Machine Learning 2

Lab 3

Oscar Teeninga

1. Zadanie wstępne (1)

Najpierw usunąłem ‘sex’ z cech i sprawdziłem wpływ na pozostałe cechy

	name	1	X1
7	sex	1.21955	1.21955
2	pclass	-0.908431	0.908431
3	age	-0.542026	0.542026
1	intercept	-0.487976	0.487976
4	sibsp	-0.317782	0.317782
6	fare	0.095779	0.095779
5	parch	0.0455476	0.0455476

Regresja logistyczna z sex

	name	1	X1
2	pclass	-0.833948	0.833948
3	age	-0.5479	0.5479
1	intercept	-0.411969	0.411969
5	parch	0.253208	0.253208
4	sibsp	-0.240451	0.240451
6	fare	0.184664	0.184664

Regresja logiczna bez sex

	name	coef	X1
1	intercept	0.408612	0.408612
7	sex	0.237728	0.237728
2	pclass	-0.147848	0.147848
3	age	-0.085454	0.085454
4	sibsp	-0.0475429	0.0475429
6	fare	0.0144862	0.0144862
5	parch	0.00801002	0.00801002

Regresja liniowa z sex

	name	coef	X1
1	intercept	0.408612	0.408612
2	pclass	-0.181983	0.181983
3	age	-0.11161	0.11161
5	parch	0.0535358	0.0535358
4	sibsp	-0.0496436	0.0496436
6	fare	0.0349719	0.0349719

Regresja liniowa bez sex

W przypadku regresji logicznej widać, że zwiększyła się istotność cechy **parch**.

W przypadku regresji liniowej występuje dokładnie to samo, cecha **parch** staje się istotniejsza niż było to wcześniej o rząd wielkości.

Ranking cech w Glass

	#	Info.gain	Gain ratio	Gini	χ^2	ReliefF	FCBF
N	...	0.537	0.269	0.101	85.202	0.200	0.346
N	Al	0.471	0.236	0.121	68.899	0.063	0.291
N	Ba	0.384	0.402	0.107	268.638	0.040	0.325
N	...	0.368	0.184	0.080	59.018	0.034	0.214
N	K	0.330	0.165	0.073	38.548	0.018	0.000
N	Ca	0.325	0.163	0.074	20.663	0.053	0.185
N	Si	0.196	0.098	0.034	18.741	0.021	0.000
N	RI	0.145	0.073	0.031	17.682	0.038	0.000
N	Fe	0.123	0.086	0.022	24.077	0.011	0.073

	name	coef	X1
1	intercept	1.54206	1.54206
10	Mg	-0.595791	0.595791
3	Na	0.588997	0.588997
4	Al	0.547345	0.547345
5	Si	0.33189	0.33189
8	Ba	0.27095	0.27095
2	RI	0.17301	0.17301
6	K	0.0670014	0.0670014
9	Fe	-0.0584218	0.0584218
7	Ca	0.012848	0.012848

	name	coef	X1
1	intercept	1.54206	1.54206
3	Na	0.90936	0.90936
4	Al	0.746692	0.746692
5	Si	0.63907	0.63907
7	Ca	0.606669	0.606669
8	Ba	0.474662	0.474662
6	K	0.33406	0.33406
2	RI	0.149168	0.149168
9	Fe	-0.0436026	0.0436026

	name	1	2	3	5	6	7	X1
10	Mg	1.65709	-0.159639	0.852644	-1.01768	-0.226414	-1.106	1.65709
4	Al	-1.59702	-0.166346	-1.31744	1.58643	0.404692	1.08968	1.59702
1	intercept	1.05666	2.24505	-0.021796	-0.952759	-1.7079	-0.619259	1.05666
3	Na	-0.575158	-0.718653	-0.274527	-0.833202	1.40474	0.996795	0.575158
2	RI	0.434145	0.247769	-1.31782	-0.266968	-0.174406	1.07728	0.434145
8	Ba	0.317639	0.0213435	-0.155131	-0.139001	-0.935247	0.890397	0.317639
5	Si	0.280381	-0.408731	-1.07559	-0.186769	0.378559	1.01215	0.280381
9	Fe	0.247795	0.484646	0.240523	0.203282	-0.825926	-0.350319	0.247795
7	Ca	-0.200563	-0.373655	0.464292	0.579175	0.271298	-0.740547	0.200563
6	K	0.128756	0.282863	-0.534485	0.799675	-1.44607	0.769264	0.128756

