Raport wykonali : Anna Zgrzebna
Sebastian Nowosielski

Budowa modelu przewidującego wystąpienie choroby serca u pacjenta.

1. Cel :
Ocena poprawności przewidywań modeli CART i lasów losowych przy przewidywaniu wystąpienia choroby serca u pacjenta.

2. Wykonanie :
W pliku heart_disease.csv zawierający dane pacjentów kardiologicznych zmienna cel jest heart_disease – brak (0) lub obecność (1) choroby serca.

Do stworzenia naszyc modeli wybraliśmy metode CART i las losowy. Są one bardziej zawansowane niż metoda k-nn i nie potrzebują dużej ilości zmiennych tak jak MLP. Metoda CART polega na rozdzielaniu obserwacji używając predyktorów. Obserwacje spełniające warunki idą do prawego węzła, a nie spełniające do lewego. Jest to drzewo binarne. Las losowy składa się z więcej niż jednego drzewa CART i jest bardziej odporny na przuczenia. Generator liczb losowych został ustawiony na wartość 308 289.

Korzystając z macierzy korelacji (tabele 1a i 1b) wybieramy najlepsze predyktory tzn. takie które mają największy współczynnik korelacji ze zmienną celu i jak najmniejsza z innymi predyktorami.

	age	sex	chest_pain_type	resting_blood_pressure	serum_cholestoral	fasting_blood_sugar
age	1.000000	-0.094401	0.096920	0.273053	0.220056	0.123458
sex	-0.094401	1.000000	0.034636	-0.062693	-0.201647	0.042140
chest_pain_type	0.096920	0.034636	1.000000	-0.043196	0.090465	-0.098537
resting_blood_pressure	0.273053	-0.062693	-0.043196	1.000000	0.173019	0.155681
serum_cholestoral	0.220056	-0.201647	0.090465	0.173019	1.000000	0.025186
fasting_blood_sugar	0.123458	0.042140	-0.098537	0.155681	0.025186	1.000000
resting_elect	0.128171	0.039253	0.074325	0.116157	0.167652	0.053499
max_heart_rate	-0.402215	-0.076101	-0.317682	-0.039136	-0.018739	0.022494
angina	0.098297	0.180022	0.353160	0.082793	0.078243	-0.004107
oldpeak	0.194234	0.097412	0.167244	0.222800	0.027709	-0.025538
slope	0.159774	0.050545	0.136900	0.142472	-0.005755	0.044076
vessel	0.356081	0.086830	0.225890	0.085697	0.126541	0.123774
thalassemia	0.106100	0.391046	0.262659	0.132045	0.028836	0.049237
heart_disease	0.212322	0.297721	0.417436	0.155383	0.118021	-0.016319

