LABORATORIUM 3

Mikołaj Sikora

Dla pełnego zestawu cech przy użyciu widgetu rank widać, że najważniejsze cechy to kolejno: *sex*, *pclass* oraz *fare*.

	#	Info. gain	Gain ratio	Gini	χ^2	ReliefF	FCBF
C sex	2	0.215	0.226	0.140	189.966	0.096	0.288
C pclass	3	0.075	0.049	0.050	62.733	0.042	0.064
N fare		0.066	0.033	0.044	79.188	0.011	0.000
C parch	8	0.034	0.028	NA	22.776	0.032	NA
C sibsp	7	0.024	0.018	0.014	0.206	0.038	0.000
N age		0.002	0.001	0.002	1.319	0.030	0.000

W regresji logistycznej trzeci najbardziej istotny atrybut to *age* - *fare* jest poniżej *sibsp*.

	name	1	X1
3	sex	1.22055	1.22055
2	pclass	-0.908223	0.908223
4	age	-0.536413	0.536413
1	intercept	-0.481484	0.481484
5	sibsp	-0.307511	0.307511
7	fare	0.0883969	0.0883969
6	parch	0.0383631	0.0383631

W regresji liniowej wyniki są podobne.

	name	coef	X1
1	intercept	0.408612	0.408612
3	sex	0.23622	0.23622
2	pclass	-0.143734	0.143734
4	age	-0.0797247	0.0797247
5	sibsp	-0.0415776	0.0415776
7	fare	0.0130529	0.0130529
6	parch	0.00420706	0.00420706

Po usunięciu płci ze zbioru rozpatrywanych cech, po użyciu widgetu rank dostajemy następujący ranking:

	#	Info. gain	Gain ratio	Gini	χ^2	ReliefF	FCBF
N pclass		0.075	0.049	0.050	62.733	0.000	0.064
N fare		0.066	0.033	0.044	79.188	0.001	0.000
N parch		0.030	0.025	0.020	30.688	0.015	0.028
N sibsp		0.019	0.015	0.013	0.417	0.004	0.000
N age		0.002	0.001	0.002	1.319	0.012	0.000

Najbardziej istotne okazują się być *pclass* oraz *fare*.

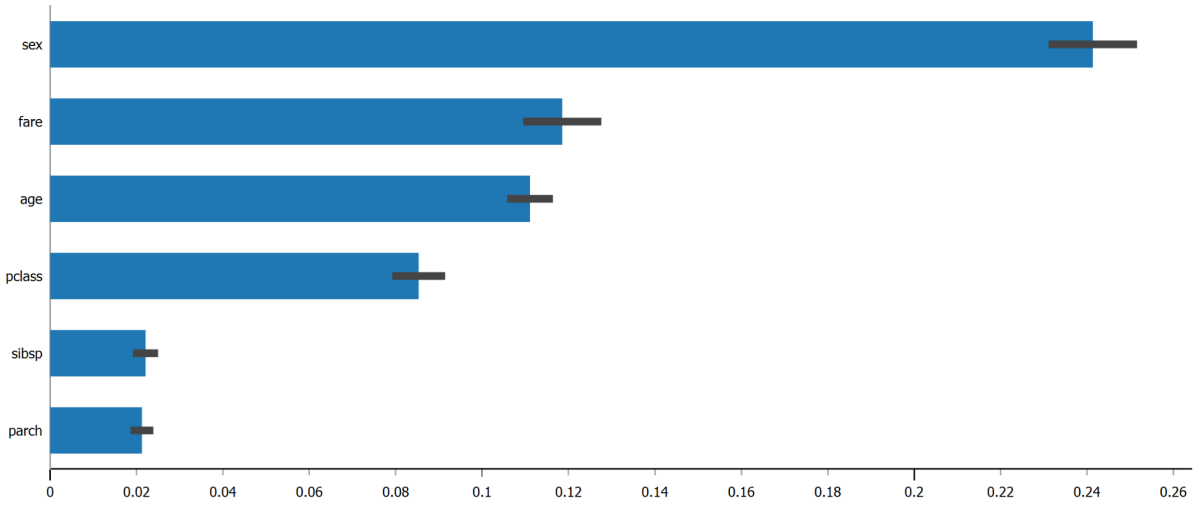
W regresji logistycznej najbardziej istotne jest *pclass* oraz *age*.

