Raport 2

Realizacja zadań przy pomocy języka Python.

PROBLEM: Klasyfikacja obserwacji ze względu na wartość zmiennej *heart_disease*.

MODEL 1 – metoda drzewo CART

Najpierw zbadamy korelację między zmiennymi. Jest to nam potrzebne do wybrania predyktorów, które będą kluczowe dla naszego modelu. One będą decydowały o klasyfikacji obserwacji naszej zmiennej celu tj. *'heart_disease'*. Przyjmuje ona wartość 0, oznaczająca brak choroby oraz 1, gdy choroba serca występuje.

W następnym kroku dokonujemy podziału na zbiór uczący (30% wszystkich obserwacji) i testowy (70% wszystkich obserwacji). Ziarno generatora liczb losowych jest równe 308272.

Macierz korelacji zmiennych wygląda następująco:

Index	age	sex	chest_pain_type	resting_blood_pressure	serum_cholestoral	fasting_blood_sugar	resting_elect	max_heart_rate	angina	oldpeak	slope	vessel	thalassemia	heart_disease
age	1	-0.0944007	0.0969198	0.273053	0.220056	0.123458	0.128171	-0.402215	0.0982965	0.194234	0.159774	0.356081	0.1061	0.212322
sex	-0.0944007	1	0.0346356	-0.0626934	-0.201647	0.0421397	0.0392535	-0.0761015	0.180022	0.0974119	0.0505448	0.0868299	0.391046	0.297721
chest_pain_type	0.0969198	0.0346356	1	-0.0431961	0.0904652	-0.0985368	0.0743252	-0.317682	0.35316	0.167244	0.1369	0.22589	0.262659	0.417436
resting_blood_pressure	0.273053	-0.0626934	-0.0431961	1	0.173019	0.155681	0.116157	-0.0391357	0.0827926	0.2228	0.142472	0.0856974	0.132045	0.155383
serum_cholestoral	0.220056	-0.201647	0.0904652	0.173019	1	0.0251859	0.167652	-0.0187392	0.0782425	0.0277092	-0.00575528	0.126541	0.0288361	0.118021
fasting_blood_sugar	0.123458	0.0421397	-0.0985368	0.155681	0.0251859	1	0.0534988	0.0224942	-0.00410716	-0.0255379	0.044076	0.123774	0.0492375	-0.0163188
resting_elect	0.128171	0.0392535	0.0743252	0.116157	0.167652	0.0534988	1	-0.0746275	0.0950984	0.120034	0.160614	0.114368	0.00733721	0.182091
max_heart_rate	-0.402215	-0.0761015	-0.317682	-0.0391357	-0.0187392	0.0224942	-0.0746275	1	-0.380719	-0.349045	-0.386847	-0.265333	-0.253397	-0.418514
angina	0.0982965	0.180022	0.35316	0.0827926	0.0782425	-0.00410716	0.0950984	-0.380719	1	0.274672	0.255908	0.153347	0.321449	0.419303
oldpeak	0.194234	0.0974119	0.167244	0.2228	0.0277092	-0.0255379	0.120034	-0.349045	0.274672	1	0.609712	0.255005	0.324333	0.417967
slope	0.159774	0.0505448	0.1369	0.142472	-0.00575528	0.044076	0.160614	-0.386847	0.255908	0.609712	1	0.109498	0.283678	0.337616
vessel	0.356081	0.0868299	0.22589	0.0856974	0.126541	0.123774	0.114368	-0.265333	0.153347	0.255005	0.109498	1	0.255648	0.455336
thalassemia	0.1061	0.391046	0.262659	0.132045	0.0288361	0.0492375	0.00733721	-0.253397	0.321449	0.324333	0.283678	0.255648	1	0.52502
heart_disease	0.212322	0.297721	0.417436	0.155383	0.118021	-0.0163188	0.182091	-0.418514	0.419303	0.417967	0.337616	0.455336	0.52502	1

Jako predyktory wybieramy zmienne, których współczynnik korelacji ze zmienną “heart_disease” jest możliwie największy.
Kandydaci na predykory to zmienne: "chest_pain_type", "angina", "slope", "oldpeak", "max_heart_rate", "vessel", "thalassemia". Można dostrzec, że zmienne "slope" oraz "oldpeak" są w korelacji, więc możemy wybrać jedną z nich w celu uproszczenia modelu.
Ostatecznie nasze predyktory to: "chest_pain_type", "angina", "slope", "max_heart_rate", "vessel", "thalassemia".