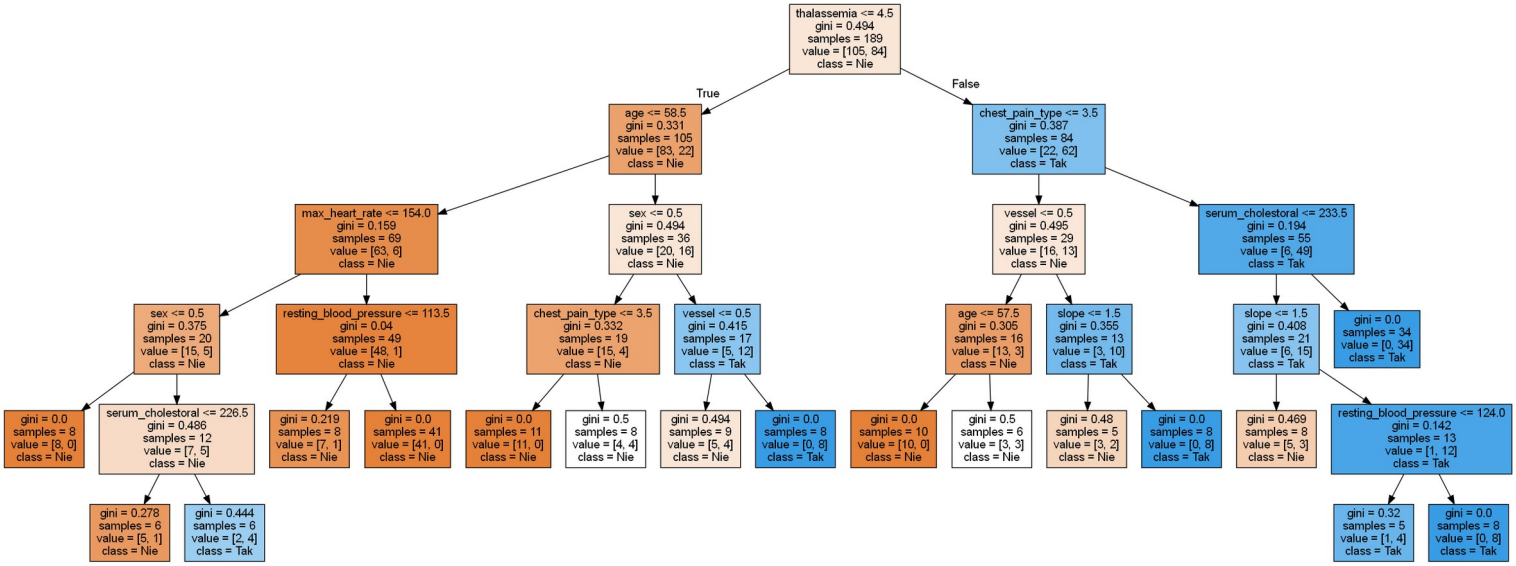
Tabela 1a

resting_elect	max_heart_rate	angina	oldpeak	slope	vessel	thalassemia	heart_disease
0.128171	-0.402215	0.098297	0.194234	0.159774	0.356081	0.106100	0.212322
0.039253	-0.076101	0.180022	0.097412	0.050545	0.086830	0.391046	0.297721
0.074325	-0.317682	0.353160	0.167244	0.136900	0.225890	0.262659	0.417436
0.116157	-0.039136	0.082793	0.222800	0.142472	0.085697	0.132045	0.155383
0.167652	-0.018739	0.078243	0.027709	-0.005755	0.126541	0.028836	0.118021
0.053499	0.022494	-0.004107	-0.025538	0.044076	0.123774	0.049237	-0.016319
1.000000	-0.074628	0.095098	0.120034	0.160614	0.114368	0.007337	0.182091
-0.074628	1.000000	-0.380719	-0.349045	-0.386847	-0.265333	-0.253397	-0.418514
0.095098	-0.380719	1.000000	0.274672	0.255908	0.153347	0.321449	0.419303
0.120034	-0.349045	0.274672	1.000000	0.609712	0.255005	0.324333	0.417967
0.160614	-0.386847	0.255908	0.609712	1.000000	0.109498	0.283678	0.337616
0.114368	-0.265333	0.153347	0.255005	0.109498	1.000000	0.255648	0.455336
0.007337	-0.253397	0.321449	0.324333	0.283678	0.255648	1.000000	0.525020
0.182091	-0.418514	0.419303	0.417967	0.337616	0.455336	0.525020	1.000000

Tabela 1b

W naszych modelach wykorzystamy wszystkie zmienne za wyjątkiem age, sex, resting_blood_pressure, serum_cholestoral ,fasting_blood_sugar ponieważ są słabo skorelowanie z zmienna heart_disease.

