Temat: Rozpoznawanie choroby serca

Skład grupy: Maksymilian Dębek s25001

Spis treści

1. Cel badania i opis zbioru danych 2
2. Metodologia i rozwiązanie 3
3. Wstępne przetwarzanie danych 4
4. Metoda oceniania jakości modelu 5
5. Wyniki eksperymentalne i wykresy 6
   5.1 Niepogrupowane dane 6
   5.2 Dane pogrupowane niebinarnie 9
   5.3 Dane pogrupowanie binarne 12
6. Podsumowanie 14

1. Cel badania i opis zbioru danych

Celem badania jest Rozpoznawanie choroby serca na podstawie podanych danych oraz określenie stopnia zaawansowania choroby jeżeli pacjent jest choroby. W projekcie zostały użyte do tego 2 klasyfikatory: **Drzewo Decyzyjne oraz Naiwny klasyfikator Bayesa**.

Źródło danych:

https://archive.ics.uci.edu/dataset/45/heart+disease

(Cleveland data)

Liczba atrybutów w zbiorze danych: 13

Opis atrybutów:

- age: wiek pacjenta
- sex: płeć pacjenta
- cp: rodzaj bólu w klatce piersiowej (4 wartości)
- trestbps: ciśnienie krwi w spoczynku
- chol: poziom cholesterolu we krwi w mg/dl
- fbs: poziom cukru we krwi na czczo > 120 mg/dl
- restecg: wyniki spoczynkowego elektrokardiogramu (wartości 0, 1, 2)

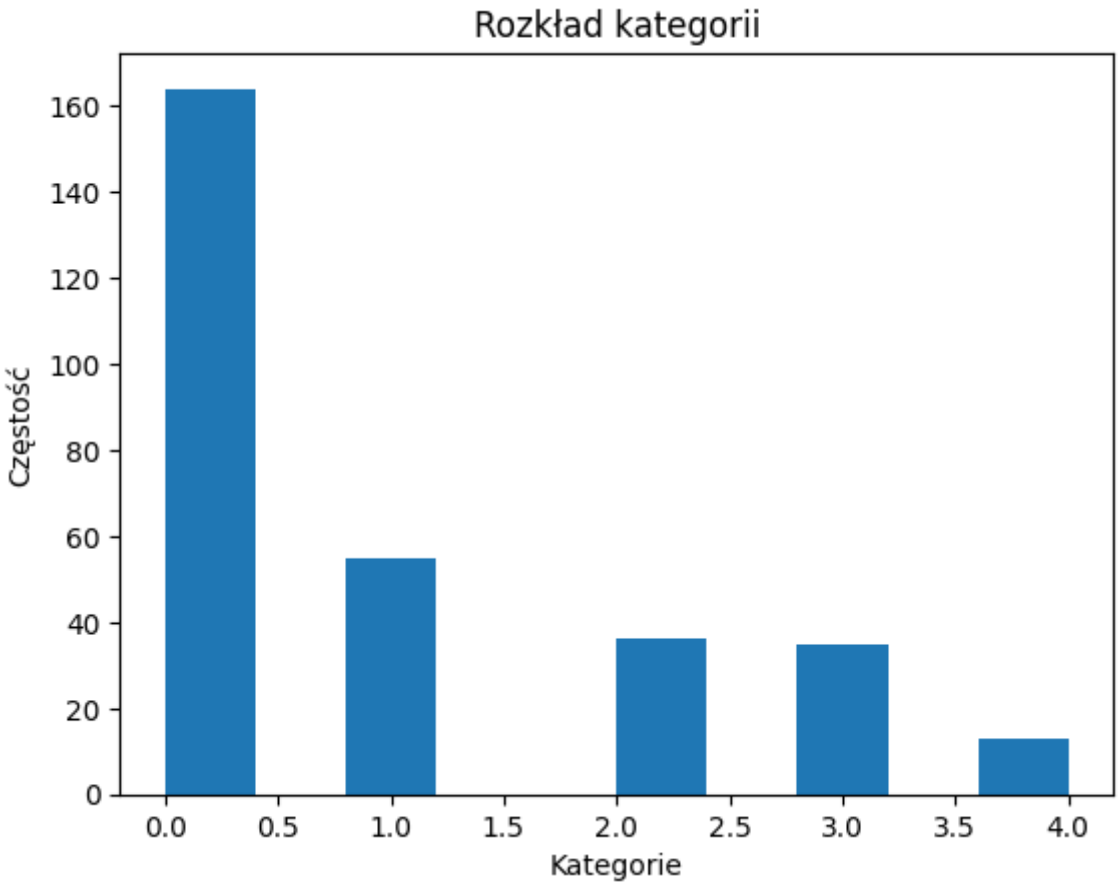
- thalach: maksymalna osiągnięta akcja serca
- exang: dławienie wysiłkowe (angina) wywołane wysiłkiem
- oldpeak: obniżenie odcinka ST wywołane wysiłkiem w stosunku do spoczynku
- slope: nachylenie szczytowego odcinka ST w czasie ćwiczeń
- ca: liczba dużych naczyń (0–3) barwionych fluoroskopią
- thal: 0 = normalne; 1 = ustalone uszkodzenie; 2 = odwracalne uszkodzenie
- num: obecność choroby serca u pacjenta(0 = zdrowy, 1,2,3,4 = chory z określonym stopniem zaawansowania)