MODEL 1.1, czyli podejście pierwsze

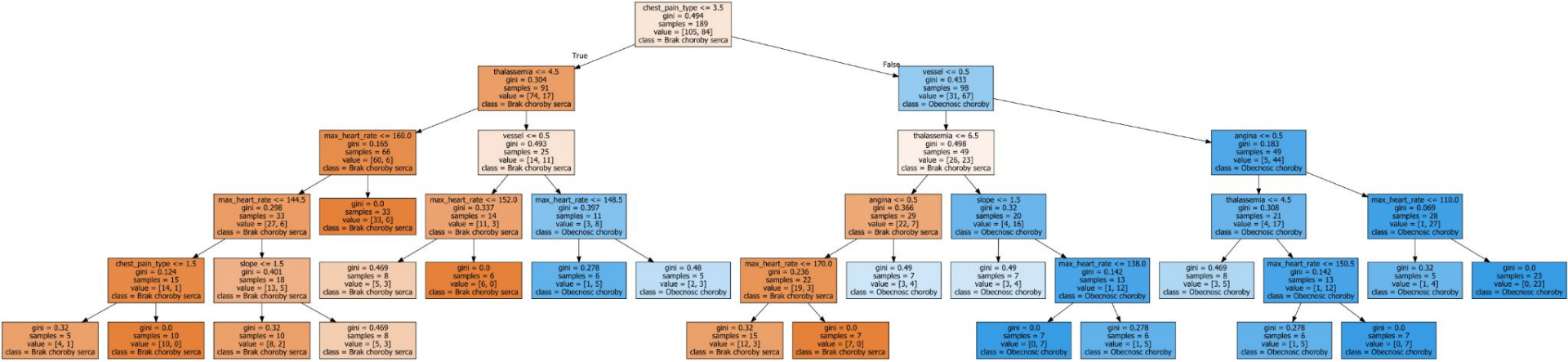
Zbiór testowy zawiera 189 obserwacji, zbiór uczący zaś 81 obserwacje.
W tym podejściu współczynniki prezentują się następująco:

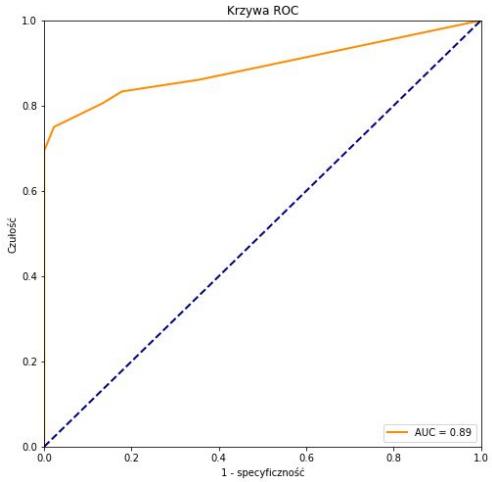
zbiór testowy	zbiór uczący
Trafność: 0.84 Czułość: 0.806 Specyficzność: 0.867	Trafność: 0.857 Czułość: 0.857 Specyficzność: 0.857

Zbiór testowy wypada gorzej (o najwięcej 5%) niż uczący, co wskazuje na lekkie przeuczenie się modelu.

Na następnej stronie przedstawiony jest wykres drzewa CART.

