Raport zespołowy nr 1 – choroby serca

Jan Łukaszewicz, numer indeksu: 308187

Dominik Sobieraj, numer indeksu: 308208

Wszystkie skrypty, opisy i logika stojąca za wykonywanymi działaniami , aby osiągnąć wyniki przestawione w tym pliku znajdują się w osobnym załączniku o nazwie ‘models.ipynb’.

Model pierwszy: metoda MLP

Ziarno generatora liczb losowych:

$$\lfloor (308187 + 308208)/2 \rfloor = 308197$$

Z danych zostały usunięte wiersze z wartościami 0 w kolumnach ‘RestingBP’ i ‘Cholesterol’.

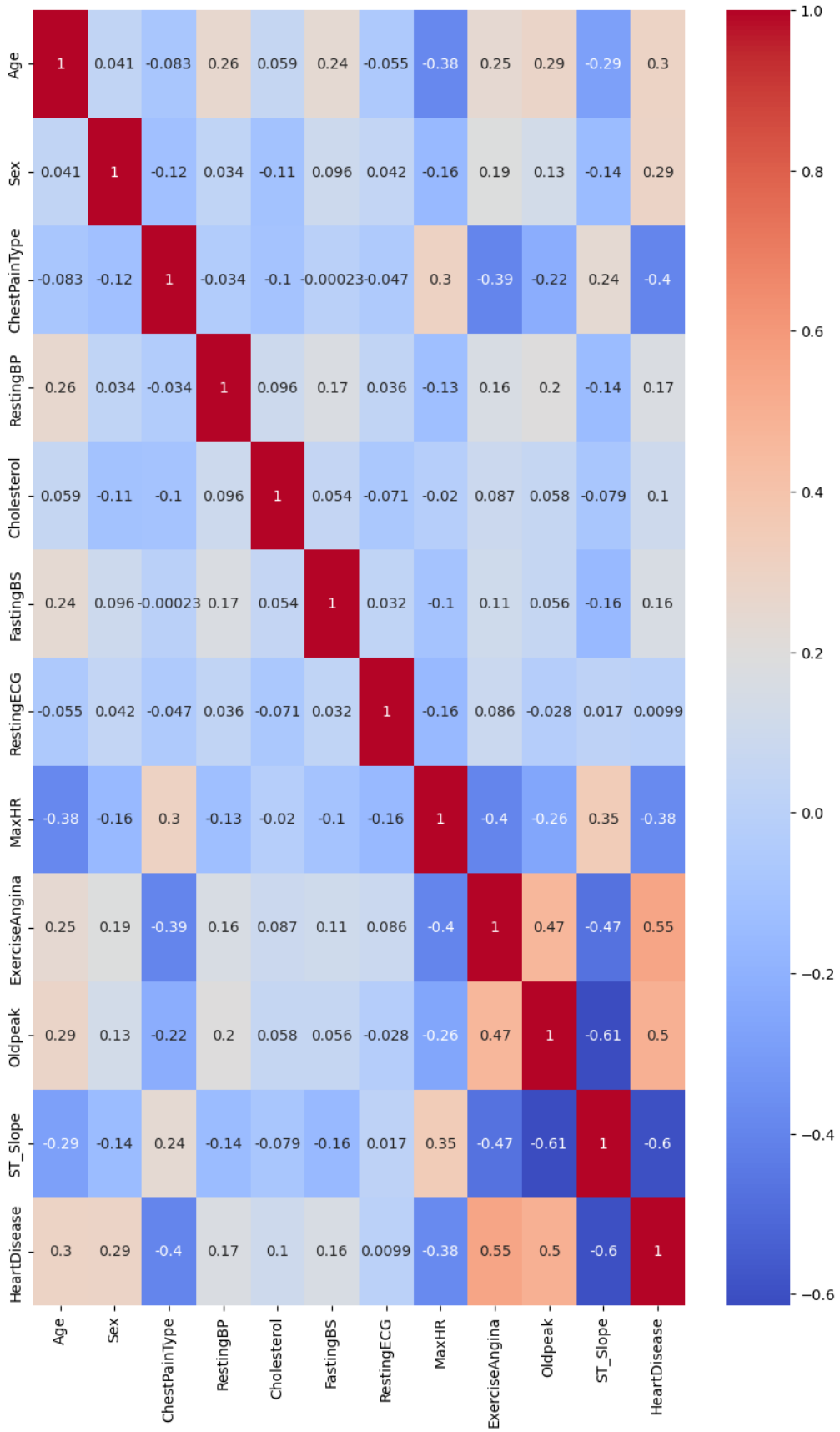
Po eksporacyjnej analizie danych kolumny ‘RestingECG’ i ‘Cholesterol’ zostały usunięte.