Liczba rekordów: 303

Rozkład klas decyzyjnych:

- 0 – zdrowy
- 1,2,3,4 – chory

Rozkład ilościowy klas decyzyjnych przed oversamplingiem:



2. Metodologia i rozwiązanie

Klasyfikatory oraz metody używane w tym modelu:

- **Decision Tree** – Metoda, która buduje model decyzyjny w formie drzewa. Węzły drzewa reprezentują testy na cechy danych, a liście zawierają etykiety klas. Służy do rozwiązywania problemów klasyfikacji i regresji oraz jest to rodzaj modelu predykcyjnego, który podejmuje decyzje w oparciu o serię warunków logicznych.

- **Naiwny Klasyfikator Bayesa** – Jest to rodzaj algorytmu klasyfikacji wykorzystywanego w nauczaniu maszynowym. Jest oparty na twierdzeniu Bayesa i zakłada naiwnie niezależność między cechami obiektu.
- **SMOTE** - Technika oversamplingu, która generuje sztuczne przykłady dla klasy mniejszościowej aby zrównoważyć zbiór danych.
- **Boosting** – Jest to technika w nauczaniu maszynowym, w której tworzone jest wiele modeli na podstawie różnych próbek (podzbiorów) zbioru treningowego, a następnie wyniki tych modeli są łączone, aby uzyskać stabilniejszy i bardziej skuteczny model.

3. Wstępne przetwarzanie danych

W podanym zbiorze danych znajduje się 6 wartości brakujących. 4 z nich są w kolumnie ‘ca’ oraz 2 w kolumnie ‘thal’. W każdym z przypadków wartości brakujące zostały uzupełnione średnią z danej kolumny. Przy dzieleniu danych zbiory testowe i treningowe, dokonaliśmy również selekcji atrybutów decyzyjnych (ostatnia kolumna). Dodatkowo atrybuty decyzyjne są niezrównoważone (164 przypadków 0, a zaledwie 13 przypadków 4) dlatego przed trenowaniem modelu stosowany jest oversampling przy użyciu SMOTE, a dodatkowo normalizacja danych. **W trakcie pracy nad projektem przez bardzo niską skuteczność zbiór danych był pogrupowany na 3 kategorie:**

- **Niezmieniony zbiór danych**
- **Atrybuty decyzyjne podzielone na 4 grupy (0 -> 0; 1 -> 1; 2 -> 1; 3 -> 3; 4 - > 4), podział ten został ustalony na podstawie analizy macierzy pomyłek**
- **Atrybuty decyzyjne podzielone na 2 grupy (0 -> 0; 1,2,3,4 -> 1)**

Dla każdej grupy danych trenowany był nowy model, a porównanie tych modeli znajdzie się w tym raporcie.

4. Metoda oceniania jakości modelu

Metody oceniania jakości modelu które zostały wykorzystane to:

- Accuracy – $\frac{\text{Liczba poprwanie sklasyfikowanych przypadków}}{\text{Wszystkie przypadki}}$
- F1 miara - $\frac{2*\text{precission}*\text{recall}}{\text{precission}+\text{recall}}$

  Gdzie:

  - precission - $\frac{TP}{TP+FP}$
  - recall - $\frac{TP}{TP+FN}$
  - TP to liczba prawdziwie pozytywnych przypadków,
  - FP to liczba fałszywie pozytywnych przypadków,
  - FN to liczba fałszywie negatywnych przypadków

- Recall – $\frac{TP}{TP+FN}$
- Precission - $\frac{TP}{TP+FP}$
- Macierz pomyłek - tablica, która przedstawia ilość poprawnych i błędnych klasyfikacji dokonanych przez model

5. Wyniki eksperymentalne i wykresy

5.1 Niepogrupowane dane

Najpierw zajmiemy się ocenianiem modelu dla niepogrupowanych danych.

Wyniki dla drzewa decyzyjnego.

Ustawione hiper parametry (max_depth=20, criterion='entropy')

Parametry te zostały znalezione poprzez algorytm Grid Search.

