Przewidywanie hepatoksyczności

Nikodem Kwaśniak
https://github.com/nkwasniak/umwpl2021-hepatoxicity-prediction

Opis problemu

Znalezienie odpowiedniej metody, a także dobór odpowiednich jej parametrów w celu określenia czy substancja jest hepatoksyczna.

Innymi słowy chcemy znaleźć jak najlepszy model, który będzie potrafił wskazać toksyczność substancji określanej parametrem ALT - czym wyższa jego wartość tym mocniej substancja jest hepatoksyczna

Dane

Dane użyte do projektu: pliki .csv w postaci fingerprintów - podstrukturalnych Klekota&Roth. - pochodzą z materiałów udostępnionych w ramach przedmiotu.

Prezentują się w następującej postaci:

Field 11																					
	KRFP1	KRFP2	KRFP3	KRFP4	KRFP5	KRFP6	KRFP7	KRFP8	KRFP9	KRFP10	KRFP11	KRFP12	KRFP13	KRFP14	KRFP15	KRFP16	KRFP17	KRFP18	KRFP19	KRFP20	KRFP21	KRFP22
32.6																						
	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37.2																						
	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2.92																						
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3.16																						
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44.1																						
	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43.21																						
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36																						
	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22.8																						
	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Przetwarzanie danych

Aby ułatwić dalszą pracę z danymi, zostały one przetworzone do następującej postaci:

ALT	KRFP1	KRFP2	KRFP3	KRFP4	KRFP5	KRFP6	KRFP7	KRFP8	KRFP9	KRFP10	KRFP11	KRFP12	KRFP13	KRFP14	KRFP15	KRFP16	KRFP17	KRFP18	K
32.6	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
37.2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2.92	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3.16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
44.1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
43.21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
36	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
22.8	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
27.1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
35.9	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
45.3	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
28.1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
40	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
80	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
273	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	
28	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	

Dodatkowo skrypt konwertujący został wzbogacony o testy.

Wstępna analiza danych

	ALT	KRFP1	KRFP2	KRFP3	KRFP4	KRFP5	KRFP6	KRFP7	KRFP8	KRFP9	...	KRFP4851	KRFP4852	KRFP4853	KRFP4854	KRFP4855	KRFP4856	KI
0	32.60	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
1	37.20	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	
2	2.92	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	1.0	0.0	0.0	0.0	1.0	
3	3.16	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	1.0	0.0	0.0	0.0	1.0	
4	44.10	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	