Parametry użyte w modelu MLP:

max_iter = 250

learning_rate = adaptive

Tabela korelacji danych heart-data.csv:



Opis sieci:

1. Sieć składa się z 3 warstw, warstwy wejściowej, jednej ukrytej i wyjściowej.

2. Liczba neuronów w warstwie wejściowej wynosi 16, ponieważ jest 5 zmiennych numerycznych, oraz 4 kategoryczne `['Sex', 'ChestPainType', 'ExerciseAngina', 'ST_Slope']` gdzie przy powyżej użytytego `OneHotEncodera`, aby można ich było użyc w sieci neuronowej zostały podzielone poszczególnie na:

a) `Sex` dzieli się na 2 zmienne, ponieważ ta zmienna zawierała dwie wartości.

b) `ChestPainType` dzieli się na 4 zmienne, ponieważ ta zmienna zawierała cztery wartości.

c) `ExerciseAngina` dzieli się na 2 zmienne, ponieważ ta zmienna zawierała dwie wartości.

d) `ST_Slope` dzieli się na 3 zmienne, ponieważ ta zmienna zawierała trzy wartości.

3. Domyślną liczbę neuronów w wartstwie ukrytej.

4. Jest to funkcja aktywacji, która aktywuje neuron.

5. Sieć ma jedną jednostkę w warstwie wyjściowej, ponieważ jest to klasyfikacja binarna.

6. Jest to funkcja sigmoidalna, która przekształca wynik na zakres [0, 1]. Została użyta, ponieważ sieć rozwiązuje problem klasyfikacji binarnej.

Macierz pomyłek, zbiór testowy:

	0	1
0	105	12
1	10	97

Macierz pomyłek, zbiór uczący:

	0	1
0	242	31
1	26	223

Patrząc na naiwny model, według którego każdy pacjent miałby choroby sercowe, co daje 50% trafności, sieć neuronowa jest zdecydowanie lepsza.

Analiza sieci dla danych testowych:

1. `Trafność` - wynosi aż 90%, co oznacza, że sieć dla pozytywnych i negatywnych przypadków, na 90% będzie zwróci poprawną odpowiedź.

2. `Czułość` - 91% świadczy, że taka jest szansa na poprawną identyfikacje pacjenta rzeczywiście chorego.

3. `Specyficzność` - 90% świadczy, że taka jest szansa na poprawną identyfikacje pacjenta niechorującego na serce.

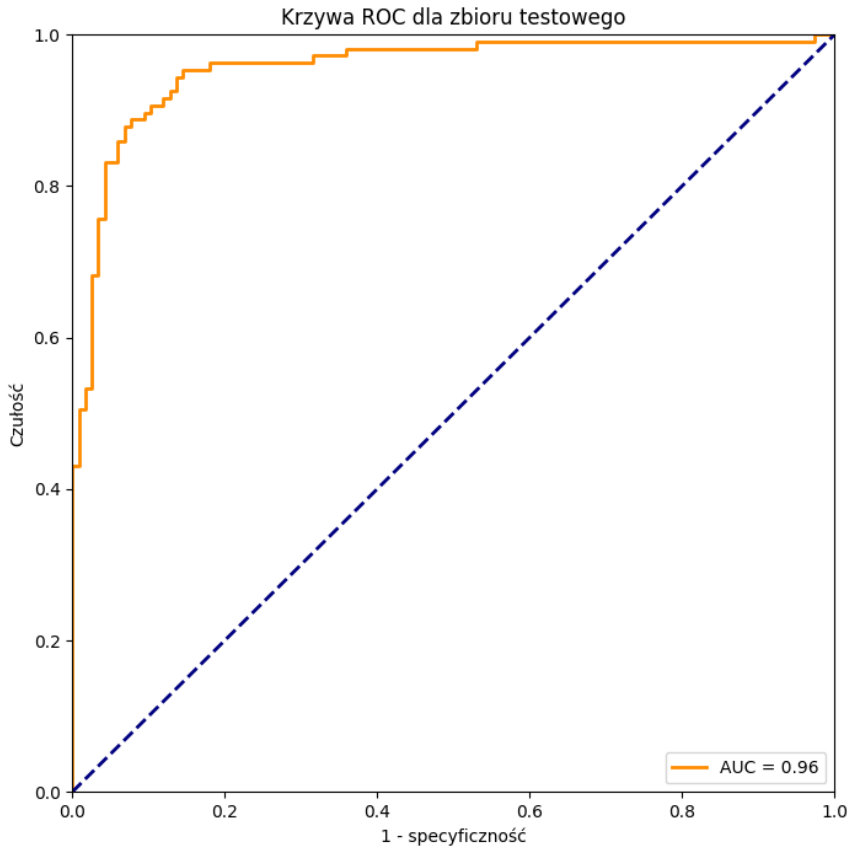
4. `Precyzja` na poziomie 89%, mówi o poprawnie zidentyfikowanych pozytywnych przypadków w porównaniu do wszystkich rzeczywistych pozytywnych przypadków.