Model 1.1 Drzewo CART

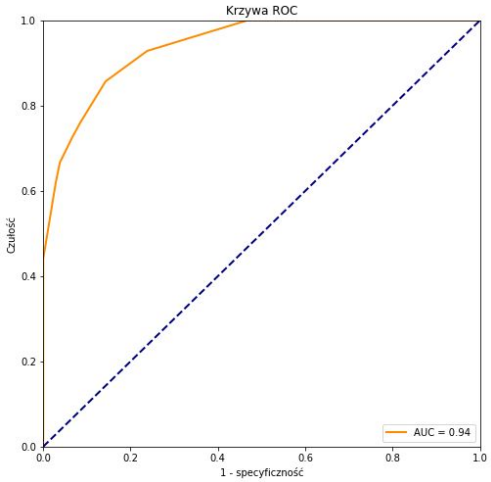




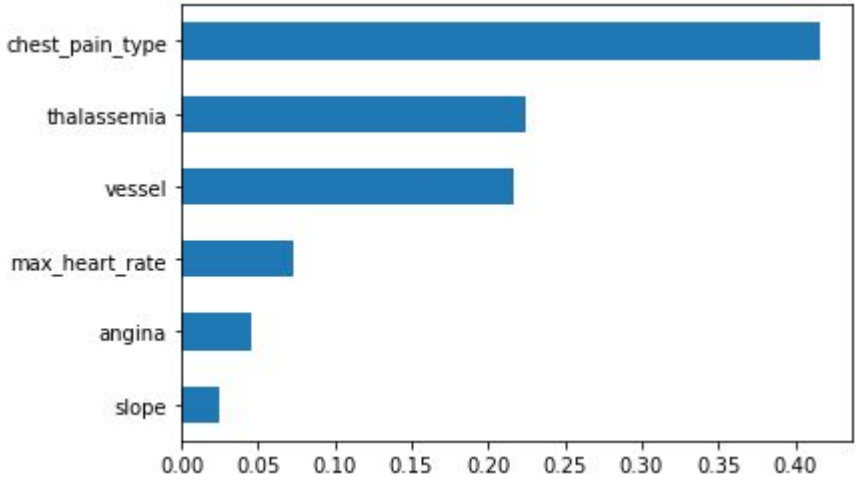
Spójrzmy jeszcze na krzywe ROC:

← *zbiór testowy*

zbiór uczący →



Spójrzmy na ważność predyktorów według modelu 1.1:



Najważniejsze predyktory to zmienne "chest_pain_type", "vessel", "thalassemia". Aby nasz model był lepszy zmienimy parametr alpha, mówiący o przycinaniu minimalnego kosztu i złożoności.

W wyniku tych działań powstanie model 1.2, czyli podejście drugie.

MODEL 1.2, czyli podejście drugie

Predyktory zostają te same jak w modelu 1.1.

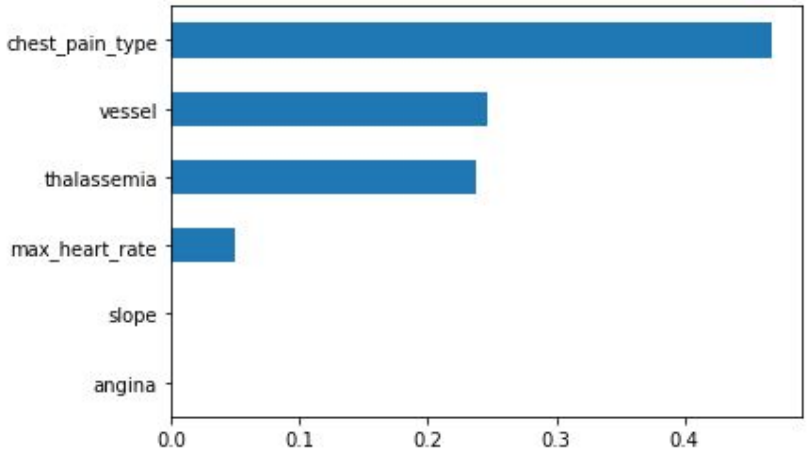
Podział na zbiór uczący i testowy także jest ten sam jak wcześniej.