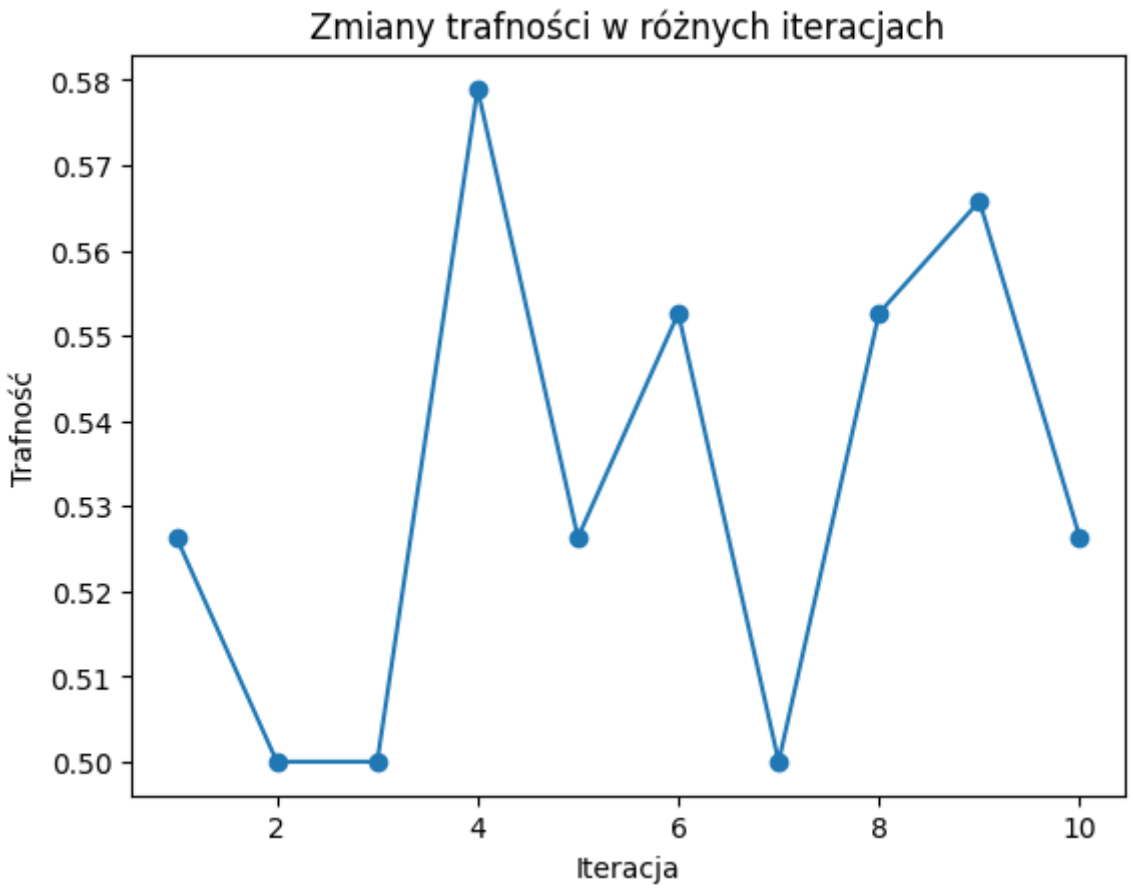
```
Accuracy: 0.6829268292682927
[[23  7  0  4  0]
 [ 5 34  7  3  1]
 [ 3  6 22  5  6]
 [ 2  0  7 32  1]
 [ 0  2  3  3 29]]
Precision: 0.6813604859568237
Recall: 0.6829268292682927
F1 measure: 0.6814818533236583
```

Co ciekawe korzystając z metody Bagging, dla pojedynczego rezultatu udało się osiągnąć aż 80% skuteczności, recall oraz precyzji i 79% F-miary.

```
Bagging Accuracy: 0.8
Bagging Confusion Matrix:
 [[27  5  0  2  0]
 [ 3 37  7  3  0]
 [ 0  2 31  4  5]
 [ 0  4  5 32  1]
 [ 0  0  0  0 37]]
Bagging Precision: 0.8001874806893277
Bagging Recall: 0.8
Bagging F1 measure: 0.7984801869616491
```

Jednak bardziej interesują nas wyniki wykonywane przez 10 epok, wyciągając średnie z tych iterowań jesteśmy w stanie więcej odczytać o naszym modelu.

```
Średnia trafność: 0.5328947368421052
Odchylenie standardowe trafności: 0.027125694905379357
Średnie precision:, 0.5479958732358695
Średni recall:, 0.5328947368421052
Średnia F1 miara: 0.2823244727248341
```

