Przewidywanie kardiotoksyczności związków

Mateusz Poleski

Plan

1. Analiza danych i pre-processing.
2. Regresja
3. Klasyfikacja
4. Porównanie regresji z klasyfikacją.

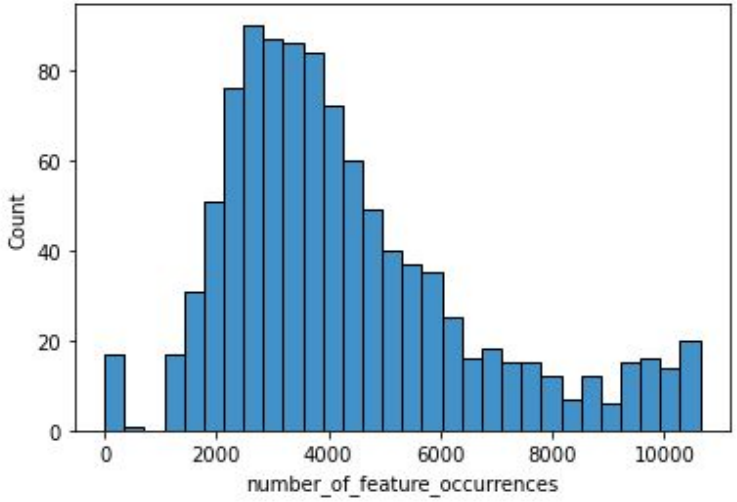
Przypomnienie projektu

- Przewidywanie kardiotoksyczności związków poprzez predykcje oddziaływania ich z kanałami potasowymi hERG.
- Identyfikacja cech odpowiedzialnych za toksyczność.
- Dostępne dane:
  - **Fingerprinty podstrukturalne**
    - **Klekota&Roth**
    - **MACCSFP (Molecular ACCess System)**
  - **Fingerprinty haszowane**

Analiza danych

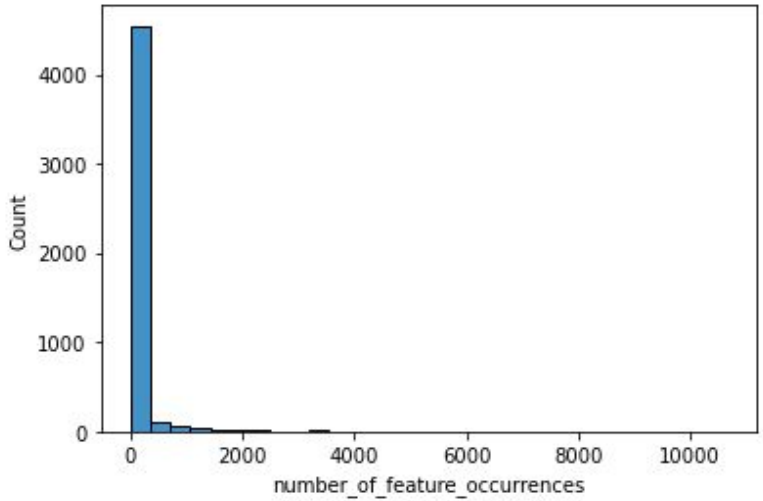
- Pliki z fingerprintami:
  - Klekota&Roth - 4860 features
  - MACCSFP (Molecular ACCess System) - 166 features
  - Hashed - 1024 features
- 11504 rekordów (substancji) w każdym z plików, z czego ~7.5% ma przypisane niepoprawne wartości IC50.
- Struktura każdego z plików jest bardzo podobna, co ułatwia łączenie ich.