5 rows × 4861 columns

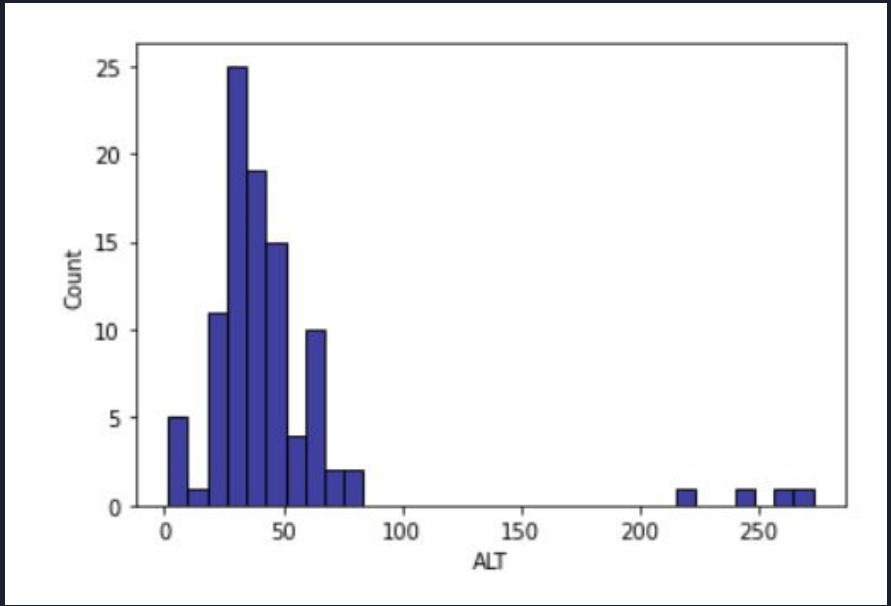
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 98 entries, 0 to 97
Columns: 4861 entries, ALT to KRFP4860
dtypes: float64(4861)
memory usage: 3.6 MB
```

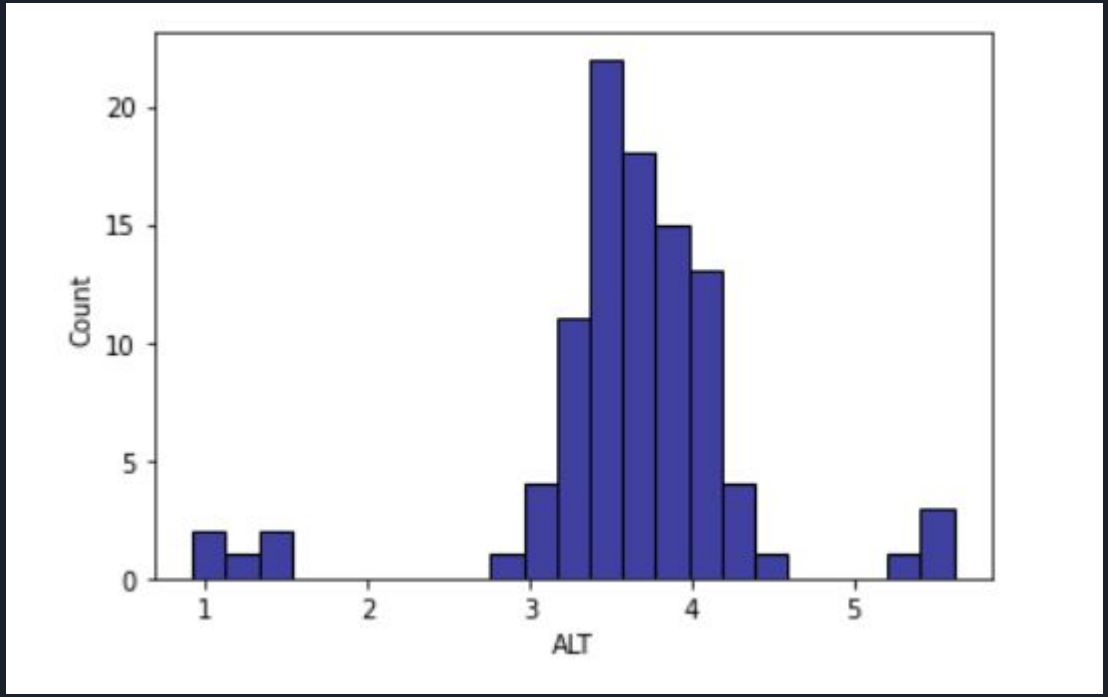
Danych jest dosyć niewiele, bo na tylko 98 rekordów

Ponowne przetwarzanie danych

Usunięcie kolumn z zerową wariancją, pozwala na odrzucenie danych nic nie wnoszących do końcowego wyniku. Ostateczny podgląd danych potwierdza że wartości ALT są zbyt daleko od siebie. Aby nie tracić danych (poprzez np. obcięcie zbioru) ponieważ nie jest ich wiele, parametr ALT ulega zlogarytmowaniu przez funkcję 1+x.



Skorygowanie parametru ALT



