

## Budowa modeli predykcji wystąpienia ataku serca

Dane, na których zostały zbudowane dwa modele predykcji, składają się z informacji o 300 osobach płci męskiej i żeńskiej. Dane osób, na których będziemy bazować to między innymi:

- Wiek
- Płeć
- Ból w klatce piersiowej
- Ciśnienie krwi
- Cholesterol
- Maksymalne tętno

Dane zostały podzielone na dwa zbiory – uczący i testowy – zawierające po 150 osób oraz zostały sprawdzone pod kątem normalizacji.

Modele, które zostały zbudowane to:

- Model regresji logistycznej

Model regresji logistycznej jest szczególnym przypadkiem uogólnionego modelu liniowego. Znajduje zastosowanie, gdy zmienna zależna jest dychotomiczna, to znaczy przyjmuje tylko dwie wartości takie jak na przykład sukces lub porażka, wystąpienie lub brak pewnej jednostki chorobowej. W zapisie matematycznym wartości te reprezentowane są jako 1 i 0.

- Drzewo decyzyjne

Korzystają one z graficznego przedstawienia decyzji i ich możliwych konsekwencji. Drzewa decyzyjne są tak skonstruowane aby pomóc w podejmowaniu decyzji.

Pierwszy model to model regresji logistycznej oparty na wszystkich zmiennych, do których mamy dostęp. Szczegółowe informacje modelu wskazują, które zmienne są mniej, a które bardziej istotne.

COEF1 RESULTS:

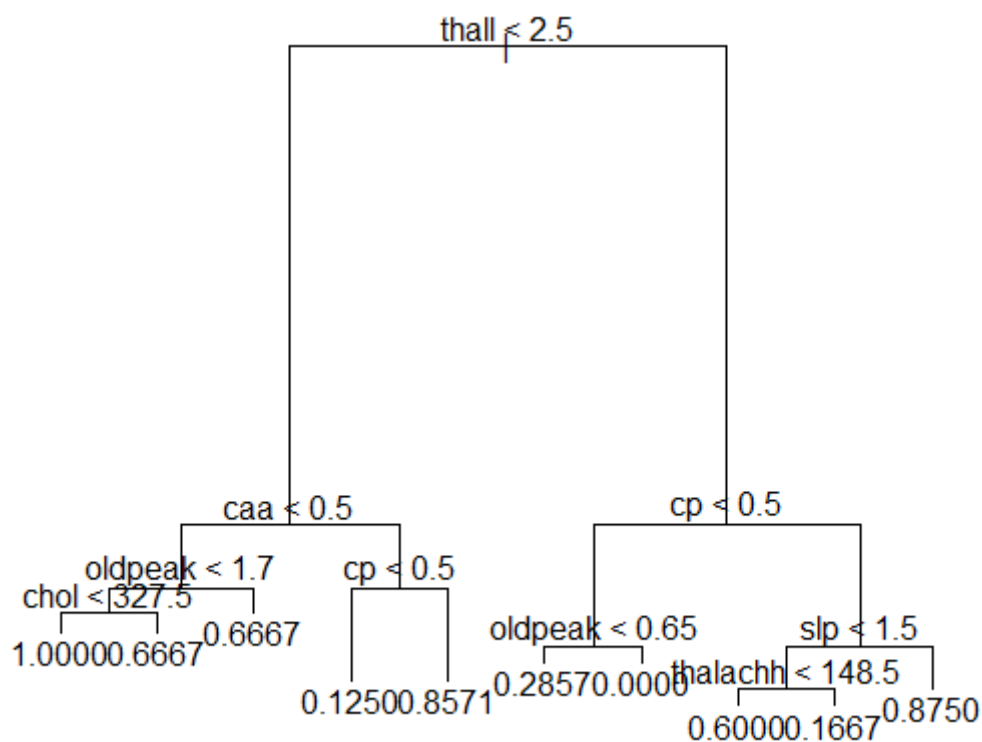
	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	7.609e+00	7.199e+00	1.057	0.29054	
dane\$age	-6.524e-02	7.322e-02	-0.891	0.37295	
dane\$sex	-3.448e+00	1.308e+00	-2.635	0.00840	**
dane\$cp	1.721e+00	6.067e-01	2.837	0.00456	**
dane\$trtbps	-1.342e-02	3.628e-02	-0.370	0.71139	
dane\$chol	-2.511e-02	1.116e-02	-2.250	0.02447	*
dane\$fbs	-4.205e-01	1.456e+00	-0.289	0.77274	
dane\$restecg	2.150e+00	1.009e+00	2.131	0.03312	*
dane\$thalachh	5.184e-02	3.322e-02	1.560	0.11866	
dane\$exng	4.817e-01	1.041e+00	0.463	0.64359	
dane\$oldpeak0.2	2.191e+00	3.417e+00	0.641	0.52141	
dane\$oldpeak0.4	4.720e+00	1.994e+00	2.367	0.01794	*
dane\$oldpeak0.5	2.041e+01	4.371e+03	0.005	0.99627	
dane\$oldpeak0.6	-1.091e+00	1.297e+00	-0.841	0.40044	
dane\$oldpeak0.8	3.425e+00	3.386e+00	1.012	0.31177	
dane\$oldpeak1	-8.775e-01	2.113e+00	-0.415	0.67794	
dane\$oldpeak1.2	-4.721e-01	1.663e+00	-0.284	0.77642	
dane\$oldpeak1.3	1.828e+01	1.075e+04	0.002	0.99864	
dane\$oldpeak1.4	-1.984e+00	2.005e+00	-0.990	0.32236	
dane\$oldpeak1.5	3.719e+00	1.153e+01	0.323	0.74702	
dane\$oldpeak1.6	1.402e+00	2.679e+00	0.524	0.60059	
dane\$oldpeak1.8	1.820e+00	1.930e+00	0.943	0.34582	
dane\$oldpeak2	-1.832e+01	7.174e+03	-0.003	0.99796	
dane\$oldpeak2.2	-1.200e+01	7.591e+03	-0.002	0.99874	
dane\$oldpeak2.3	1.929e+01	7.023e+03	0.003	0.99781	
dane\$oldpeak2.4	-1.318e+01	1.075e+04	-0.001	0.99902	
dane\$oldpeak2.5	-1.781e+01	7.076e+03	-0.003	0.99799	
dane\$oldpeak2.6	-4.916e-01	5.702e+00	-0.086	0.93130	
dane\$oldpeak2.8	-1.423e+01	4.720e+03	-0.003	0.99759	
dane\$oldpeak3	-9.500e-01	2.056e+00	-0.462	0.64401	
dane\$oldpeak3.1	-1.494e+01	1.075e+04	-0.001	0.99889	
dane\$oldpeak3.2	-1.916e+01	1.075e+04	-0.002	0.99858	
dane\$oldpeak3.4	-1.070e+01	1.075e+04	-0.001	0.99921	
dane\$oldpeak3.5	1.485e+01	1.075e+04	0.001	0.99890	
dane\$oldpeak3.6	-2.141e+01	4.700e+03	-0.005	0.99636	
dane\$oldpeak4	-1.171e+01	7.253e+03	-0.002	0.99871	
dane\$oldpeak5.6	-1.475e+01	1.075e+04	-0.001	0.99891	
dane\$oldpeak6.2	-1.519e+01	1.075e+04	-0.001	0.99887	
dane\$slp	1.241e+00	8.331e-01	1.490	0.13629	
dane\$caa	-9.147e-01	5.157e-01	-1.774	0.07612	.
dane\$thall	-2.150e+00	8.360e-01	-2.572	0.01011	*