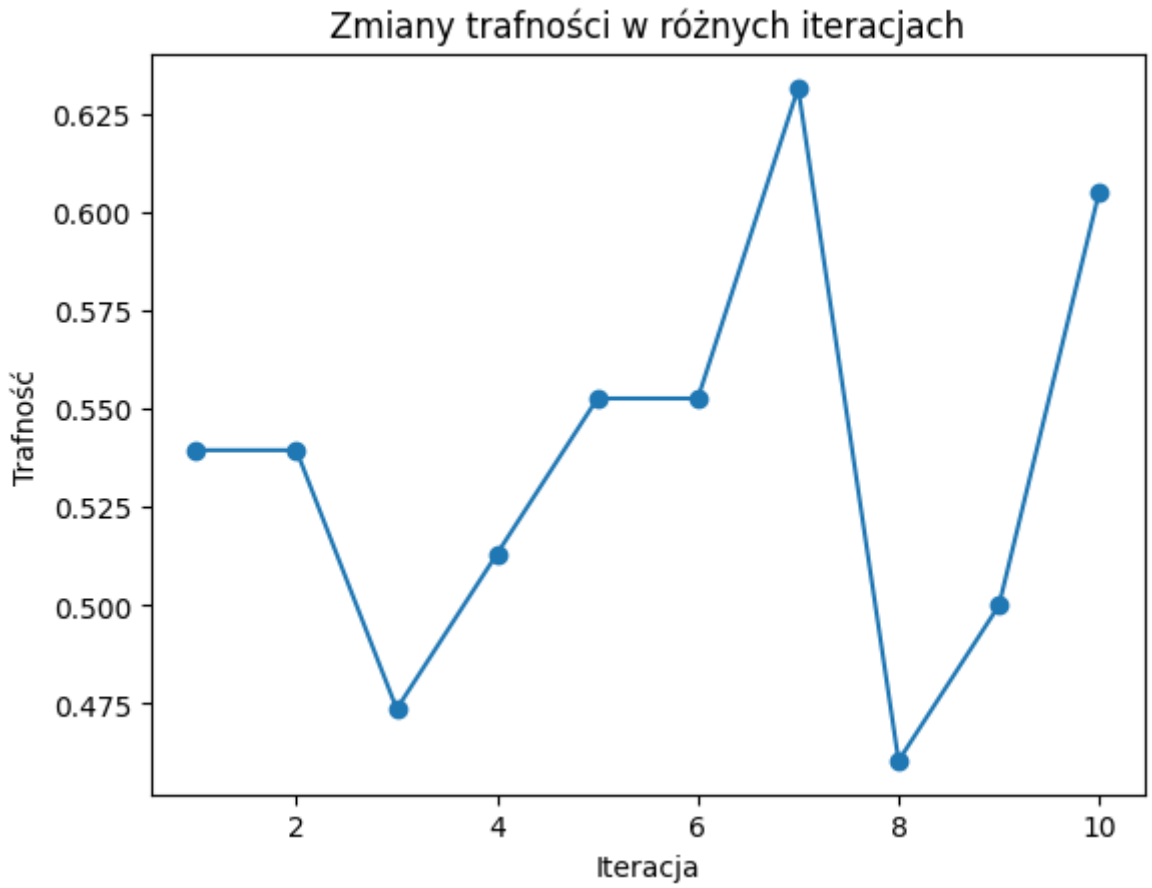


Wyniki dla Klasyfikatora Bayesa:

```
Accuracy: 0.5526315789473685
Precision: 0.6698716488190173
Recall: 0.5526315789473685
F1 measure: 0.5871746674378253
[[35  5  0  0  4]
 [ 1  1  2  0  5]
 [ 0  3  2  1  3]
 [ 0  1  1  1  6]
 [ 0  1  0  1  3]]
              precision    recall  f1-score   support

           0       0.97      0.80      0.88        44
           1       0.09      0.11      0.10         9
           2       0.40      0.22      0.29         9
           3       0.33      0.11      0.17         9
           4       0.14      0.60      0.23         5

    accuracy                           0.55        76
   macro avg       0.39      0.37      0.33        76
weighted avg       0.67      0.55      0.59        76
```

Wyniki dla 10 iteracji:

```
Średnia trafność: 0.536842105263158
Odchylenie standardowe trafności: 0.05089231475214136
Średnie precision:, 0.5881403646947594
Średni recall:, 0.536842105263158
Średnia F1 miara: 0.32687302798267137
```



W tym przypadku średnia trafność jest podobna do naszego jednorazowego testowania.