CART



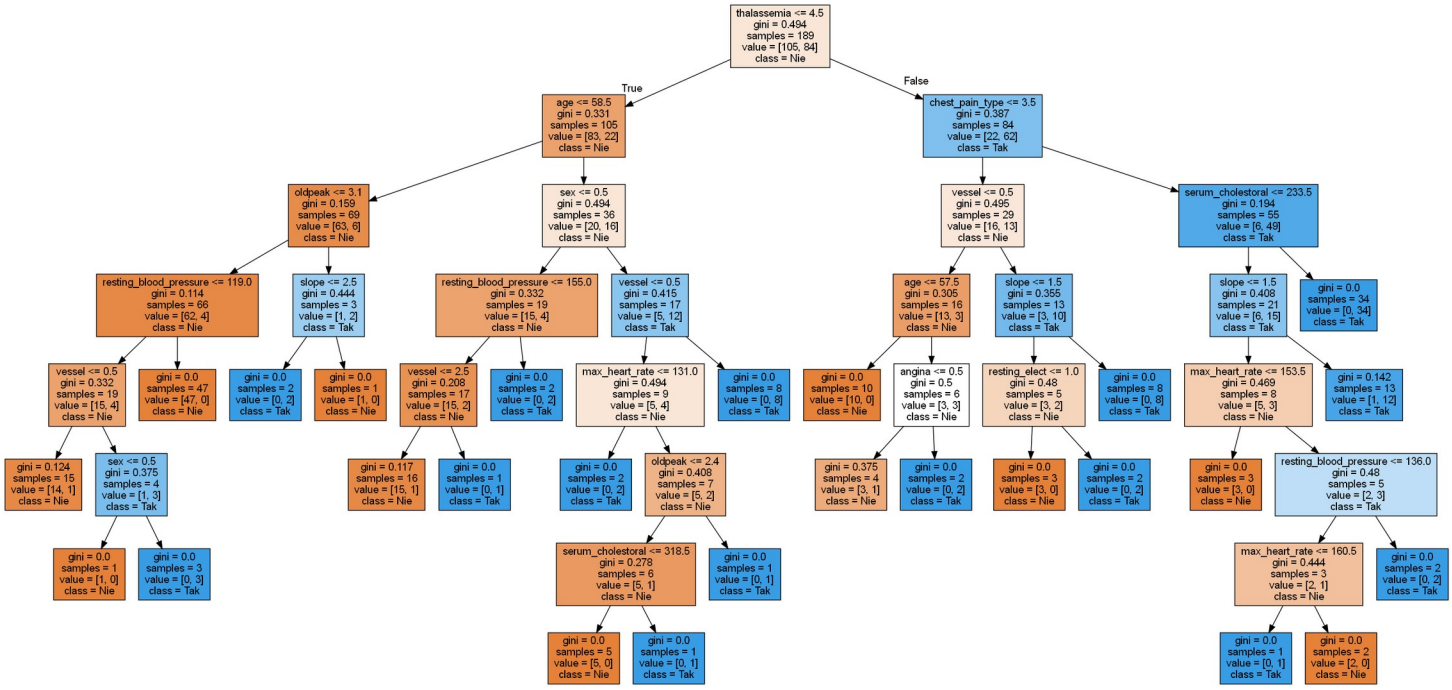
Rysunek 1

Rysunek 1 przedstawia drzewo CART zbudowane na podstawie powyższych predyktorów. Parametry określające trafność naszego testu:

Zbiór uczący	Zbiór testowy
Trafność: 0.89	Trafność: 0.77
Czułość: 0.79	Czułość: 0.61
Specyficzność: 0.97	Specyficzność: 0.89

F1 score:
Zbiór uczący ~0.8627
Zbiór testowy ~0.6984

Model jest przeuczony. Spróbujmy więc przyciąć nasze drzewo z parametrem ccp_alpha=0.005:



Zbiór uczący

Trafność: 0.98
Czułość: 0.96
Specyficzność: 0.99

Zbiór testowy

Trafność: 0.73
Czułość: 0.7
Specyficzność: 0.75

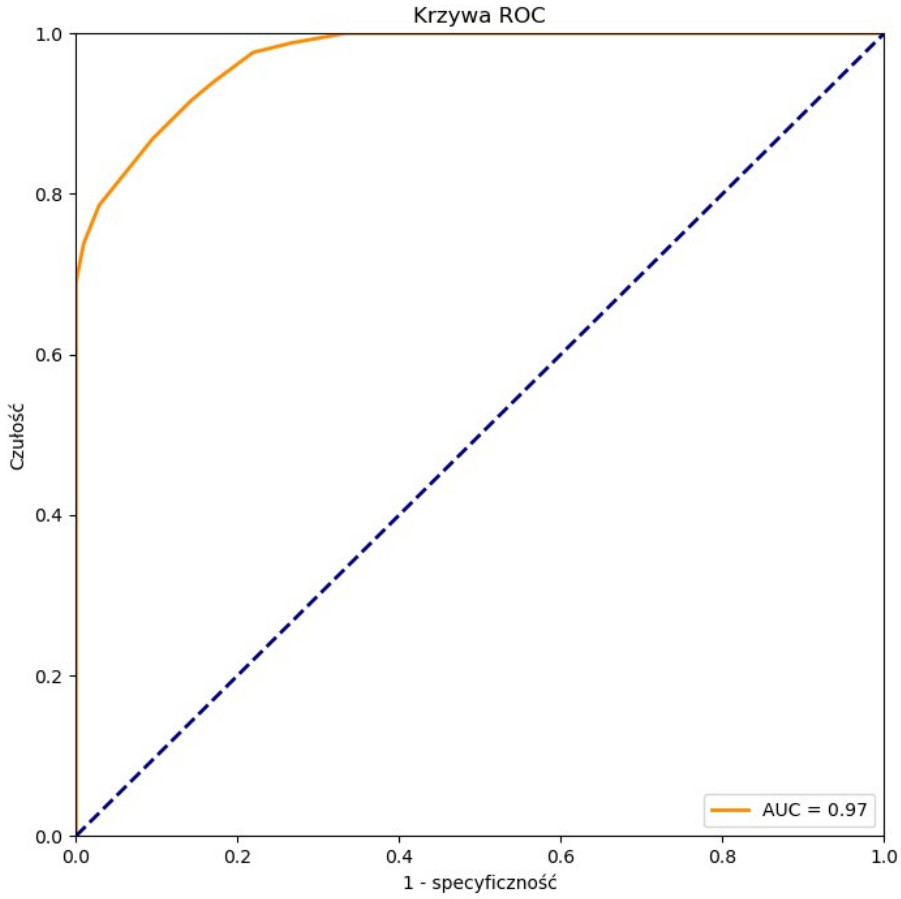
F1 score:
Zbiór uczący ~0.9759
Zbiór testowy ~0.7142

Porównując wartości F1 score ten model wypada lepiej, ale nadal jest przeuczony. Zostańmy więc przy pierwszym modelu i przeanalizujemy jego krzywe ROC.

Zbiór testowy

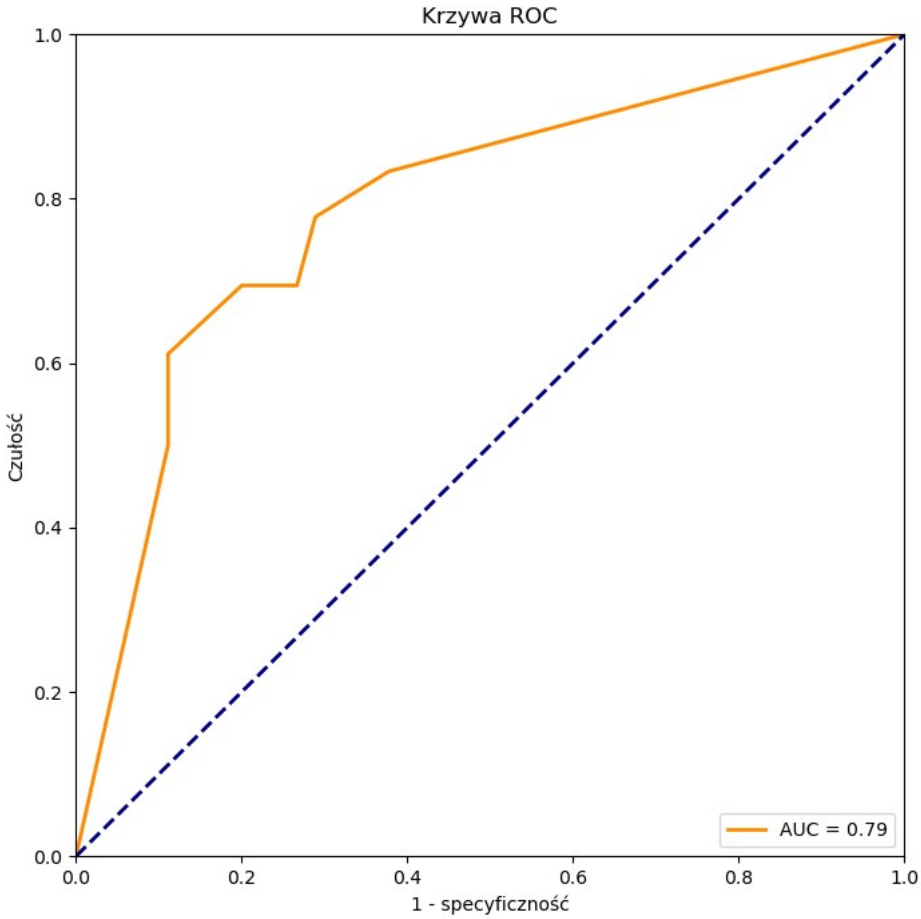
Krzywa ROC dla zbioru testowego wskazuje na dobre przewidywanie wartosci zmiennej celu na zbiorze testowym. Zbliża się ona do punktu (1,0).
Pole pod krzywą to 0,97 więc bardzo duże.