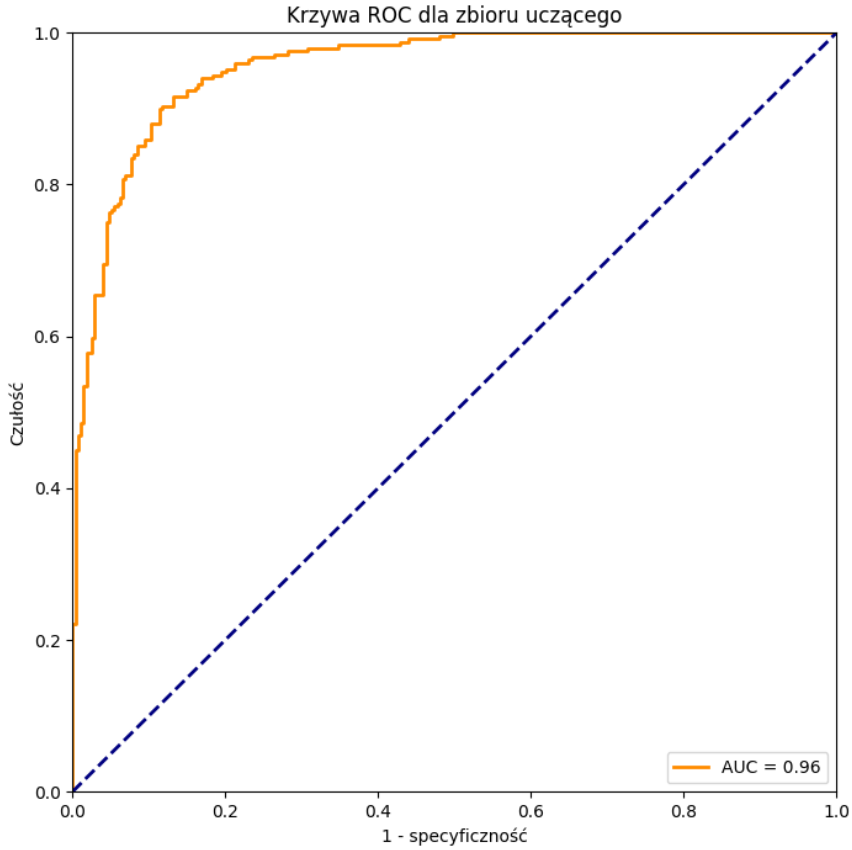
5. `Współczynnik F1` - czyli średnia harmoniczna między precyzją a czułością na poziomie 90%, świadczy że sieć dobrze radzi sobię z false posiitives i false negatives.

Ocena modelu dla zbioru testowego:
Trafność: 0.9
Całkowity współczynnik błędu 0.1
Czułość: 0.91
Wskaźnik fałszywie negatywnych: 0.09
Specyficzność: 0.9
Wskaźnik fałszywie pozytywnych: 0.1
Precyzja: 0.89
Wynik F1: 0.9

Ocena modelu dla zbioru uczącego:
Trafność: 0.89
Całkowity współczynnik błędu 0.11
Czułość: 0.9
Wskaźnik fałszywie negatywnych: 0.1
Specyficzność: 0.89
Wskaźnik fałszywie pozytywnych: 0.11
Precyzja: 0.88
Wynik F1: 0.89

Można zauważyć ciekawą zależność, otóż sieć radzi sobie delikatnie lepiej na danych testowych niż na uczących może to być spowodowane, małym zbiorem uczącym. Sieć łatwo przystosowywuje się do danych, ale może mieć trudność w generalizacji na nowe dane. W tym przypadku, zbiór testowy może być bardziej reprezentatywny dla rzeczywistego rozkładu danych, co pozwala na lepszą generalizację i w rezultacie lepsze wyniki.