	name	1	X1
2	pclass	-0.833164	0.833164
3	age	-0.542671	0.542671
1	intercept	-0.40729	0.40729
5	parch	0.248599	0.248599
4	sibsp	-0.232677	0.232677
6	fare	0.178289	0.178289

W regresji liniowej podobnie:

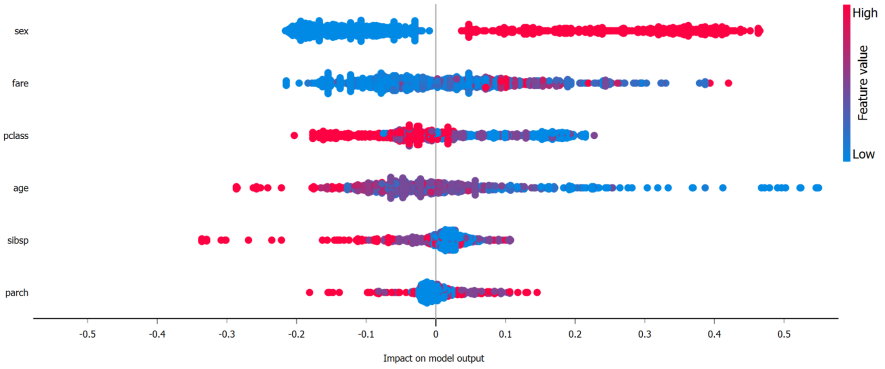
	name	coef	X1
1	intercept	0.408612	0.408612
2	pclass	-0.177653	0.177653
3	age	-0.105713	0.105713
5	parch	0.0494443	0.0494443
4	sibsp	-0.0436635	0.0436635
6	fare	0.0334082	0.0334082

Poniżej przedstawiono ranking cech dla pełnego zbioru (z płcią) dla widgetu *Feature Importance*:

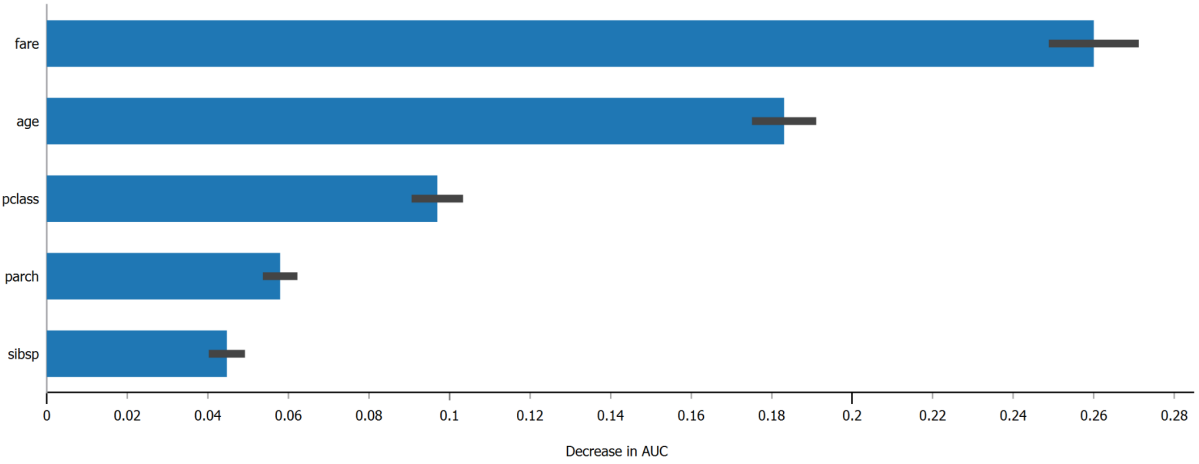


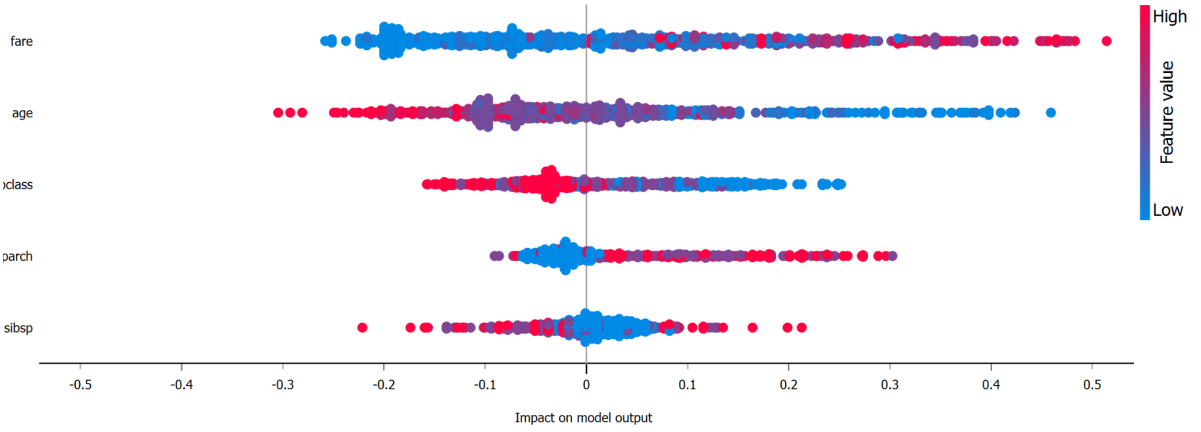
Podobnie jak dla klasycznych metod płeć okazała się najbardziej skorelowana z wartością *survived*. Kolejne istotne cechy to *fare, age* oraz *pclass*. Pozostałe dwa nie mają wielkiego wpływu.

Poniżej znajduje się wykres z *ExplainModel* (Random Forest). Można odczytać następujące zależności, że na to czy ktoś przeżył, wpływ ma: czy był mężczyzną, czy miał wysokie *fare*, czy miał niskie *pclass* i jak młody był (im młodszy - tym większe szanse).



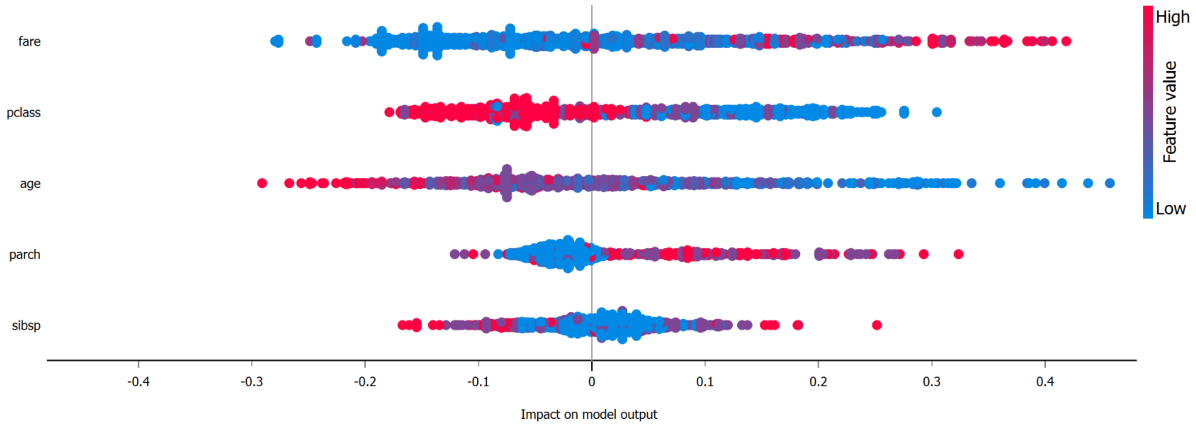
Po usunięciu płci najbardziej istotne okazało się być kolejno *fare, age* oraz *pclass*.