Częstotliwości występowania cech - Hashed FP



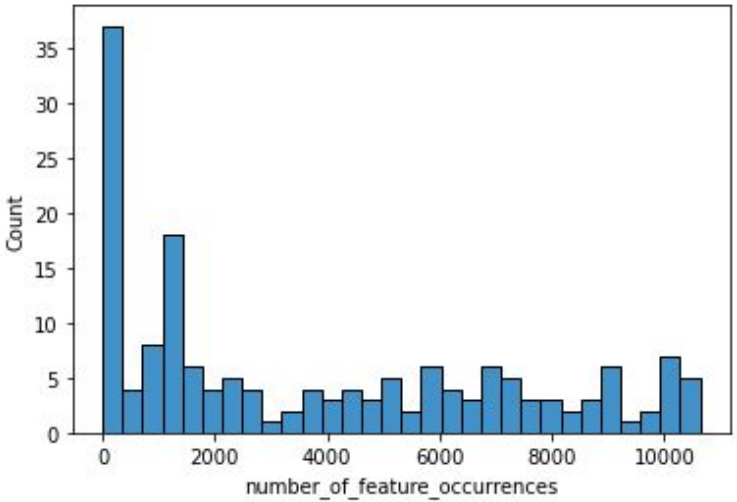

```
1.37% cech nie występuje w żadnym ze związków.
1.56% cech występuje w mniej niż 0.1% wszystkich związków.
1.66% cech występuje w mniej niż 1.0% wszystkich związków.
1.76% cech występuje w mniej niż 10.0% wszystkich związków.
74.32% cech występuje w mniej niż 50.0% wszystkich związków.
```

Częstotliwości występowania cech - Klek&Roth



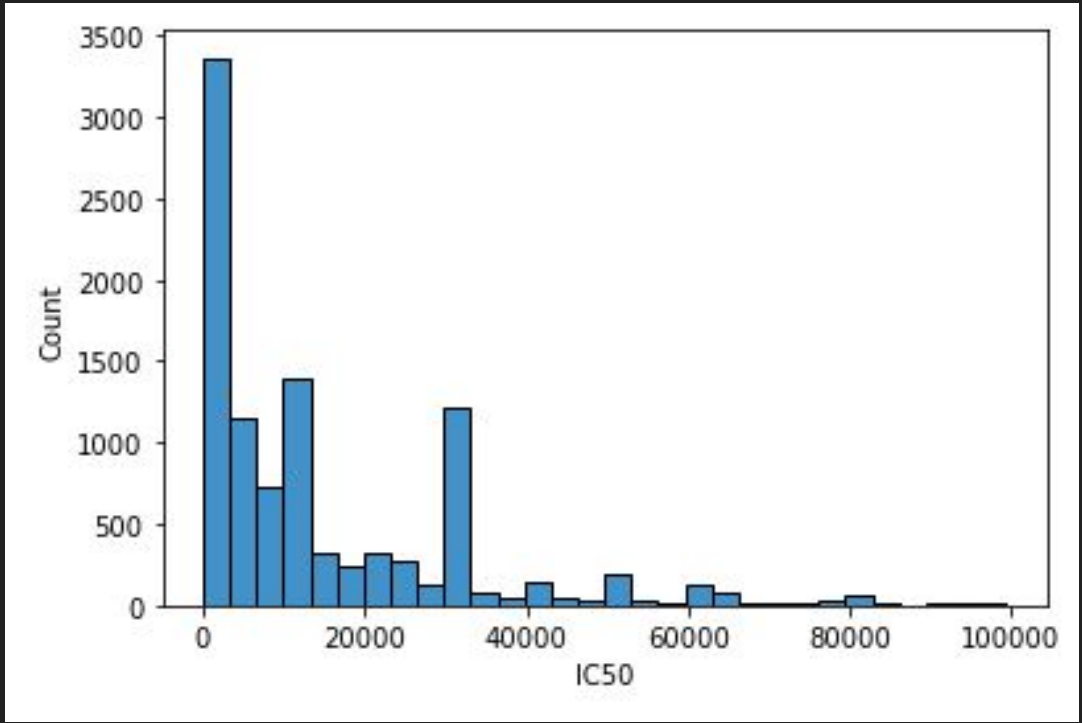

```
56.79% cech nie występuje w żadnym ze związków.
73.87% cech występuje w mniej niż 0.1% wszystkich związków.
87.33% cech występuje w mniej niż 1.0% wszystkich związków.
96.73% cech występuje w mniej niż 10.0% wszystkich związków.
99.28% cech występuje w mniej niż 50.0% wszystkich związków.
```