5.2 Dane pogrupowane niebinarnie

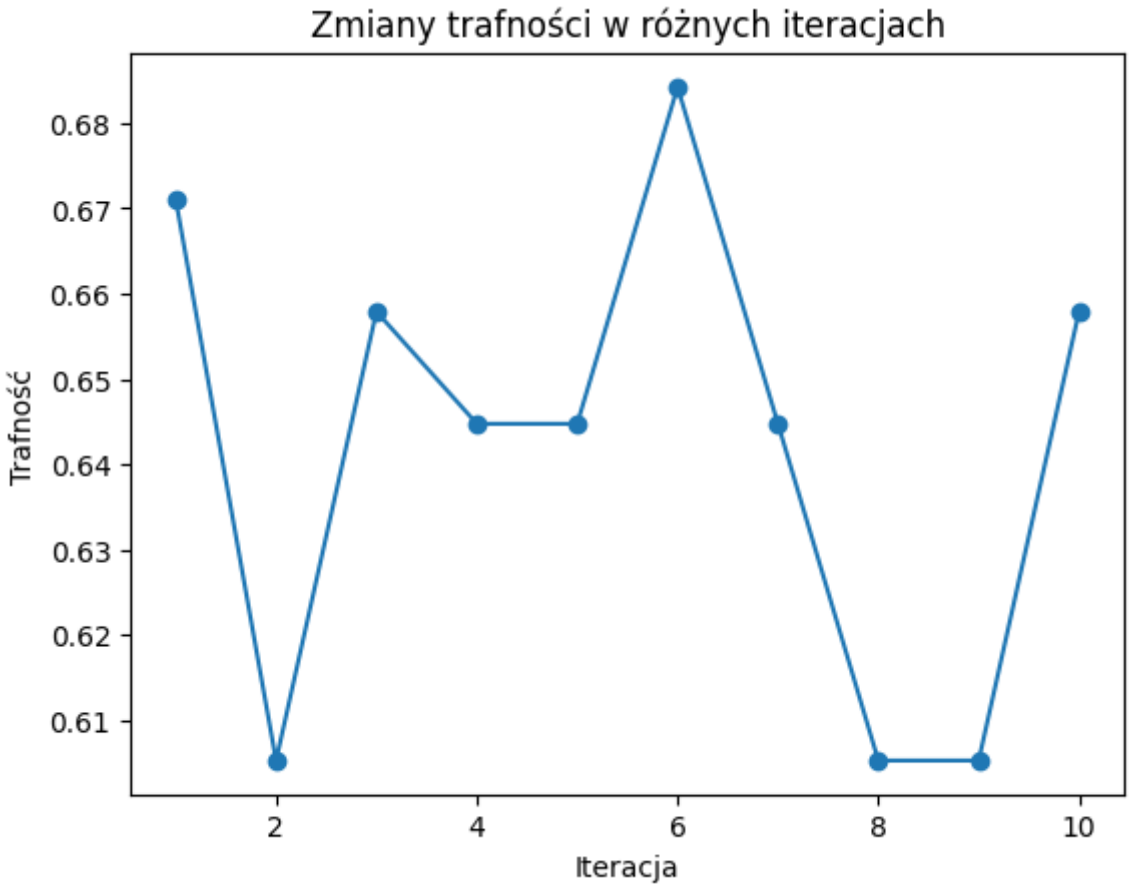
Teraz skupimy się na testowaniu danych pogrupowanych w następujący sposób (0 -> 0, 1 -> 1, 2 -> 1, 3 -> 3, 4 -> 4)

Wyniki dla Drzewa Decyzyjnego: (hiper parametry znalezione przez algorytm Grid Search: max_depth=2, criterion='entropy')

```
Accuracy: 0.6578947368421053
Precision: 0.5651147098515521
Recall: 0.6578947368421053
F1 measure: 0.6079514625363784
[[40  5  0  0]
 [10 10  0  0]
 [ 1  6  0  0]
 [ 1  3  0  0]]
```

Co ciekawe dla pojedynczego testowania modelu osiągamy wynik gorszy od modelu działającego na niepogrupowanych danych. Jednak jak zaraz zobaczymy w przypadku testowania przez 10 epok uzyskamy o wiele lepszy efekt.

```
Średnia trafność: 0.6421052631578947
Odchylenie standardowe trafności: 0.026836944808383084
Średnia Precyzja: 0.47982456140350893
Średnia Recall: 0.3458919530857305
Średnia F1: 0.31651609133367764
```