Wyszukiwanie najlepszych hiperparametrów dla modeli

Użycie gridsearch do wyszukania najlepszych hiperparametrów przy pomocy podwójnej 5krotnej walidacji krzyżowej, dla modeli regresji grzbietowej, SVR, LinearRegerssion i Lasso.

	mean_fit_time	std_fit_time	mean_score_time	std_score_time	param_alpha	param_random_state	params	split0_test_score	split1_test_score	split2_te
0	0.018450	0.007042	0.000599	0.000489	0.001	12345	{'alpha': 0.001, 'random_state': 12345}	0.052402	-0.025968	
1	0.018945	0.007786	0.000304	0.000464	0.001	123	{'alpha': 0.001, 'random_state': 123}	0.052402	-0.025968	
2	0.020097	0.008939	0.000399	0.000489	0.001	666	{'alpha': 0.001, 'random_state': 666}	0.052402	-0.025968	
3	0.018568	0.007762	0.000299	0.000457	0.001	123123	{'alpha': 0.001, 'random_state': 123123}	0.052402	-0.025968	
4	0.018550	0.007083	0.000299	0.000457	0.001	777	{'alpha': 0.001, 'random_state': 777}	0.052402	-0.025968	

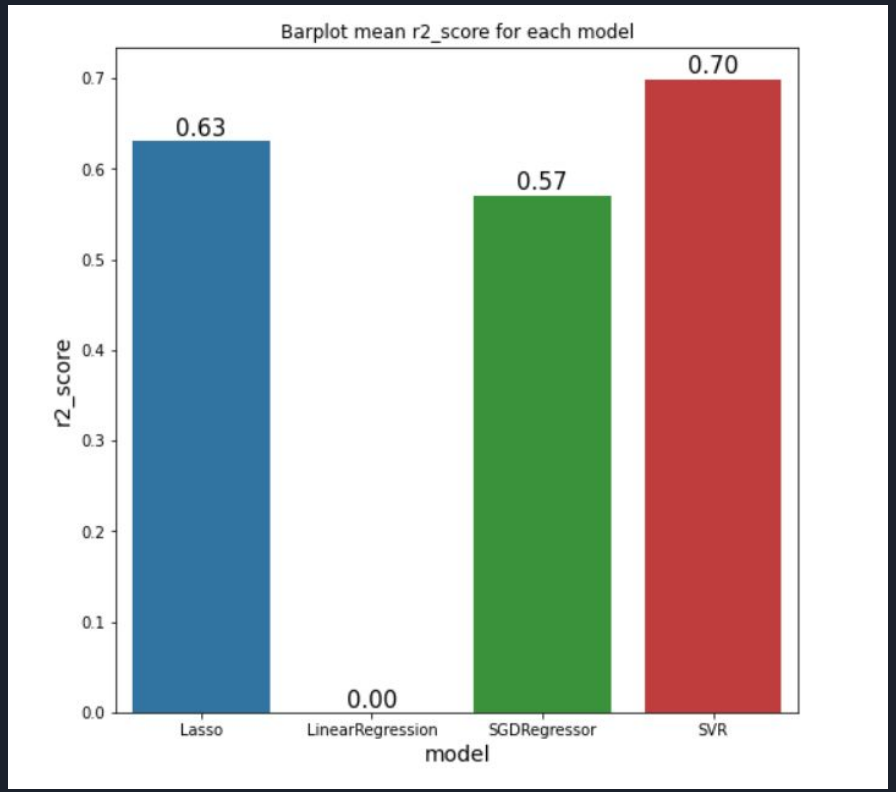
Znalezienie najlepszych hiperparametrów nie uwzględniając parametru random_state

param_C	param_coef0	param_degree	param_gamma	param_kernel	param_tol	mean_test_score	std_test_score
0.0001	0.01	3	auto	linear	0.00001	-0.039704	0.063404
					0.00010	-0.039705	0.063403
					0.00100	-0.039720	0.063451
					0.01000	-0.039778	0.063227
					0.10000	-0.036781	0.061831
...	...	...	...	...	...	...	...
25.0000	10.00	8	scale	sigmoid	0.00001	-0.056269	0.063646
					0.00010	-0.056269	0.063646