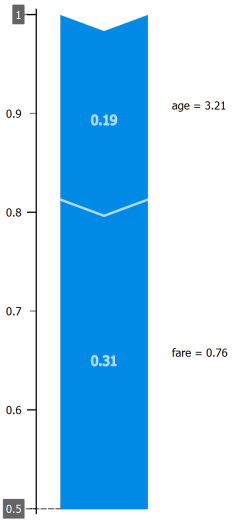


Przykładowy osobnik: miał 76 lat, sibsp = 1, parch = 0, fare = 78.85, pclass = 1.

ExplainModel dla RandomForest (65% accuracy) (survived == 1)

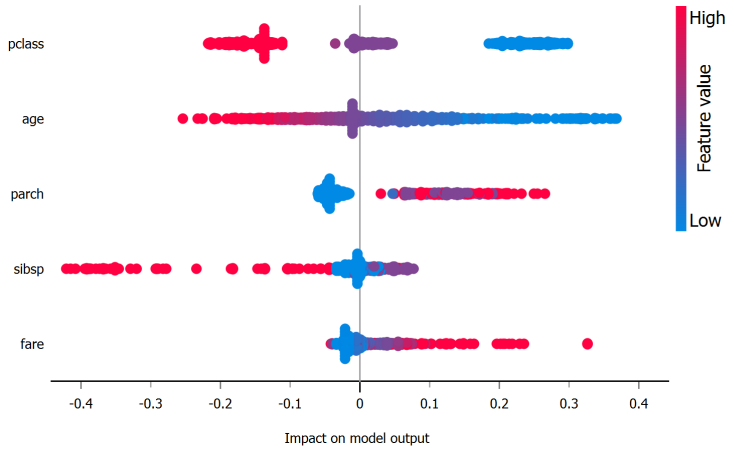


ExplainPrediction zwróciło 0.5.

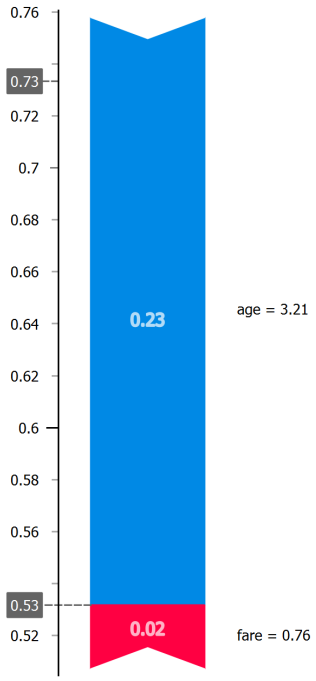


Wysokie BaseValue (wynoszące 1), ale spadek do 0.5 spowodował duży wiek oraz niezbyt wysokie *fare*.

ExplainModel dla Neural Network (80% accuracy) (survived == 1)



Widać, że ten model bardzo precyzyjnie odseparował od siebie kategoryczne wartości *pclass*.
Sklasyfikowano osobnika jako survived z prawdopodobieństwem 0.53.