Współczynniki dla modelu 1.2 wynoszą:

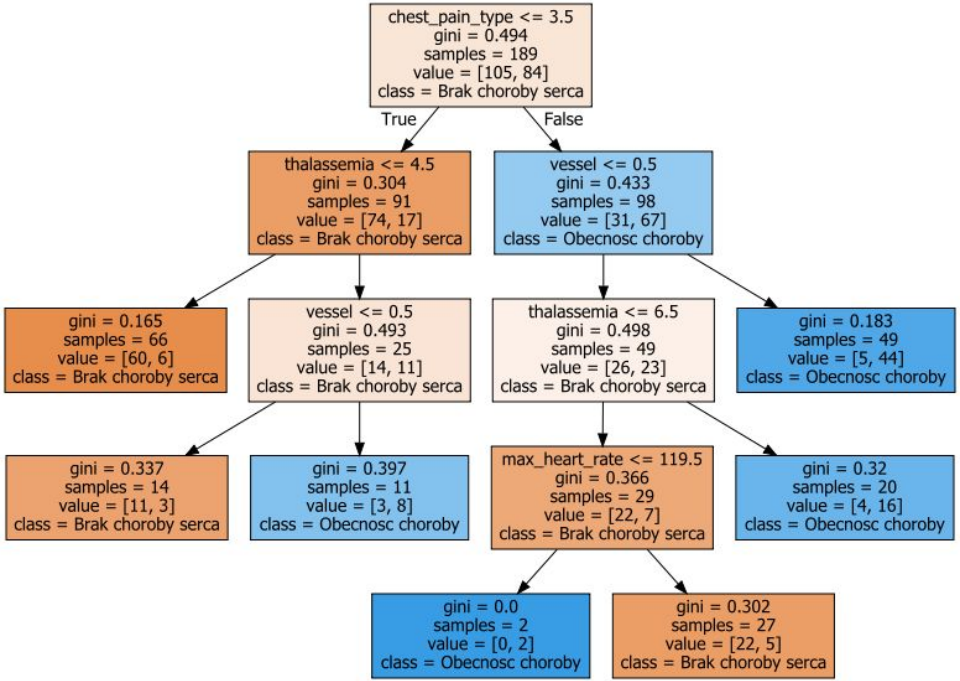
zbiór testowy	zbiór uczący
Trafność: 0.889 Czułość: 0.806 Specyficzność: 0.956	Trafność: 0.862 Czułość: 0.833 Specyficzność: 0.886

Widzimy, że dla zbioru testowego otrzymaliśmy lepsze wyniki dla trafności i specyficzności. Z tego wynika, że udało nam się polepszyć nasz model. Krzywe ROC wyglądają tak samo, jak w modelu 1.1.

Model 1.2 Drzewo CART



Na wykresie słupkowym widzimy ważność zmiennych dla Modelu 1.2



W metodzie klasyfikacji drzewa CART lepszy okazał się model 1.2.

MODEL 2 – metoda MPL

Założenia modelu:

- Dane numeryczne muszą być poddane skorygowanej normalizacji $x_i^S = 2 \cdot \frac{x_i - min}{max - min} - 1$
- Dane kategoryczne należy zamienić na wartości liczbowe, jeśli są łańcuchami

By móc porównać dwa modele, wykorzystujemy te same predyktory. Są to "chest_pain_type", "angina", "slope", "max_heart_rate", "vessel", "thalassemia".

Dokonujemy podziału na zbiór uczący (30% wszystkich obserwacji) i testowy (70% wszystkich obserwacji). Ziarno generatora liczb losowych jest równe 308272. Następnie dzielimy predyktory na zmienne numeryczne i kategoryczne. Dane numeryczne poddajemy skorygowanej normalizacji. W naszym wyjściowym pliku nie trzeba zamieniać danych kategorycznych, ponieważ są liczbami.