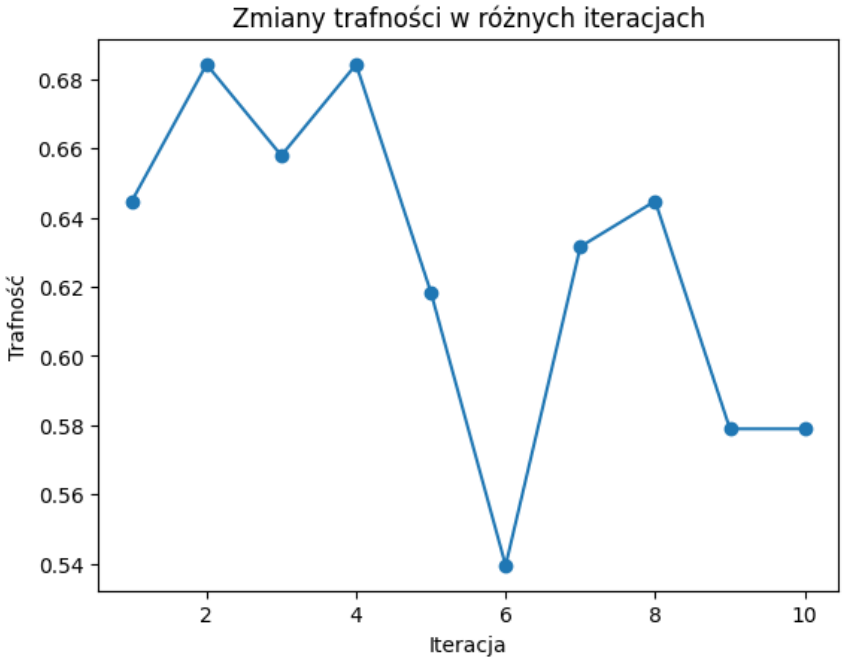


Wyniki dla klasyfikatora Bayesa:

```
Accuracy: 0.5921052631578947
Precision: 0.6736842105263158
Recall: 0.5921052631578947
F1 measure: 0.6045883940620783
[[41  1  0  3]
 [ 4  3  3 10]
 [ 0  1  0  6]
 [ 0  1  2  1]]
              precision    recall  f1-score   support

           0       0.91      0.91      0.91        45
           1       0.50      0.15      0.23        20
           3       0.00      0.00      0.00         7
           4       0.05      0.25      0.08         4

    accuracy                           0.59        76
   macro avg       0.37      0.33      0.31        76
weighted avg       0.67      0.59      0.60        76
```

Wyniki dla 10 iteracji:

```
Średnia trafność: 0.6263157894736843
Odchylenie standardowe trafności: 0.045275396142329616
Średnia Precyzja: 0.41431438315461444
Średnia Recall: 0.4028998858704851
Średnia F1: 0.3999613484429579
```



5.3 Dane pogrupowanie binarne

W ostatnim modelu dane grupujemy w następujący sposób (0 -> 0 1,2,3,4 -> 1), oceniamy jedynie czy pacjent jest chory czy zdrowy bez określenia stanu rozwinięcia choroby.

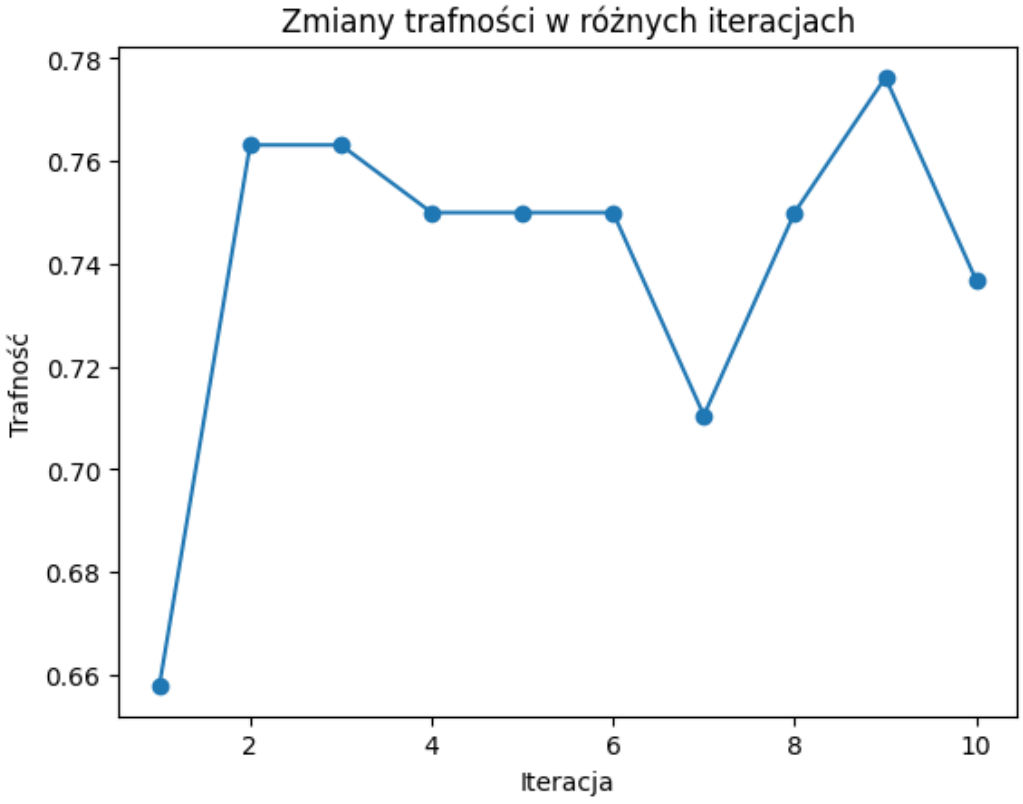
Wyniki dla drzewa decyzyjnego (hiper parametry - max_depth=6, criterion='entropy') :

```
Accuracy: 0.7631578947368421
Precision: 0.7641388417279759
Recall: 0.7631578947368421
F1 measure: 0.7614888171803716
[[34  7]
 [11 24]]
```