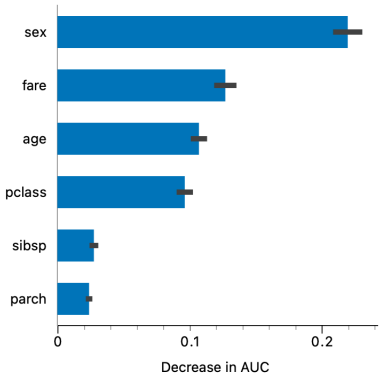
	name	1	2	3	5	6	7	X1
4	Al	-1.91913	-0.0613028	-1.45249	1.81955	0.346404	1.26696	1.91913
7	Ca	-1.37383	-0.199198	-0.0198402	1.24538	0.412936	-0.0654433	1.37383
1	intercept	1.24131	2.06943	0.0477239	-0.967189	-1.84713	-0.544152	1.24131
3	Na	-1.05423	-0.714521	-0.450178	-0.55884	1.4198	1.35796	1.05423
2	RI	0.410434	0.160717	-1.3657	-0.176823	-0.184173	1.15554	0.410434
5	Si	-0.338338	-0.39376	-1.31305	0.139603	0.469939	1.43561	0.338338
9	Fe	0.239552	0.474475	0.234806	0.219746	-0.818399	-0.350179	0.239552
6	K	-0.195927	0.281617	-0.554302	0.973458	-1.45133	0.946483	0.195927
8	Ba	0.0120869	0.0282359	-0.231995	0.0205723	-0.969758	1.14086	0.0120869

Wyrzuciłem z cech **Magnez (Mg).**

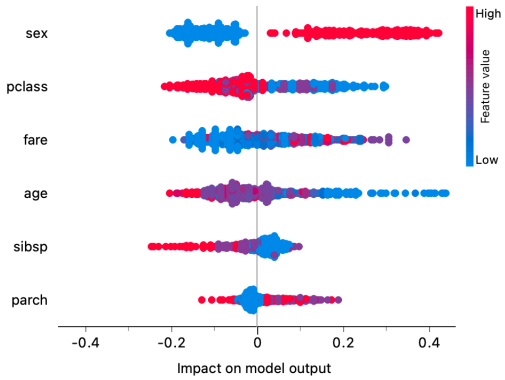
Znacząco w przypadku regresji logicznej wzrasta ranking wapnia (**Ca**). Inne cechy również zmieniają swoją istotność, ale w mniejszym stopniu. Jednym słowem - widać różnicę.

Analogicznie wygląda to dla regresji liniowej. Wapń (**Ca**) awansuje w rankingu. Inne cechy również zmieniają swoją pozycję jak np. **Ba**, który spada w rankingu zna ostatnie miejsce.

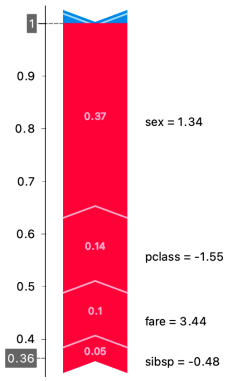
2. Zadanie wstępne (2)



Feature importance



Model



Prediction

Model	Train time [s]	Test time [s]	AUC	CA	F1	Precision	Recall	LogLoss	Specificity
Tree	0.621	0.001	0.767	0.760	0.756	0.758	0.760	4.329	0.721
SVM	0.529	0.029	0.823	0.807	0.803	0.808	0.807	0.468	0.767
Random Forest	0.260	0.054	0.835	0.782	0.780	0.780	0.782	1.239	0.752
CN2 rule inducer	72.076	0.056	0.762	0.735	0.733	0.733	0.735	0.589	0.704

Porównanie metod klasyfikacji

Płeć (**sex**) jest w tym przypadku najbardziej istotną cechą, zgodnie w poprzednim zadaniem. Natomiast ciekawe jest, że znacznie wyżej jest **fare**.
Explain model bada, które cechy mają największy wpływ w zależności od ich wielkości dla każdego przykładu. Bardzo fajnie widać to po wieku (**age**) (zbiór mniej dyskretny niż np. sex). Widzimy, że dla dużego bardzo dużego wieku i małego wieku wpływ na model jest największy.

Następnie wybrałem dwa przykłady. Są to bardzo małe dzieci, podróżujące pierwszą klasą, będące chłopcami. Jest to ciekawe ze względu na to, że dzieci mają większą szansę na przeżycie, ale wiek jednego z dzieci to było 2 lata, więc może to dać ciekawe wnioski w jaki sposób model podchodzi do przykładów ekstremalnych (cechy powinny wręcz gwarantować przeżycie, a dziecko tak małe nie przeżyło).

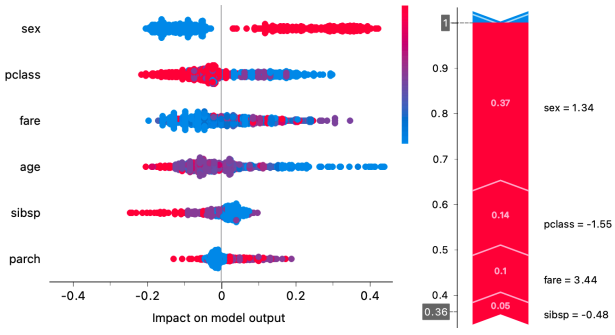
	survived	pclass	sex	age	sibsp	parch	fare
1	0	1	1	2	1	2	151.55
2	1	1	1	14	1	2	120

Wybrane dwa przykłady

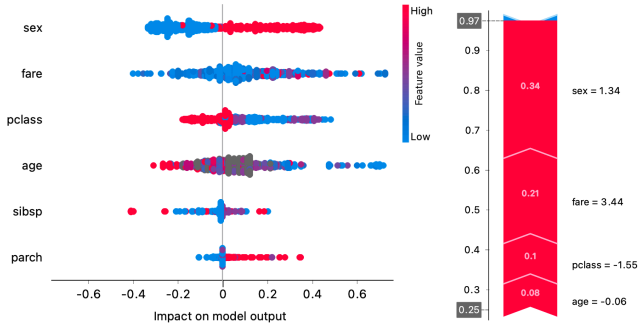
Co ciekawe, CN2 rule inducer zadziałał i zorientował się, że jedno z dzieci było na tyle młode, żeby nie przeżyć.