Częstotliwości występowania cech - MACCSFP

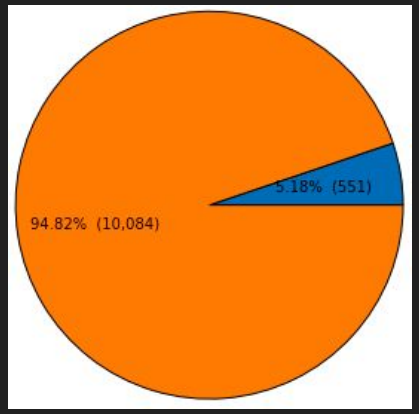


8.43% cech nie występuje w żadnym ze związków.
11.45% cech występuje w mniej niż 0.1% wszystkich związków.
17.47% cech występuje w mniej niż 1.0% wszystkich związków.
29.52% cech występuje w mniej niż 10.0% wszystkich związków.
65.06% cech występuje w mniej niż 50.0% wszystkich związków.

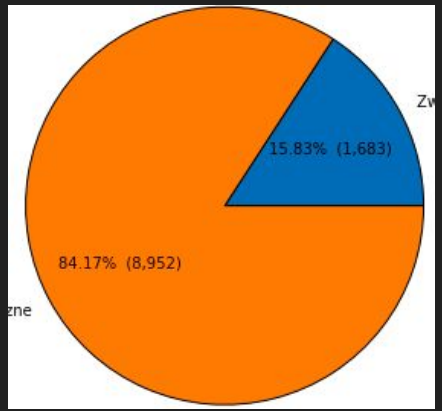
Histogram rozkładu wartości IC50



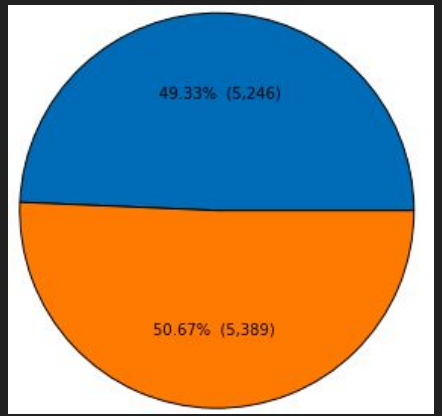
Rozkład klasyfikacji w zależności do progu IC50