Na tej podstawie został zbudowany drugi model regresji logistycznej, składający się tylko z najbardziej istotnych zmiennych. Model ten został poddany próbie predykcji ataku serca na zbiorze testowym. Skuteczność wyniosła 76%. Szczegółowe informacje tego modelu wskazały dwie, mniej istotne zmienne.

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	7.264047	1.986587	3.657	0.000256	***
dane\$sex	-1.240954	0.599214	-2.071	0.038362	*
dane\$cp	0.950666	0.246399	3.858	0.000114	***
dane\$chol	-0.008897	0.005360	-1.660	0.096930	.
dane\$restecg	1.407323	0.497355	2.830	0.004660	**
dane\$thall	-2.183768	0.450208	-4.851	1.23e-06	***

Na tej podstawie został zbudowany kolejny model, aby możliwie jak najbardziej poprawić skuteczność wcześniejszego modelu. Wynik praktycznie się nie zmienił – 75%.

Do porównania wyników został stworzony inny model – drzewo decyzyjne. Inicjalizacja drzewa została oparta na wszystkich zmiennych. Ostatecznie algorytm wybrał tylko 7 z 13. Zmienne, które zostały wzięte pod uwagę różnią się od tych, które zostały użyte w modelu regresji logistycznej. Wspólne zmienne obu modeli to cholesterol oraz ból w klatce piersiowej.



Ten model również został poddany testowi predykcji ataku serca u pewnych osób. Jego skuteczność również wynosiła 76%.

Oba modele – model regresji logistycznej i drzewo decyzyjne – sprawdziły się tak samo dobrze, osiągając taką samą skuteczność. Chcąc przewidzieć atak serca pacjenta, mając taki sam zestaw informacji, który został wykorzystany przy budowie obu modeli, mamy 76% szans na to, że u pacjenta faktycznie wystąpi atak serca.