wykres 1



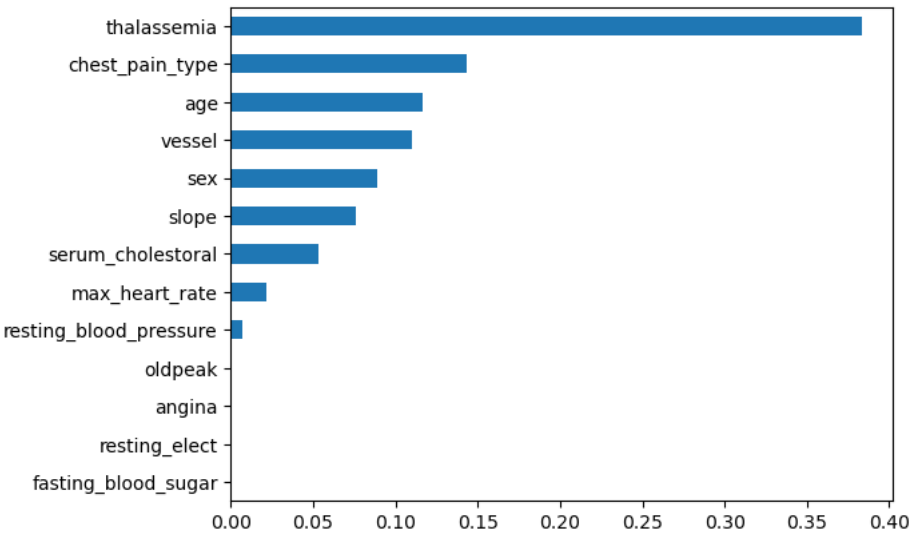
Zbiór uczący

Jak można było przewidzieć krzywa nie zbliża się bardzo do punktu (1,0).
Pole pod krzywą to 0,79.



wykres 2

Ważności:



thalassemia, chest_pain_type, age i vessel są zmiennymi które model używał najczęściej do podziału na grupy. Oldpeak, angina, resting_elect, fasting_blood_sugar model nie użył ani razu. Jeśli usuniemy te predyktory z modelu, wynik się nie zmieni.

Las losowy

Weźmy teraz 100 drzew CART i stwórzmy z nich las losowy. Parametry określające trafność naszego testu wyglądają następująco :