					0.00100	-0.056269	0.063646
					0.01000	-0.056274	0.063647
					0.10000	-0.051496	0.058165

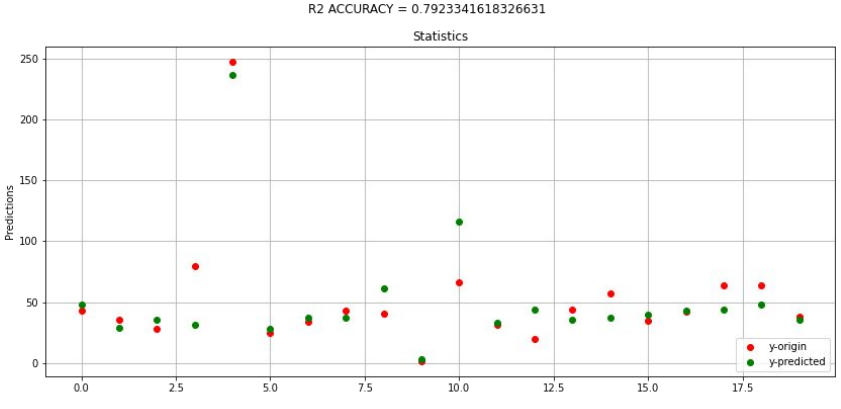
	model	params
0	SVR	(0.1, 0.01, 3, auto, linear, 1e-05)
1	Lasso	0.01
2	SGDRegressor	(0.0, invscaling)
3	LinearRegressor	0

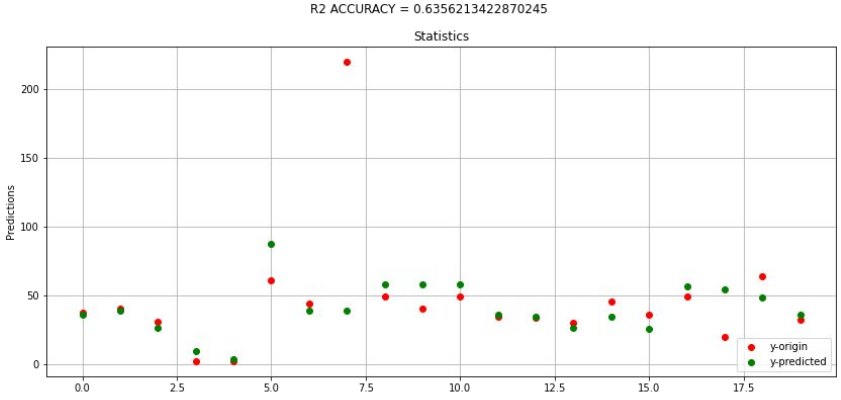
param_alpha	mean_test_score	std_test_score
0.001	0.206384	0.617849
0.005	0.291197	0.512019
0.010	0.322450	0.456979
0.050	0.202680	0.258652
0.100	-0.008026	0.101629
0.500	-0.074849	0.068368
1.000	-0.074849	0.068368

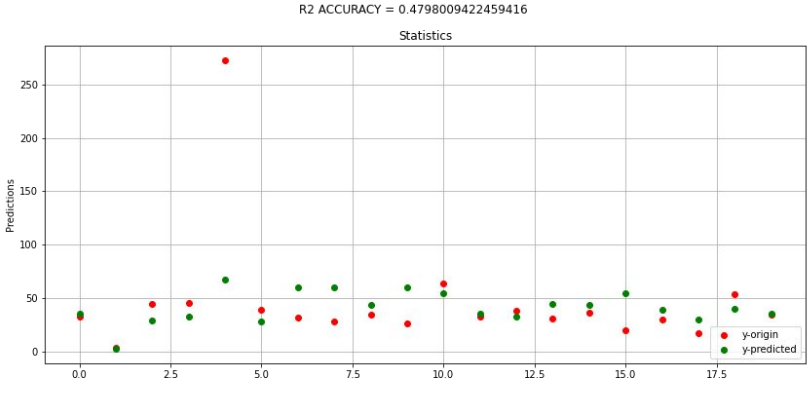
Wyniki predykcji modeli - współczynnik r2

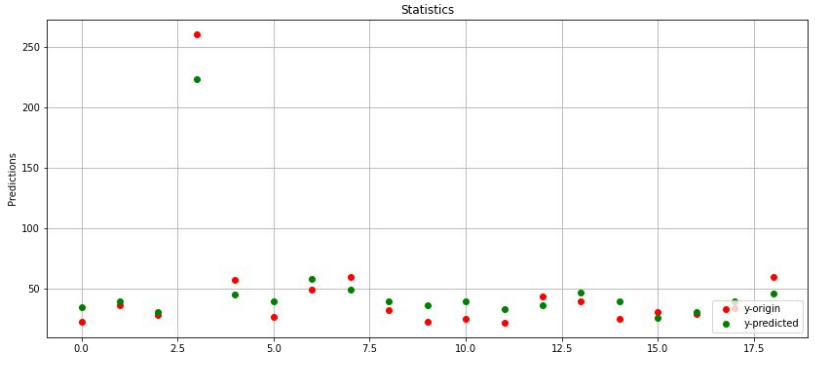


Współczynnik R^2 każdego modelu