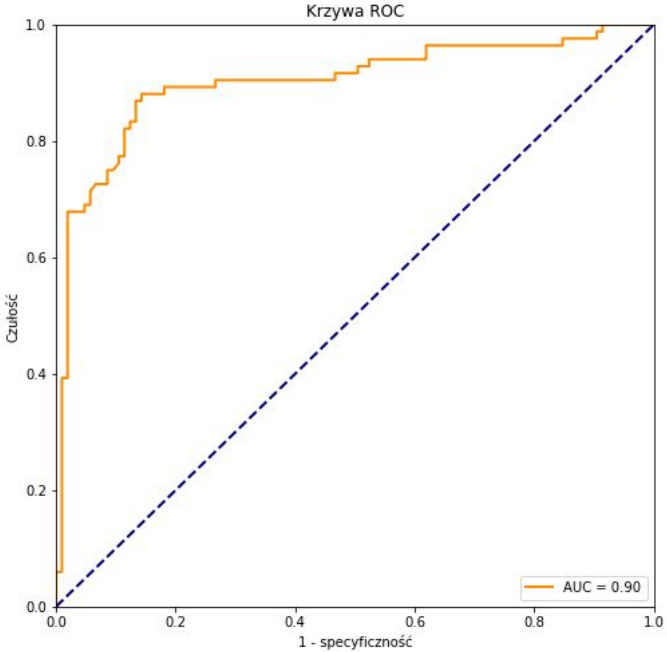
MODEL 2.1

Zbiór testowy zawiera 189 obserwacji, zbiór uczący zaś 81 obserwacje.
Parametr kary wynosił 0,0001 oraz warstwa ukryta (1,).
W tym podejściu współczynniki prezentują się następująco:

zbiór testowy	zbiór uczący
Trafność: 0.864 Czułość: 0.722 Specyficzność: 0.978	Trafność: 0.847 Czułość: 0.798 Specyficzność: 0.886

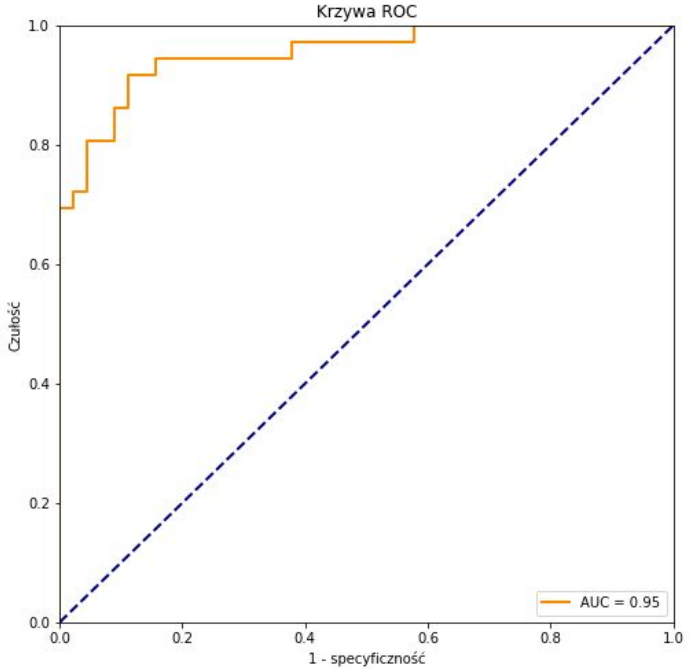
Zbiór testowy wypada gorzej (o najwięcej 7%) niż uczący, co wskazuje na lekkie przeuczenie się modelu. Spróbujmy polepszyć nasz model.