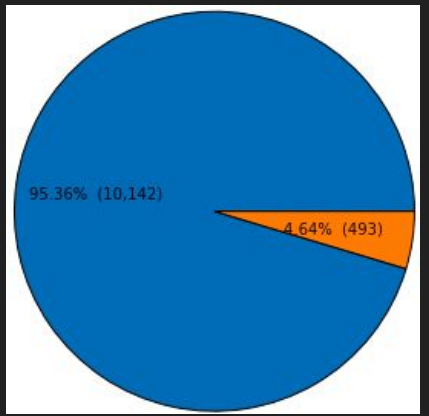
IC50: 100



IC50: 1000

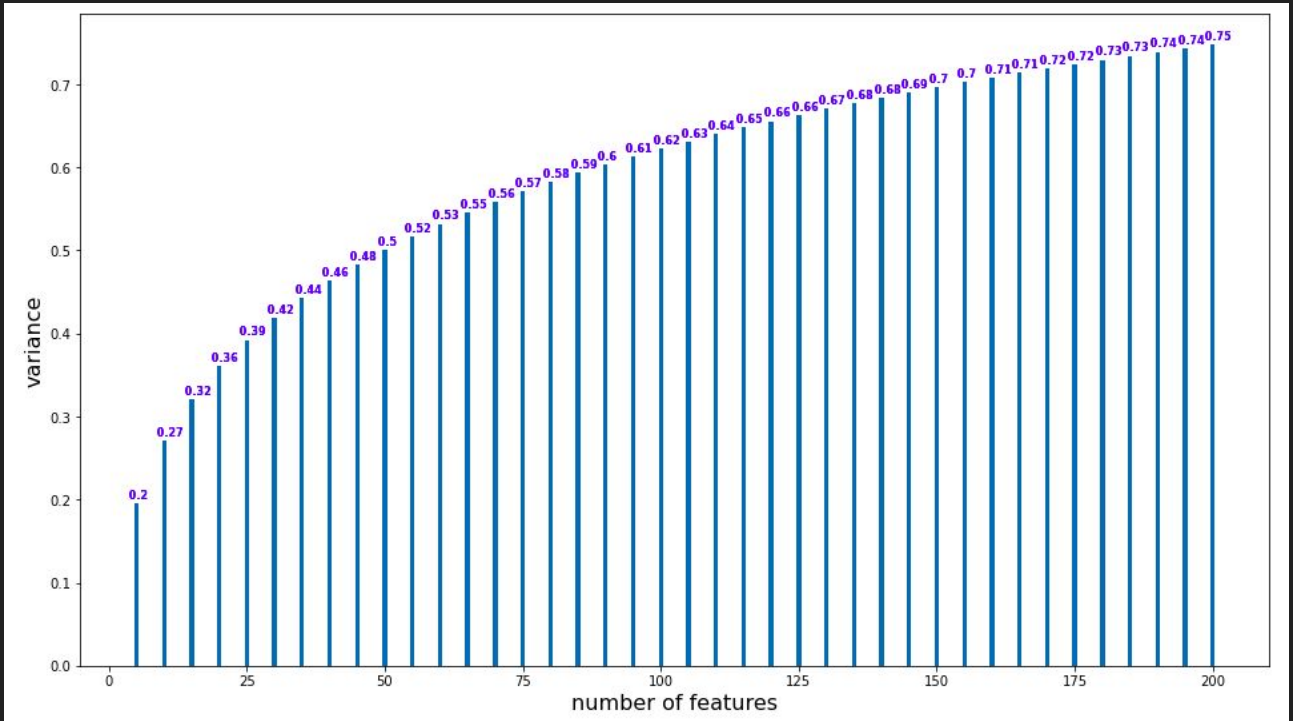


IC50: 10 000



IC50: 100 000

Analiza istotności featerów - PCA



Regresja

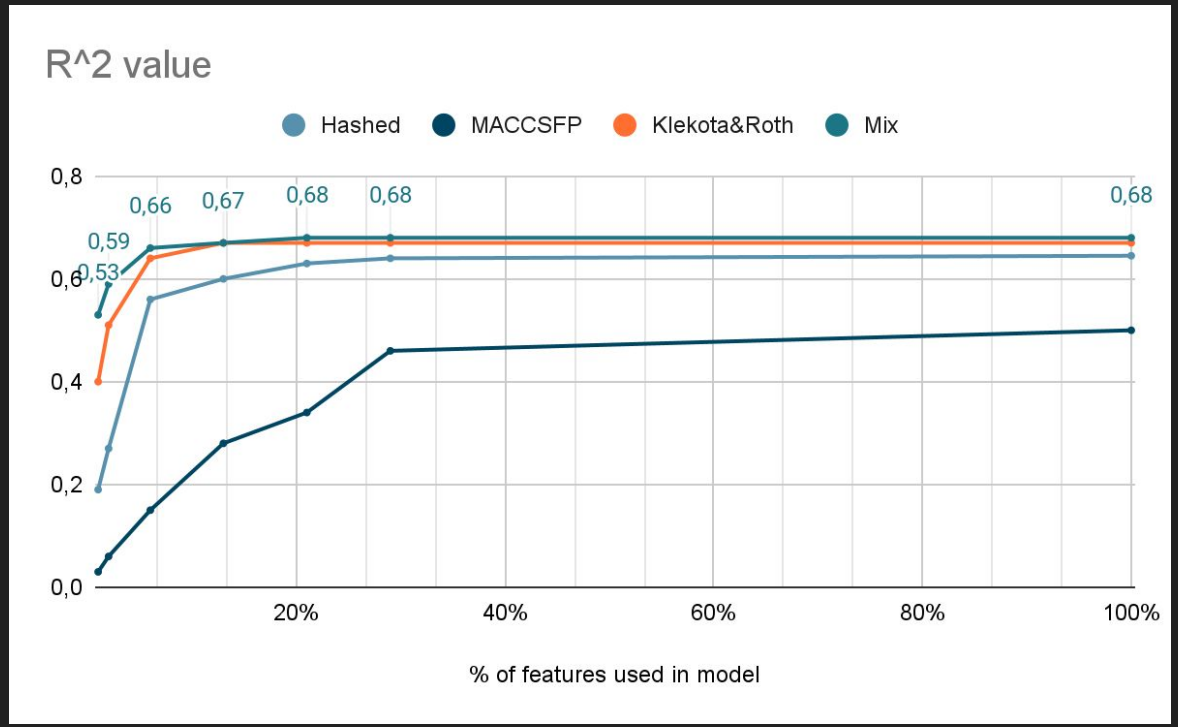
Regresja

- Wykorzystałem kilka standardowych modeli oraz prostą sieć neuronową.
- Modele były testowane na różnych plikach z fingerprintami osobno oraz na połączeniu wszystkich trzech plików.
- Modele były także testowane wybierając tylko jakąś część najbardziej istotnych featerów.