Pole powierzchni pod ROC dla zbioru testowego : 0.956

Pole powierzchni pod ROC dla zbioru uczącego : 0.955

Model drugi: drzewa CaRT

Ziarno generatora liczb losowych:

$$\lfloor (308187 + 308208)/2 \rfloor = 308197$$

Z danych zostały usunięte wiersze z wartościami 0 w kolumnach ‘RestingBP’ i ‘Cholesterol’.

Po eksporacyjnej analizie danych kolumny ‘RestingECG’ i ‘Cholesterol’ zostały usunięte.

Drzewo zostało stworzone przy pomocy:

DecisionTreeClassifier

Random_state = 308197

Zbiór danych został podzielony na dwa podzbiory uczący 70% i testowy 30%.

Po dostarczeniu danych do drzewa macierze pomyłek wyglądały w ten sposób:

Macierz pomyłek, zbiór testowy:

	0	1
0	96	21
1	16	91

Macierz pomyłek, zbiór uczący:

	0	1
0	273	0
1	0	249

Widać, że na zbiorze uczącym algorytm nie popełnił prawie żadnych błędów, jednak jak spojrzymy na zbiór testowy to zwrócił tam o wiele gorsze wyniki.

Celność:

Zbiór uczący : 1.0
Zbiór testowy: 0.8348214285714286

Różnica wynosi prawie 17 punktów procentowych. Może to być spowodowane przeuczeniem klasyfikatora. Nie zastosowaliśmy ograniczeń na wzrost drzewa przez co prawdopodobnie się ono rozrosło.

Sprawdzamy wielkość drzewa:

Liczba wszystkich węzłów: 183
Liczba liści: 92
Głębokość: 14

Okazało się, że nie nadaliśmy ograniczenia na możliwość rozrastania się drzewa, przez co drzewo rozrosło się do dużych rozmiarów i wszystkie wyniki zostały zapamiętane przez drzewo w danych uczących.

Po ograniczeniu rozrostu drzewa otrzymaliśmy takie wyniki:

Macierz pomyłek, zbiór testowy:

	0	1
0	93	24
1	15	92

Macierz pomyłek, zbiór uczący:

	0	1
0	211	62
1	30	219

Jak widać wyniki są bardzo zbliżone do siebie tym razem i celność prezentuję się w taki sposób:

Zbiór uczący : 0.8237547892720306
Zbiór testowy: 0.8258928571428571

Patrząc na naiwny model, według którego każdy pacjent miałby choroby sercowe, co daje 50% trafności, trzewo Cart jest troche lepsze.

Jednak nie jest to idealne rozwiązanie

Analiza sieci dla danych testowych:

1. `Trafność` - wynosi około 83%, co oznacza, że dla pozytywnych i negatywnych przypadków, na 83% będzie zwróci poprawną odpowiedź.

2. `Czułość` - 86% świadczy, że taka jest szansa na poprawną identyfikacje pacjenta rzeczywiście chorego.

3. `Specyficzność` - 79% świadczy, że taka jest szansa na poprawną identyfikacje pacjenta niechorującego na serce.

4. `Precyzja` na poziomie 79%, mówi o poprawnie zidentyfikowanych pozytywnych przypadków w porównaniu do wszystkich rzeczywistych pozytywnych przypadków.