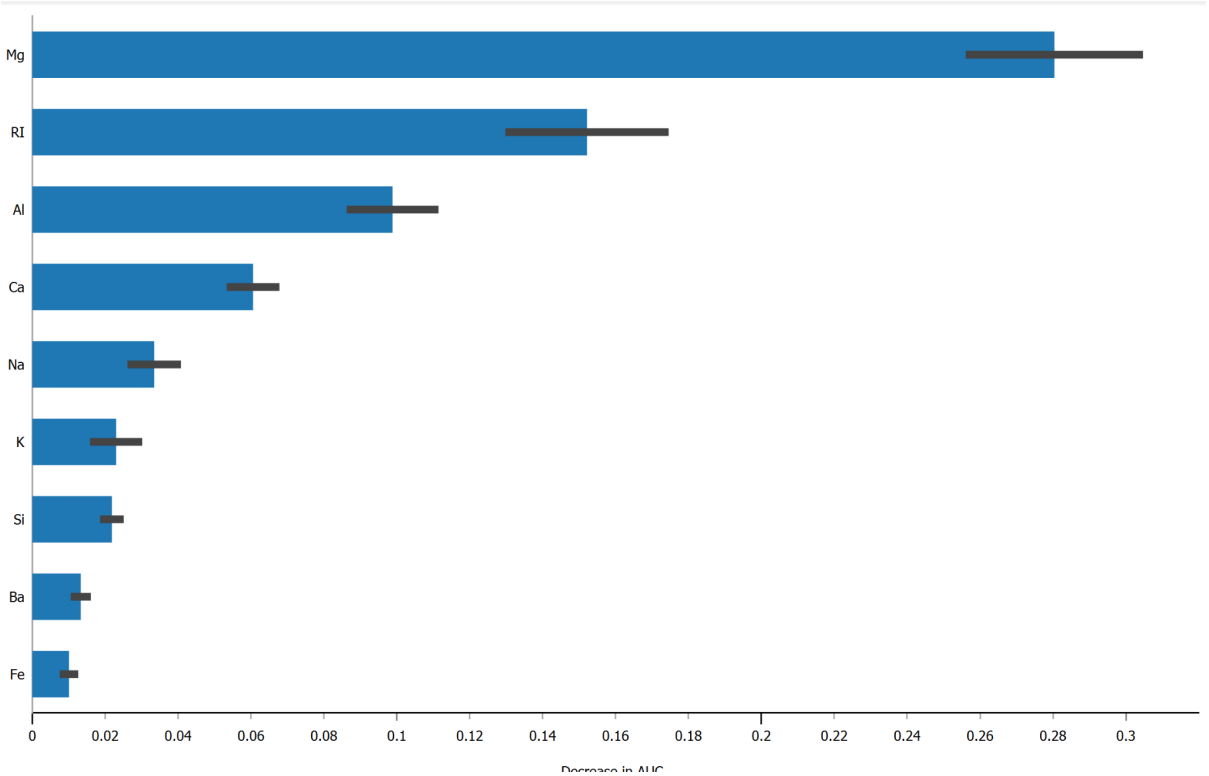
W tym przypadku dużo lepiej dostosowano *fare*, wpłynęło ono dodatnio na klasyfikację.
Również wiek okazał się mieć większe znaczenie niż w przypadku Random Forest.

Ćwiczenie powtórzono dla zbioru **Glass**.

Najważniejsze parametry (widget *Rank*):

	#	Info. gain	Gain ratio	Gini	χ^2	ReliefF	FCBF
Mg		0.537	0.269	0.101	85.202	0.261	0.346
Al		0.471	0.236	0.121	68.899	0.069	0.291
Ba		0.384	0.402	0.107	268.638	0.063	0.325
Na		0.368	0.184	0.080	59.018	0.036	0.214
K		0.330	0.165	0.073	38.548	0.024	0.000
Ca		0.325	0.163	0.074	20.663	0.070	0.185
Si		0.196	0.098	0.034	18.741	0.014	0.000
RI		0.145	0.073	0.031	17.682	0.049	0.000
Fe		0.123	0.086	0.022	24.077	-0.013	0.073