Model	Train time [s]	Test time [s]	AUC	CA	F1	Precision	Recall	LogLoss	Specificity
Tree	0.101	0.001	0.500	0.500	0.333	0.250	0.500	1.849	0.500
SVM	0.070	0.002	1.000	0.500	0.333	0.250	0.500	0.736	0.500
Random Forest	0.023	0.007	1.000	0.500	0.333	0.250	0.500	0.458	0.500
CN2 rule inducer	13.674	0.005	1.000	1.000	1.000	1.000	1.000	0.166	1.000

Porównanie wyników klasyfikacji dwóch przykładów



Random forest

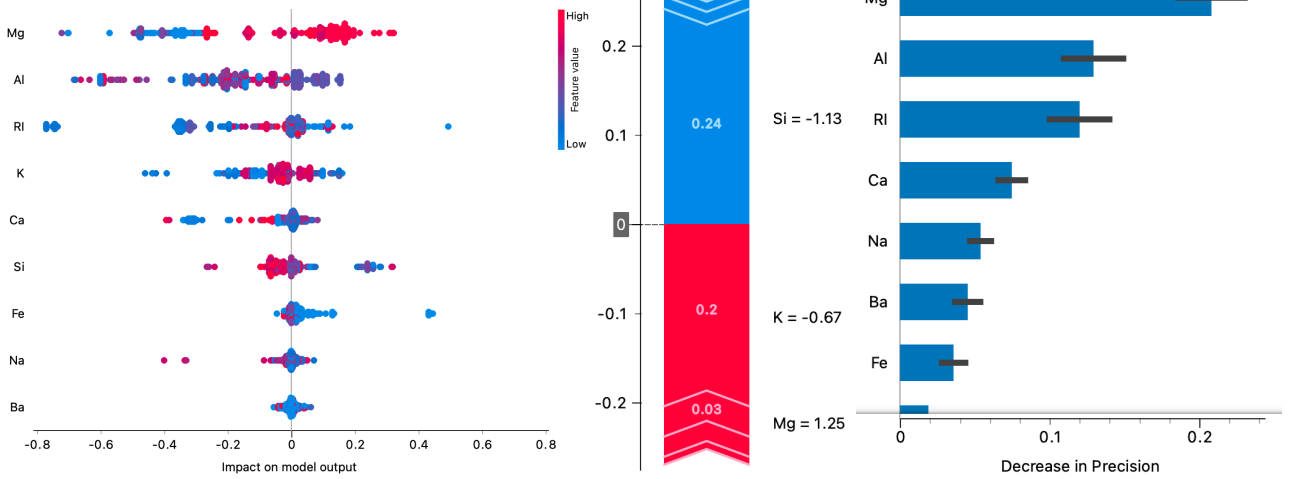


Tree

3. Zadanie finalne (Glass)

Jak widać, CN2 i SVM są wyraźnie lepsze (Precision) od pozostałych.

	y	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
1	2	1.53125	10.73	0.00	2.10	69.81	0.58	13.30	3.15	0.28
2	5	1.51514	14.01	2.68	3.50	69.89	1.68	5.87	2.20	0.00
3	7	1.51831	14.39	0.00	1.82	72.86	1.41	6.47	2.88	0.00



Model	Train time [s]	Test time [s]	AUC	CA	F1	Precision	Recall	LogLoss	Specificity
Tree	0.030	0.001	1.000	0.667	0.556	0.500	0.667		0.833
SVM	0.016	0.003	1.000	1.000	1.000	1.000	1.000		1.000
Random Forest	0.017	0.005	1.000	0.667	0.556	0.500	0.667		0.833
CN2 rule inducer	2.046	0.002	1.000	1.000	1.000	1.000	1.000		1.000

Predicted

Actual	1	2	3	5	6	7	Σ
1	0	0	0	0	0	0	0
2	0	1	0	0	0	0	1
3	0	0	0	0	0	0	0
5	0	0	0	1	0	0	1
6	0	0	0	0	0	0	0
7	0	0	0	0	0	1	1
Σ	0	1	0	1	0	1	3

Predicted

Actual	1	2	3	5	6	7	Σ
1	0	0	0	0	0	0	0
2	0	1	0	0	0	0	1
3	0	0	0	0	0	0	0
5	0	0	0	0	0	1	1
6	0	0	0	0	0	0	0
7	0	0	0	0	0	1	1
Σ	0	1	0	0	0	2	3

Jak widzimy jeden przykład został zakwalifikowany jako 7 zamiast 5 (poprawnie dla CN2 i SVM, niepoprawnie dla Tree i Random Forest).