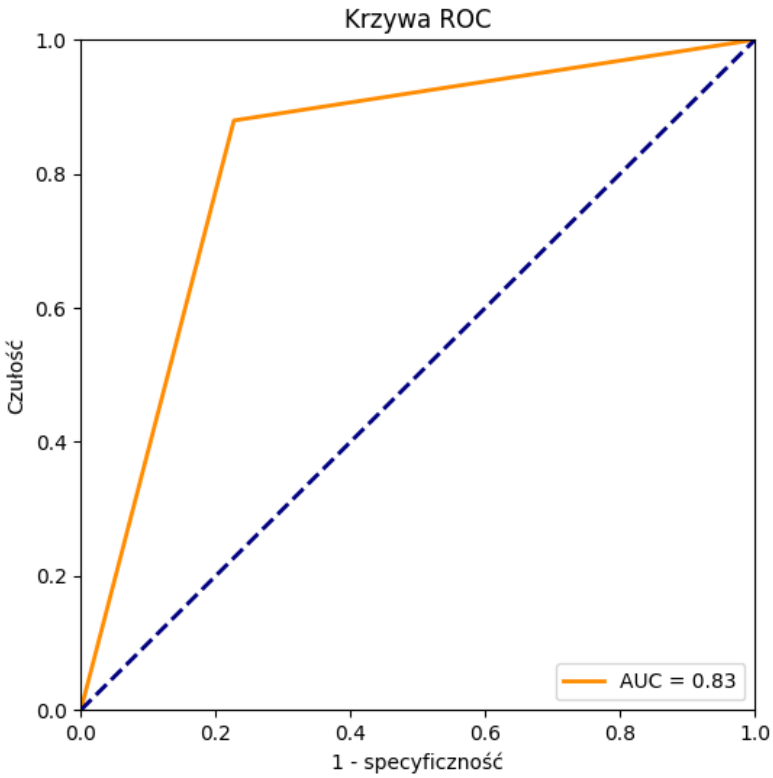
5. `Współczynnik F1` - czyli średnia harmoniczna między precyzją a czułością na poziomie 83%, świadczy że sieć dobrze radzi sobię z false posiitives i false negatives.

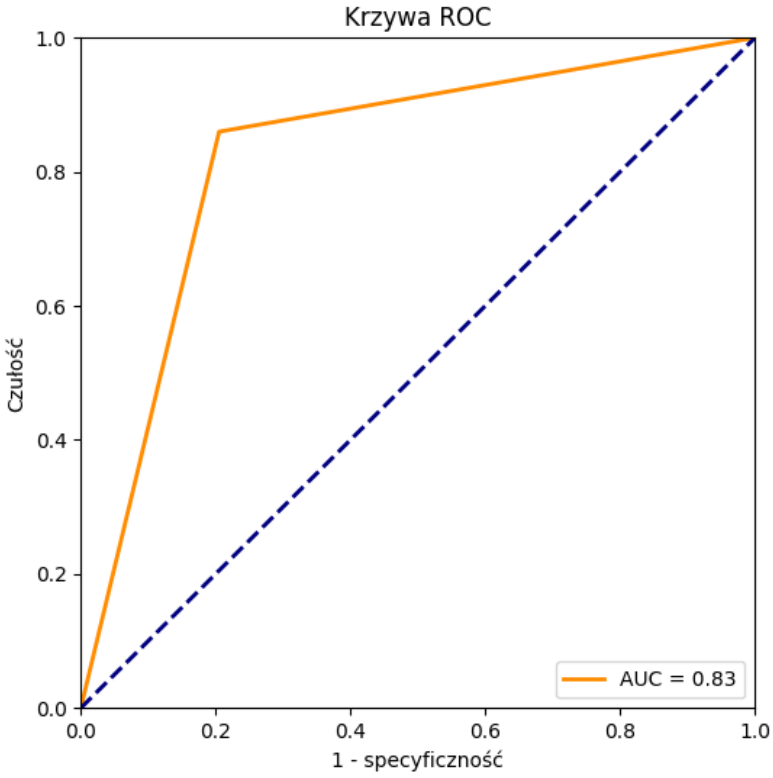
Zbiór uczący

Trafność: 0.82
Całkowity współczynnik błędu 0.18
Czułość: 0.88
Wskaźnik fałszywie negatywnych: 0.12
Specyficzność: 0.77
Wskaźnik fałszywie pozytywnych: 0.23
Precyzja: 0.78
Wynik F1: 0.83

Zbiór testowy

Trafność: 0.83
Całkowity współczynnik błędu 0.17
Czułość: 0.86
Wskaźnik fałszywie negatywnych: 0.14
Specyficzność: 0.79
Wskaźnik fałszywie pozytywnych: 0.21
Precyzja: 0.79
Wynik F1: 0.83





Pole powierzchni pod ROC dla zbioru testowego : 0.827
Pole powierzchni pod ROC dla zbioru uczącego : 0.826

Podsumowanie

Po zastosowaniu modelu MLP jak i drzewa CaRT, stwierdzamy, że skuteczniejszym wyborem w przypadku tych danych do ocenienia i przewidywania choroby serca jest model MLP.

Cechuje się on wyższa celnością jak i wiekszością innych statystyk w porównaniu do drzewa CaRT.