Zbiór uczący
Trafność: 0.87
Czułość: 0.73
Specyficzność: 0.96

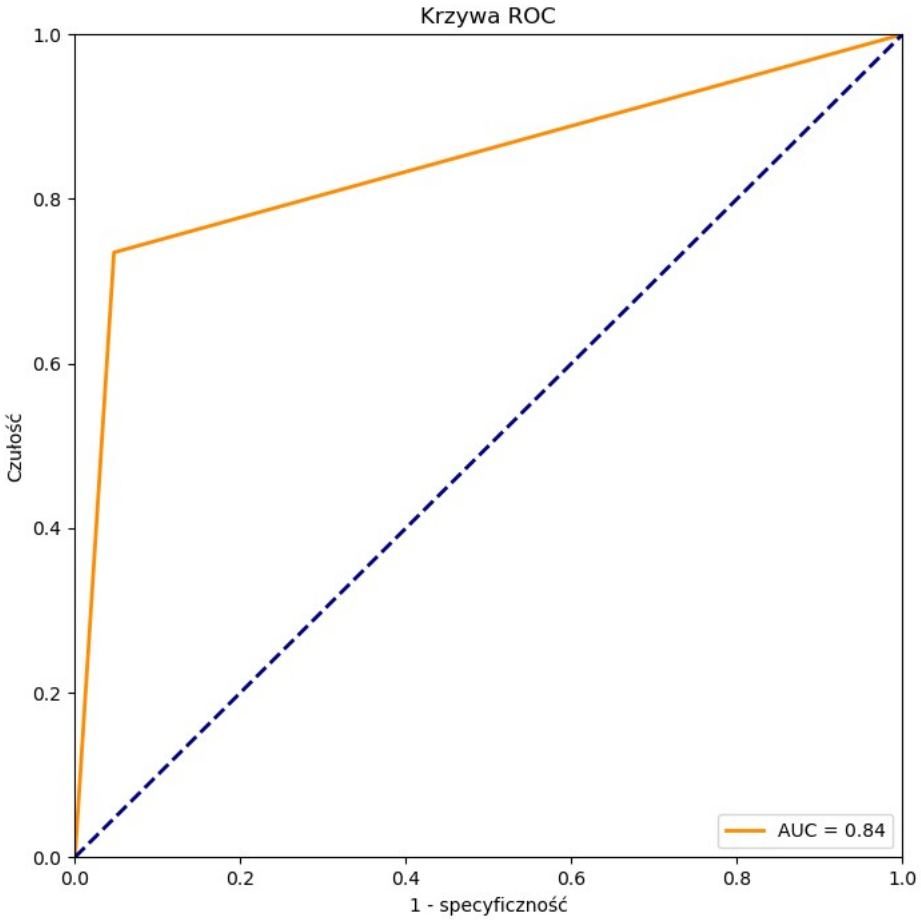
Zbiór testowy
Trafność: 0.81
Czułość: 0.74
Specyficzność: 0.90

F1 score:
Zbiór uczący ~0.8201
Zbiór testowy ~0.8051

Model nie jest przeuczony. Trafność, czułość i specyficzność są na podobnym poziomie dla obu zbiorów. Nasz model jest w stanie poprwanie przewidzieć obecność choroby serca u ok. 81% pacjentów i brak choroby u ok. 90%.