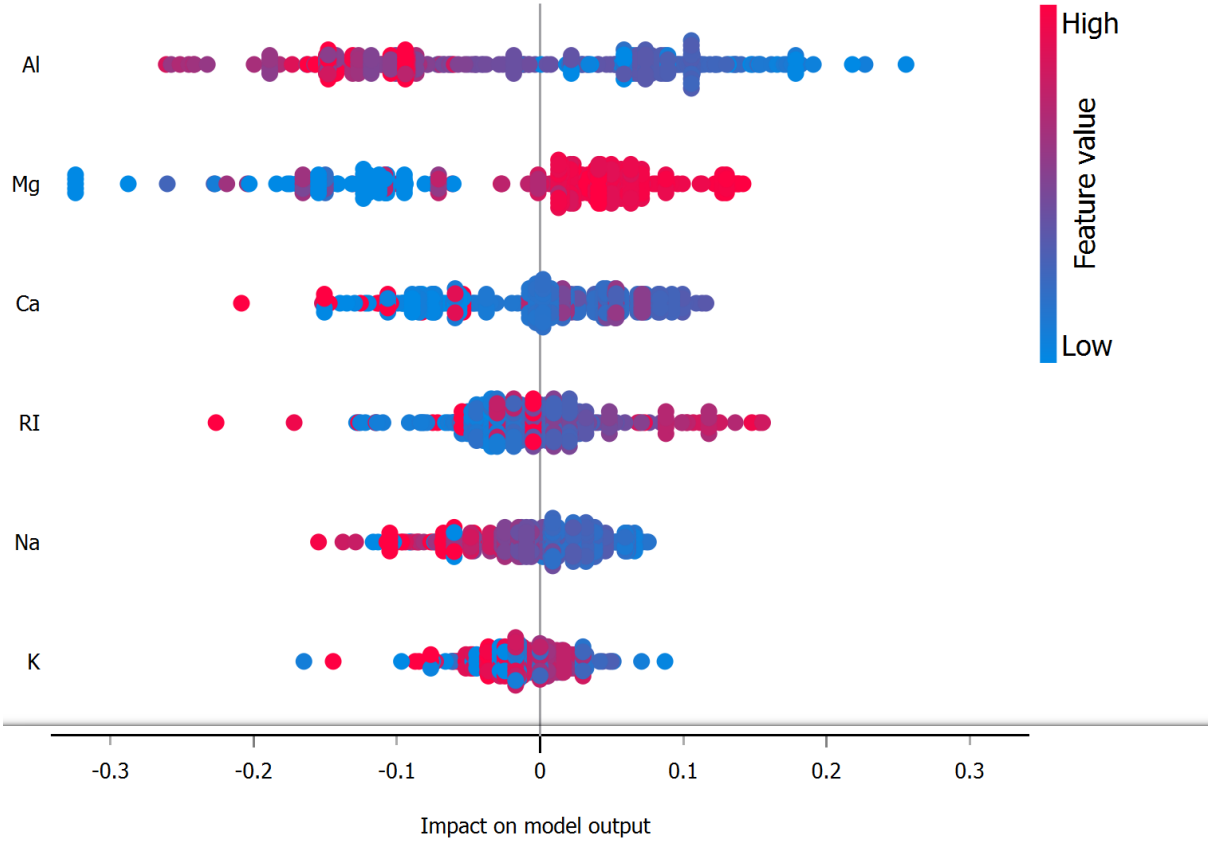
Wykres z *Feature Importance*:



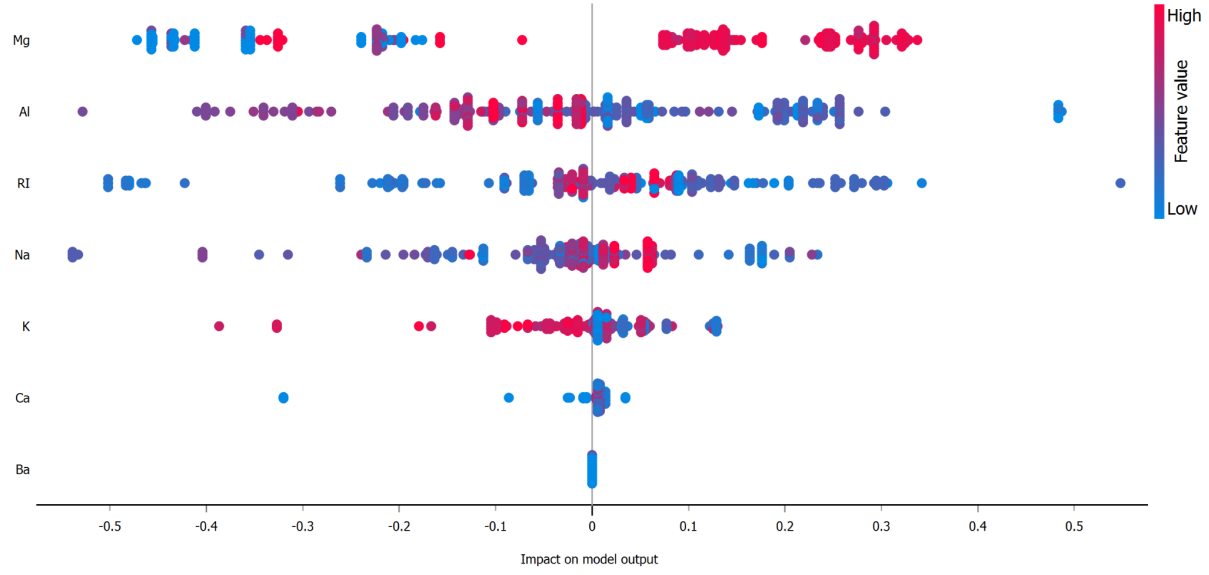
Na podstawie powyższych dwóch postanowiono wyłączyć krzem i żelazo.

