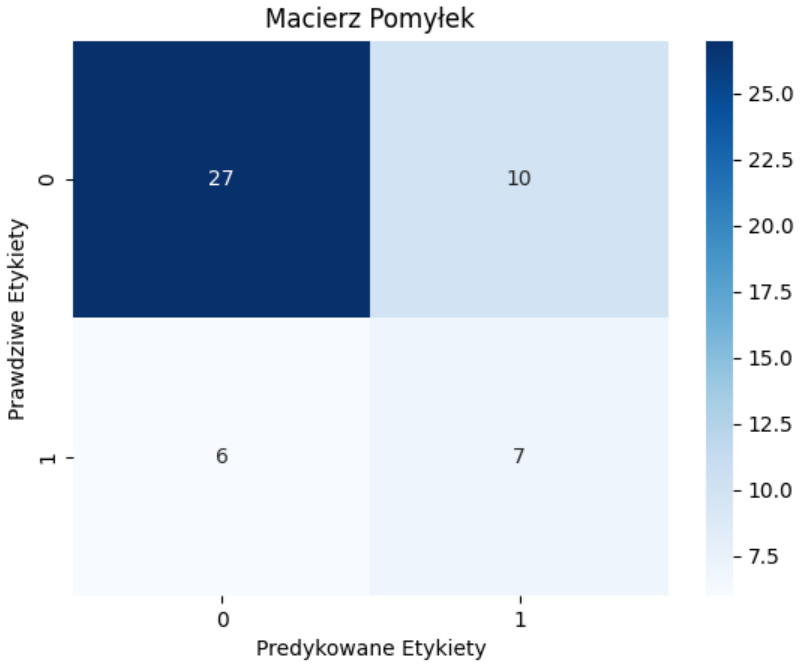


Macierz Pomyłek

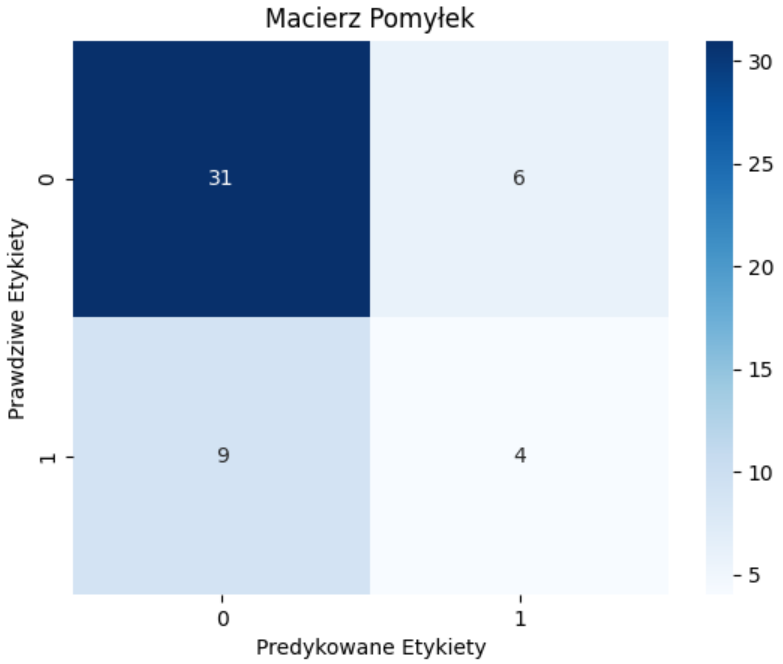
- Macierz Pomyłek dla random_state równego 75



- Macierz Pomyłek dla random_state rówego 50



- Macierz Pomyłek dla random_state równego 85



Raport klasyfikacji

- Raport klasyfikacji dla random_state równego 75

```
Accuracy: 0.8
              precision    recall  f1-score   support

           0       0.86      0.86      0.86        37
           1       0.62      0.62      0.62        13

    accuracy                           0.80        50
   macro avg       0.74      0.74      0.74        50
weighted avg       0.80      0.80      0.80        50
```

- Raport klasyfikacji dla random_state równego 50

```
Accuracy: 0.68
              precision    recall  f1-score   support

           0       0.82      0.73      0.77        37
           1       0.41      0.54      0.47        13

    accuracy                           0.68        50
   macro avg       0.61      0.63      0.62        50
weighted avg       0.71      0.68      0.69        50
```

- Raport klasyfikacji dla random_state równego 85

```
Accuracy: 0.7
              precision    recall  f1-score   support

           0       0.78      0.84      0.81        37
           1       0.40      0.31      0.35        13

    accuracy                           0.70        50
   macro avg       0.59      0.57      0.58        50
weighted avg       0.68      0.70      0.69        50
```

Podsumowanie: Własne Komentarze, Wnioski

Eksperyment ten pozwolił nam na głębsze zrozumienie skuteczności klasyfikatora Bayesowskiego w prognozwowaniu przypadków raka piersi na podstawie dostępnych cech diagnostycznych. Poniżej przedstawiam własne komentarze oraz główne wnioski.

Skuteczność

Klasyfikator Bayesowski, mimo swojej prostoty, wykazał się umiarkowaną skutecznością w rozróżnianiu między przypadkami związanymi z rakiem a przypadkami niezwiązanymi z rakiem. Dokładność wynosząca 80% została osiągnięta przy parametrze random_state ustawionym na 75. Jakiekolwiek odejścia od tej wartości w góre bądź w dół powodowały pogorszenie wyników klasyfikatora.

Wpływ Brakujących Wartości

Problem brakujących wartości w kolumnie “Lymph_Node_Status” miał wpływ na proces trenowania klasyfikatora. Zastosowana metoda zastępowania brakujących wartości medianą okazała się praktycznym rozwiązaniem.

Różnice w Wynikach

Analiza raportu klasyfikacyjnego wykazała, że klasyfikator miał trudności w identyfikowaniu przypadków rzeczywistych. Szczególnie zauważalne są niższe wartości precision, recall i f1-score dla klasy 1 (recur). Jest to prawdopodobnie spowodowane dysproporcją w ilości rekordów recur oraz norecur w datasecie.

Potencjalne Obszary Doskonalenia

Optymalizacja hiperparametrów klasyfikatora Bayesowskiego oraz przegląd dodatkowych cech diagnostycznych mogą poprawić ogólną skuteczność modelu.

Podsumowanie Ogólne

Mimo pewnych ograniczeń, eksperyment ten stanowi cenny krok w kierunku opracowania modelu klasyfikacyjnego dla prognozowania raka piersi. Wnioski te stanowią punkt wyjścia do dalszych badań i doskonalenia modelu w przyszłości.