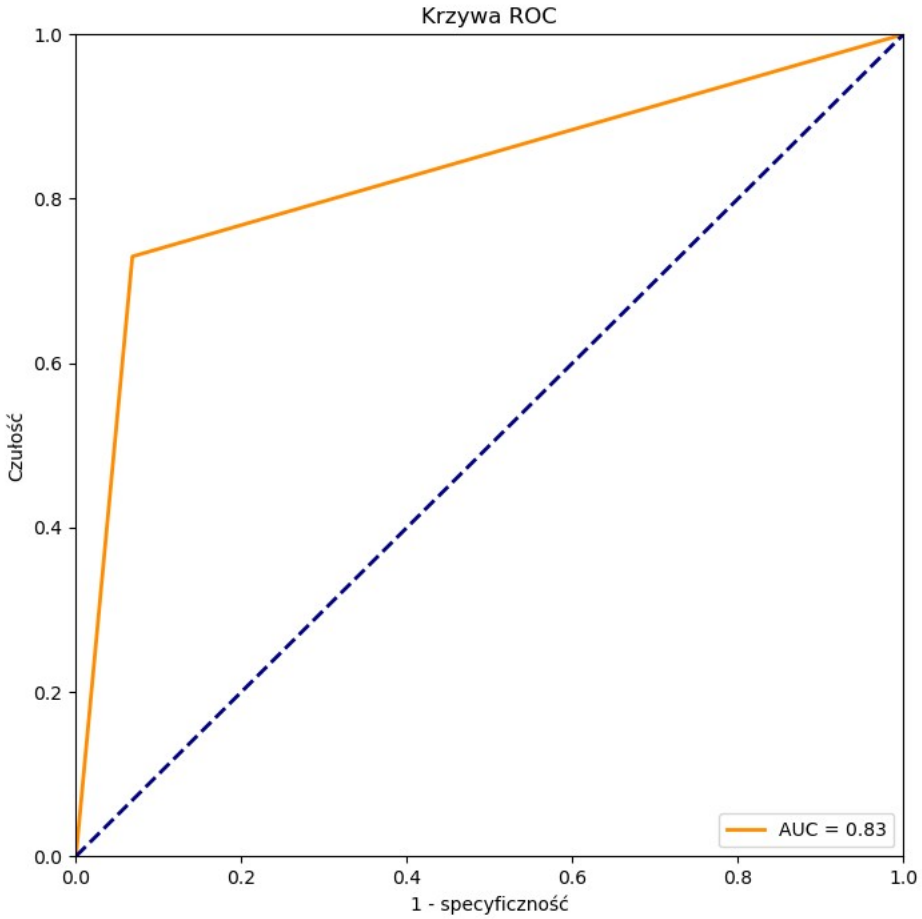
Zbiór uczący

Krzywa wypada gorzej w porównaniu do krzwej ROC z drzewa CART. Jest tak dlatego że model CART zbyt dobrze dopasował się do obserwacji ze zbioru testowego, a las losowy nie. Pole pod krzywa to 0.84.

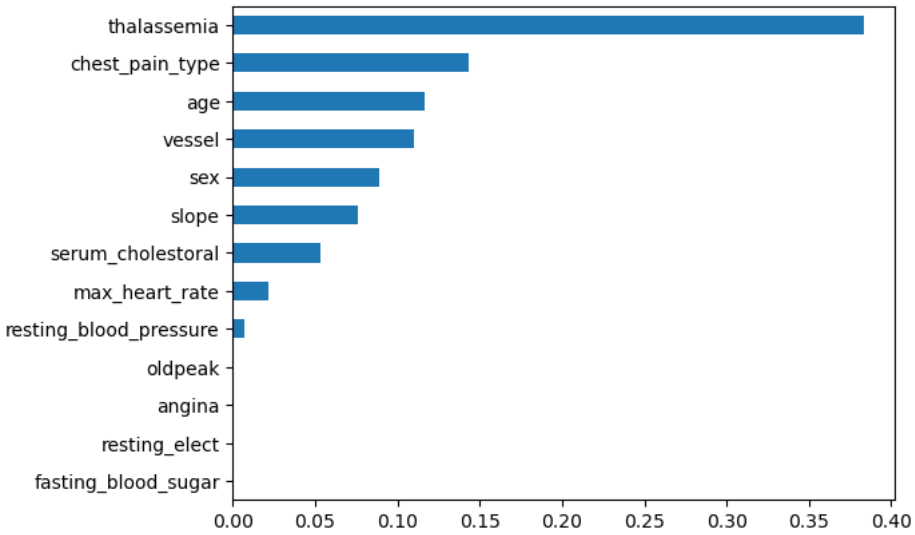


Zbiór uczący

Obie krzywe są niemalże identyczne. Ich pola różnią się o 0.01 czyli model dobrze przewiduje wartości zmiennej celu.



Ważności:



Te same predyktory co w drzewie CART zostały uznane za najlepiej dzielące obserwacje.

Podsumowanie:

Drzewo CART przeuczyło się, czego nie mogliśmy zmienić przycięciem go. Stworzenie za jego pomocą lasu losowego złożonego ze 100 drzew dało znacznie lepsze wyniki bo lasy są bardziej odporne na przeuczenie. Oba modele dały lepsze rezultaty niż przypisanie wszystkich pacjętów do grupy chorych, która ok. 44% osob wskazała by poprawnie jako chorych.