Usunięcie tych dwóch atrybutów poprawiło jakość sieci neuronowej o 0.01%.

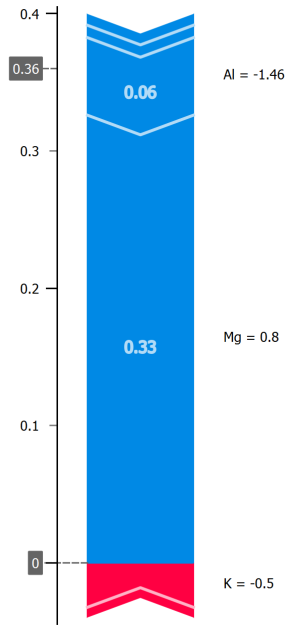
Explain Model dla typu szkła = 1 dla sieci neuronowej (skuteczność 89%) przedstawiono poniżej. Największy wpływ ma jak najmniejsza zawartość *Al*, do tego jak największa zawartość *Mg*. Kolejnymi ważnymi czynnikami jest kolejno *Ca, RI, Na* oraz *K*.



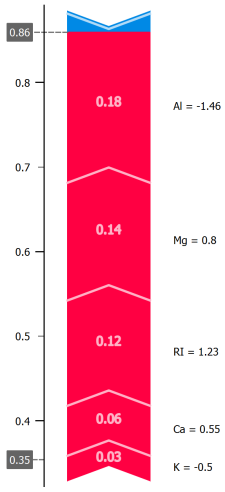
Natomiast dla Random Forest (skuteczność 70%) widać, że rozkład jest dużo bardziej chaotyczny. Tutaj *Mg* okazał się dużo bardziej istotny. Kolejny jest jak najmniejsze *Al* oraz jak najmniejsze *RI.*



Test dla przykładu o atrybutach:
y=1, RI=1.52210, Na=13.73, Mg=3.84, Al=0.72, K=0.17, Ca=9.74, Ba=0.00



Dla Random Forest prawdopodobieństwo, że y == 0 oceniono na 0!
Widać, że największy wpływ na ten wynik miały niskie wartości Al, Mg (które są najbardziej istotne w tym modelu).



W przypadku sieci neuronowej klasyfikacja przebiegła prawidłowo - otrzymano prawdopodobieństwo 86% na przynależność do klasy nr 1.
Największy wpływ na ten wynik miały Al, Mg oraz RI. Zatem właściwie bardzo podobne parametry, do tych co miały wpływ na błędny wynik klasyfikatora Random Forest. Model utworzony z pomocą sieci neuronowej zatem jest dużo lepiej dopasowany do danego zbioru, co zresztą było już widać po wykresach wyprodukowanych przez *ExplainModel,* gdzie wykres dla sieci neuronowej był bardziej uporządkowany i mieścił się w mniejszym przedziale (-0.3;0.3) niż ten dla Random Forest (-0.5;0.5).