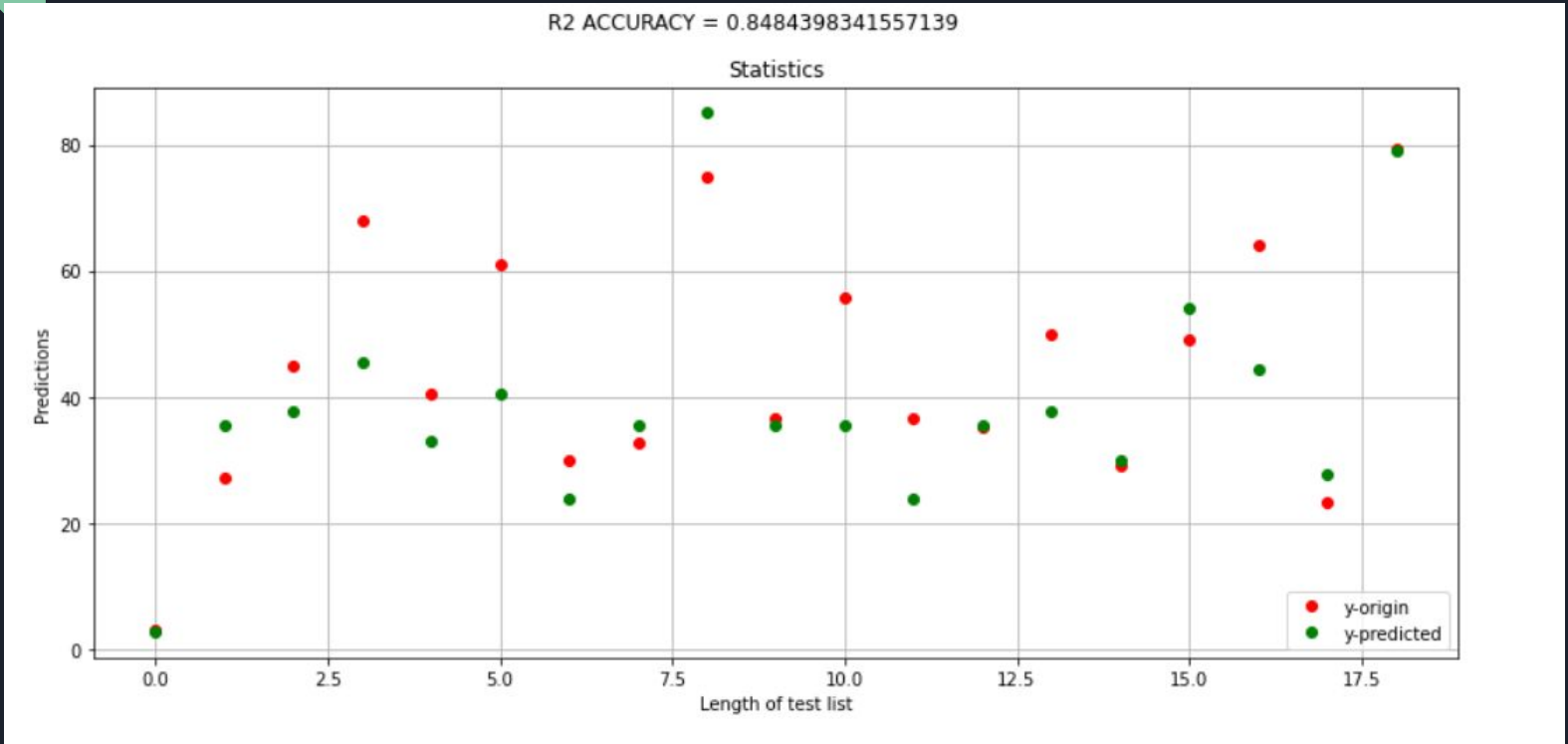








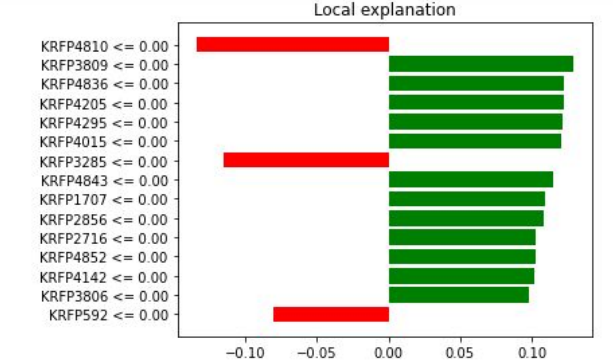
Analiza predykcji najlepszego modelu

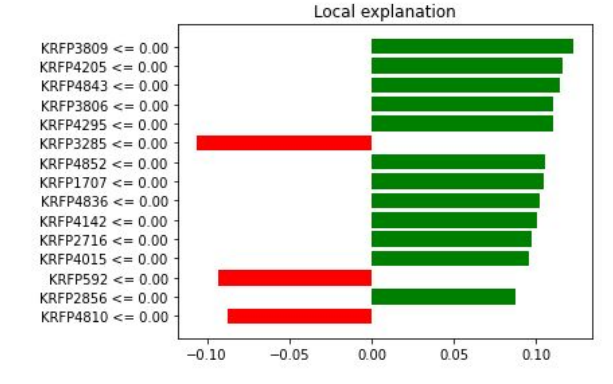


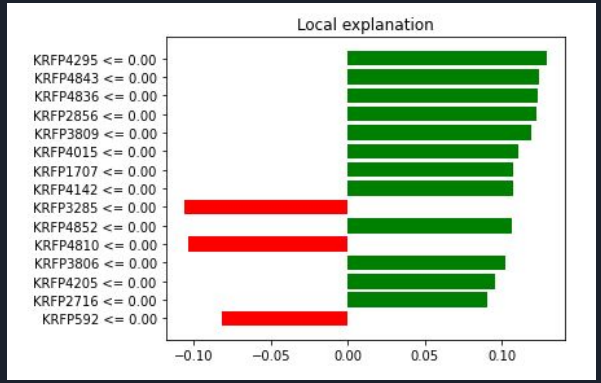
Lista bitów ważnych i nieważnych w kontekście wpływu na wynik

	unimportant_bits	important_bits
0	KRFP1750 <= 0.00	KRFP4142 <= 0.00
1	KRFP852 <= 0.00	KRFP1799 <= 0.00
2	KRFP1524 <= 0.00	KRFP1566 > 0.00
3	KRFP670 <= 0.00	KRFP2988 <= 0.00
4	KRFP3177 <= 0.00	KRFP4852 > 0.00
5	KRFP232 <= 1.00	KRFP2716 <= 0.00
6	KRFP3915 <= 0.00	KRFP4142 > 0.00
7	KRFP55 <= 0.00	KRFP1707 <= 0.00
8	KRFP3222 <= 1.00	KRFP4836 > 0.00
9	KRFP2714 <= 0.00	KRFP4295 > 0.00

Analiza wpływu poszczególnych fingerprintów na predykcję







KONIEC