Krzywa ROC dla modelu 2.1



← *zbiór uczący*

zbiór testowy →



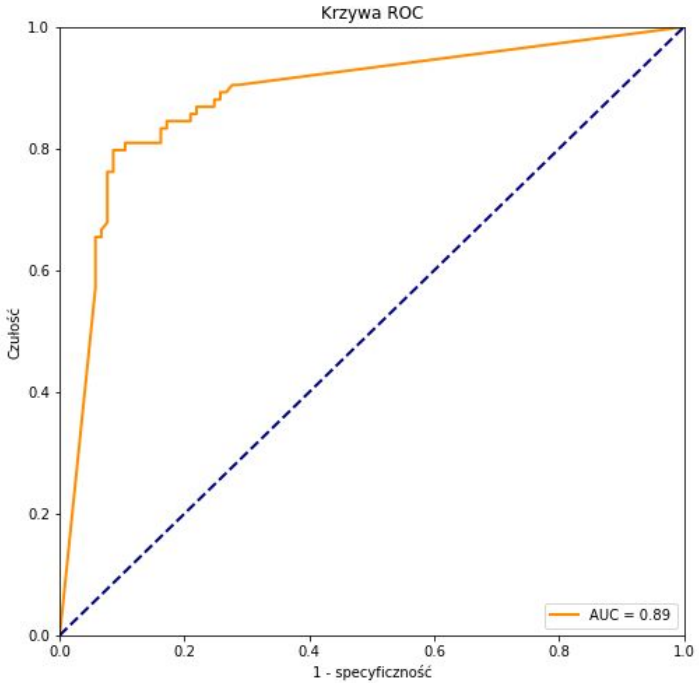
Model 2.2

Zbiór testowy zawiera 189 obserwacji, zbiór uczący zaś 81 obserwacje.
W tym podejściu został zmieniony parametr kary na (1,2)
W tym podejściu współczynniki prezentują się następująco:

zbiór testowy	zbiór uczący
Trafność: 0.889 Czułość: 0.75 Specyficzność: 1.0	Trafność: 0.862 Czułość: 0.798 Specyficzność: 0.914

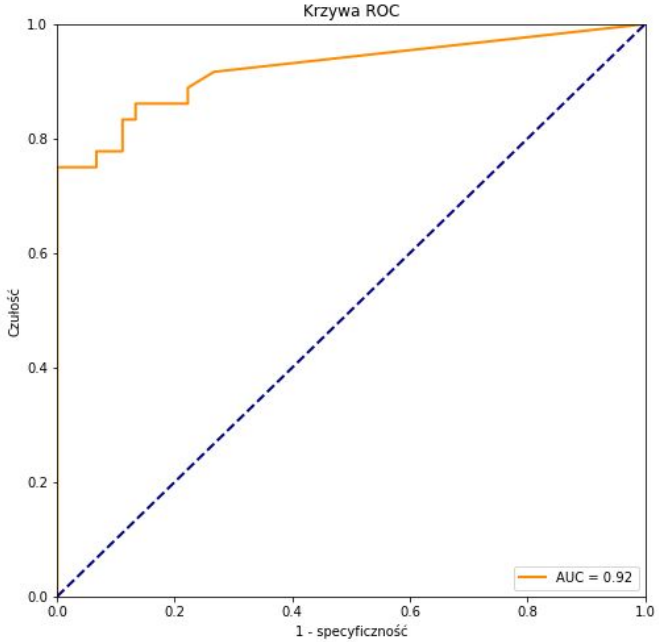
Widzimy, że dla zbioru testowego polepszyły się wszystkie współczynniki. Z tego wynika, że udało nam się polepszyć nasz model.

Krzywa ROC dla modelu 2.2



← *zbiór uczący*

zbiór testowy →



W metodzie klasyfikacji za pomocą sieci neuronowych lepszy okazał się model 2.2.

PODSUMOWANIE:

1) Zarówno model 1.2, jak i 2.2 są dobre.
2) Ciężko stwierdzić jednoznacznie, który z tych modeli jest lepszy.
3) Pojawiły się problemy z ulepszaniem modeli, ale dzięki konsultacjom zostały rozwiązane.
4) Inne problemy były, ale jakoś poszło...

Raport wykonały:

Aleksandra Grzegórska, Anna Cabaj