Zdecydowana poprawa w wynikach, wszystkie miary skuteczności na poziomie aż 76%.

Wyniki dla 10 iteracji:

```
Średnia trafność: 0.7407894736842104
Odchylenie standardowe trafności: 0.032256975452976996
Średnia Precyzja: 0.7593117939110426
Średnia Recall: 0.7269005669114994
Średnia F1: 0.7241781165940486
```



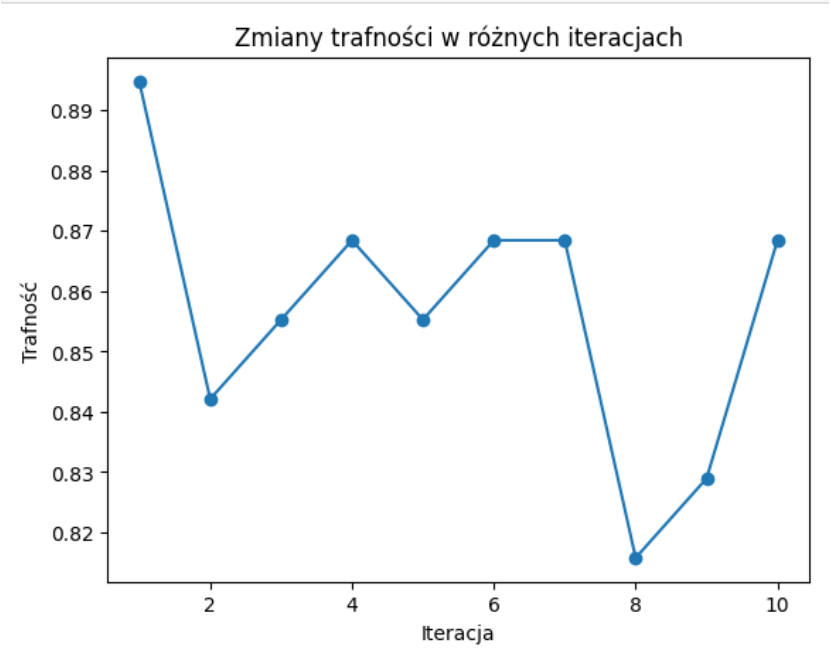
Wyniki dla Klasyfikatora Bayesa:

```
Accuracy: 0.8157894736842105
Precision: 0.817953420135705
Recall: 0.8157894736842105
F1 measure: 0.8163054695562435
[[35  8]
 [ 6 27]]
              precision    recall  f1-score   support

           0       0.85      0.81      0.83        43
           1       0.77      0.82      0.79        33

    accuracy                           0.82        76
   macro avg       0.81      0.82      0.81        76
weighted avg       0.82      0.82      0.82        76
```

Duża poprawa, 81% skuteczności, F1 miary, precyzji oraz recall jest to satysfakcjonujący nas wynik.

Wyniki dla 10 iteracji:

```
Średnia trafność: 0.8565789473684211
Odchylenie standardowe trafności: 0.02158055193007465
Średnia Precyzja: 0.8585293611828009
Średnia Recall: 0.8511031921103527
Średnia F1: 0.8523453318296903
```



Aż średnio 85% na przestrzeni wszystkich miar jest to satysfakcjonujący nas wynik i zdecydowanie lepszy od początkowego rezultatu.

6. Podsumowanie

Poniżej tabelki ze średnimi statystykami dla każdego pogrupowania danych:

- Niepogrupowane dane

No.	Model	Accuracy	Recall	Precision	F1 Score
1	Decision Tree	0.5236842105263159	0.5236842105263159	0.5375224833127467	0.2887365676154981
2	Bayes	0.536842105263158	0.536842105263158	0.5881403646947594	0.32687302798267137

- Dane pogrupowane niebinarnie

No.	Model	Accuracy	Recall	Precision	F1 Score
1	Decision Tree	0.6421052631578947	0.3458919530857305	0.47982456140350893	0.31651609133367764
2	Bayes	0.6263157894736843	0.4028998858704851	0.41431438315461444	0.3999613484429579

- Dane pogrupowane binarnie

No.	Model	Accuracy	Recall	Precision	F1 Score
1	Decision Tree	0.7407894736842104	0.7269005669114994	0.7593117939110426	0.7241781165940486
2	Bayes	0.8565789473684211	0.8511031921103527	0.8585293611828009	0.8523453318296903

Podsumowując grupowanie danych zdecydowanie pomogło nam polepszyć naszą skuteczność. W podsumowaniu będę się skupiał na wynikach uzyskanych przez 10 iteracji, ponieważ lepiej jesteśmy w stanie zauważyć jak zmieniają się wyniki w zależności od grupowania atrybutu decyzyjnego. Dla niepogrupowanych danych osiągamy wyniki, które nie spełniają naszych oczekiwań. Dla Drzewa decyzyjnego: 52% accuracy, 52% recall, 53% precyzji oraz jedynie 28% F1 score. Dla Bayesa: 53% accuracy, 53% recall, 58% precyzji oraz 32% miary F1. Pomimo wszystkich technik użytych przy przetwarzaniu danych są to słabe wyniki.

Przy testowaniu modelu na danych pogrupowanych niebinarnie, osiągamy trochę lepszą skuteczność jednak w dalszym ciągu dosyć przeciętną, warto zwrócić na dalej niską miarę F1. Osiągnięte wyniki to dla Drzewa decyzyjnego: 64% accuracy, 34% recall, 47% precyzji oraz 31% miary F1. Natomiast dla Naiwnego Klasyfikatora Bayesa są to: 62% accuracy, 40% recall, 41% precyzji i 39% F1 score. Dopiero w przypadku danych pogrupowanych binarnie bez uwzględnienia stanu choroby pacjenta osiągamy dobre rezultaty, szczególnie przy użyciu Bayesa. Uzyskane wyniki dla drzewa decyzyjnego to: 74% accuracy, 72% recall, 75% precision, 72% F1 score. Dla Bayesa: 85% accuracy, 85% recall, 85% precison, 85% F1 score.

Porównując stworzone modele możemy bardzo dobrze zauważyć jak bardzo grupowanie danych wpływa na wyniki naszych modeli. Pomimo zastosowania oversamplingu oraz normalizacji danych pierwszy model pracujący na niepogrupowanych danych osiąga nie najlepsze wyniki. Po zmapowaniu atrybutów decyzyjnych na 4 klasy zmiast 5 możemy bardzo dobrze zauważyć różnicę w wynikach drugiego modelu. 64% accuracy dla drzewa decyzyjnego oraz 62% dla klasyfikatora Bayesa nie są to już aż tak złe wyniki, natomiast bardzo dobrze obrazuje to jak bardzo tak małe grupowanie poprawia wyniki modelu. Ostatni trenowany model pracujący na danych pogrupowanych binarnie zdecydowanie wypadł najlepiej. Co ciekawe dopiero w przypadku tego modelu możemy zauważyć znaczną różnicę w skuteczności dwóch badanych klasyfikatorów i Naiwny Klasyfikator Bayesa sprawdził się zdecydowanie lepiej. 85% osiągnięte w każdej mierze możemy nazwać satysfakcjonującym wynikiem. Klasyfikator ten sprawdza się najlepiej dla analizy tego zbioru danych.

Podsumowując najlepsze rezultaty osiągamy dla danych pogrupowanych binarnie (w ten sposób nie określamy niestety stopnia zaawansowania choroby) używając klasyfikatora Bayesa. W przypadku tego klasyfikatora nie korzystamy z żadnych hiper parametrów więc jedyną możliwa poprawa wyniku była by możliwa przy użyciu zbioru danych z większą liczbą przypadków.

Niestety pomimo różnego rodzaju próby pracy z danymi, pierwszy model działający na danych oryginalnych nie był w stanie osiągnąć satysfakcjonujących nas danych . (najlepszy uzyskany wynik to 68% z użyciem drzewa decyzyjnego jednak na przestrzeni 10 iteracji, wynik ten bardzo spadł). Warto zauważyć bardzo niską F1 miarę w przypadku pierwszego modelu co może oznaczać że model ma trudności z identyfikacją negatywnych przypadków. Uzyskanie lepszych wyników byłoby być może możliwe w przypadku posiadania więcej danych lub być może przy użyciu innego modelu.