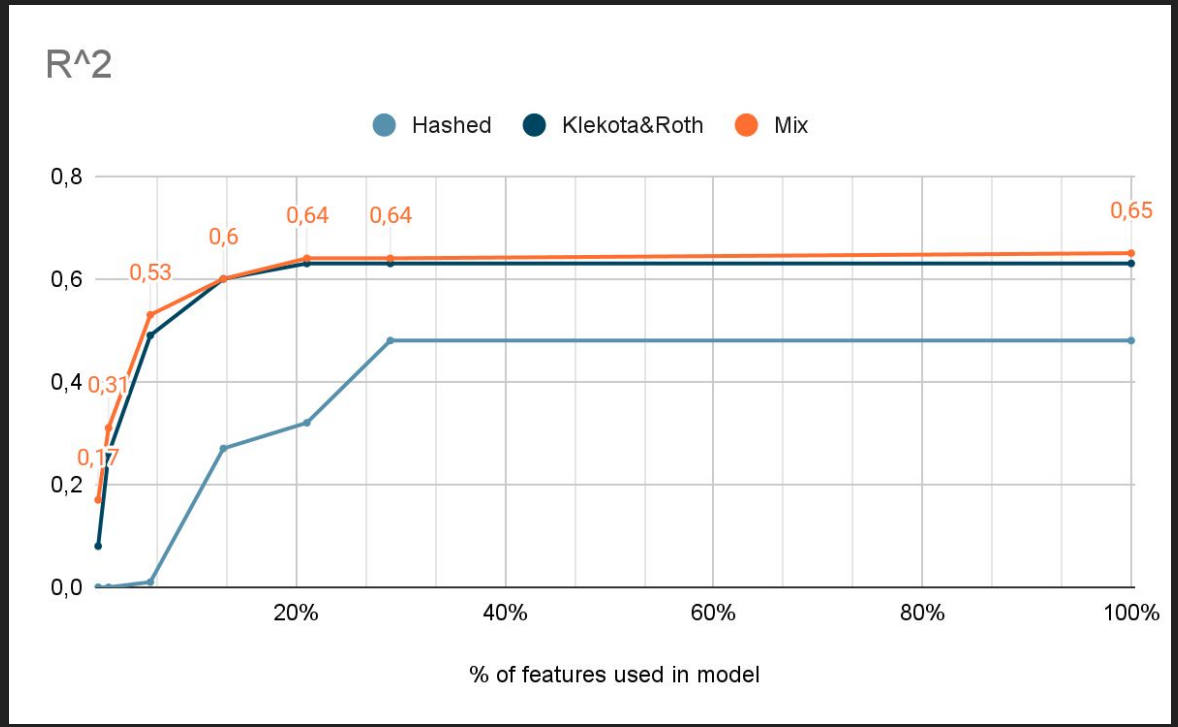
Zastosowane metody

Regression model:	R^2:	
Ridge	0.521879	{'alpha': 10.0, 'max_iter': 50000}
Lasso	0.508381	{'alpha': 1e-09, 'max_iter': 10000}
Elastic Net	0.508381	{'alpha': 1e-09, 'max_iter': 10000}
Bayesian	0.498869	{'alpha_1': 10.0, 'alpha_2': 1e-09}
SGD	0.476612	{'alpha': 0.0001}
SVM	0.686291	{'C': 10.0}
Decision Tree	0.359572	{'ccp_alpha': 0}
K Neighbors	0.563361	{'algorithm': 'auto'}
Best		
SVM	0.686291	{'C': 10.0}

Regresja - metody klasyczne



Regresja - sieć neuronowa



Klasyfikacja

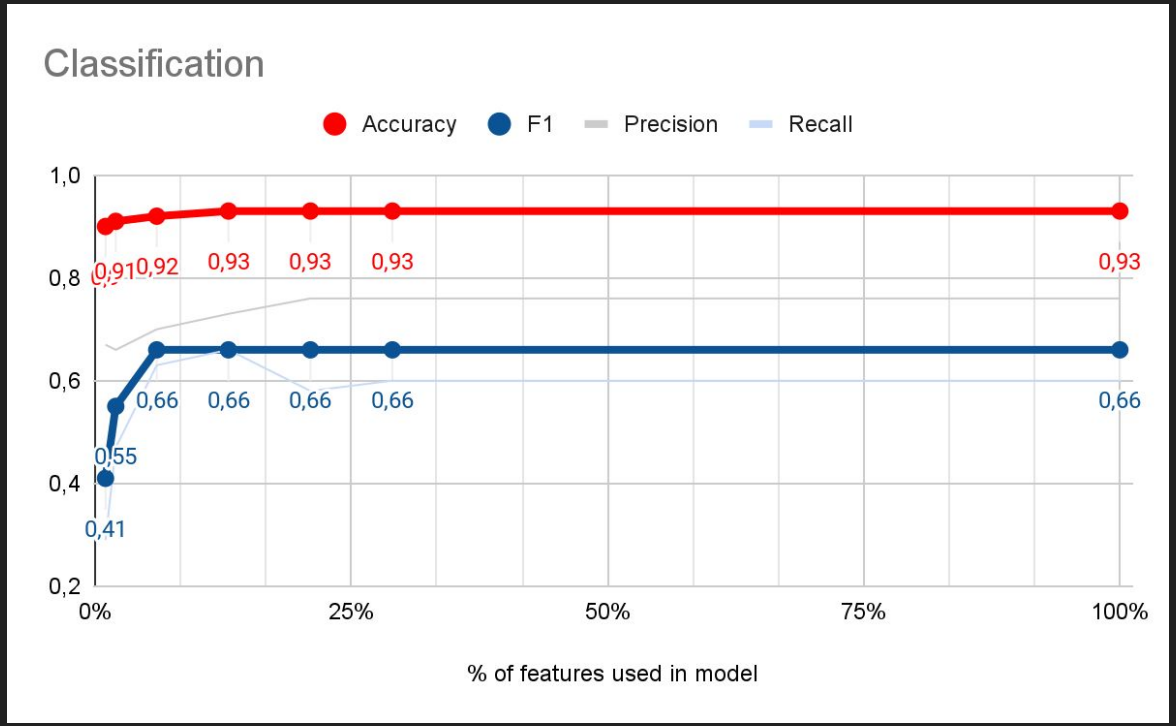
Klasyfikacja

- Wykorzystałem kilka standardowych modeli oraz prostą sieć neuronową.
- Modele były testowane tylko na połączeniu wszystkich trzech plików z fingerprintami.
- Modele były także testowane wybierając tylko jakąś część najbardziej istotnych featerów.
- Sprawdziłem również jak modele zachowują w zależności od wartości progu IC50.

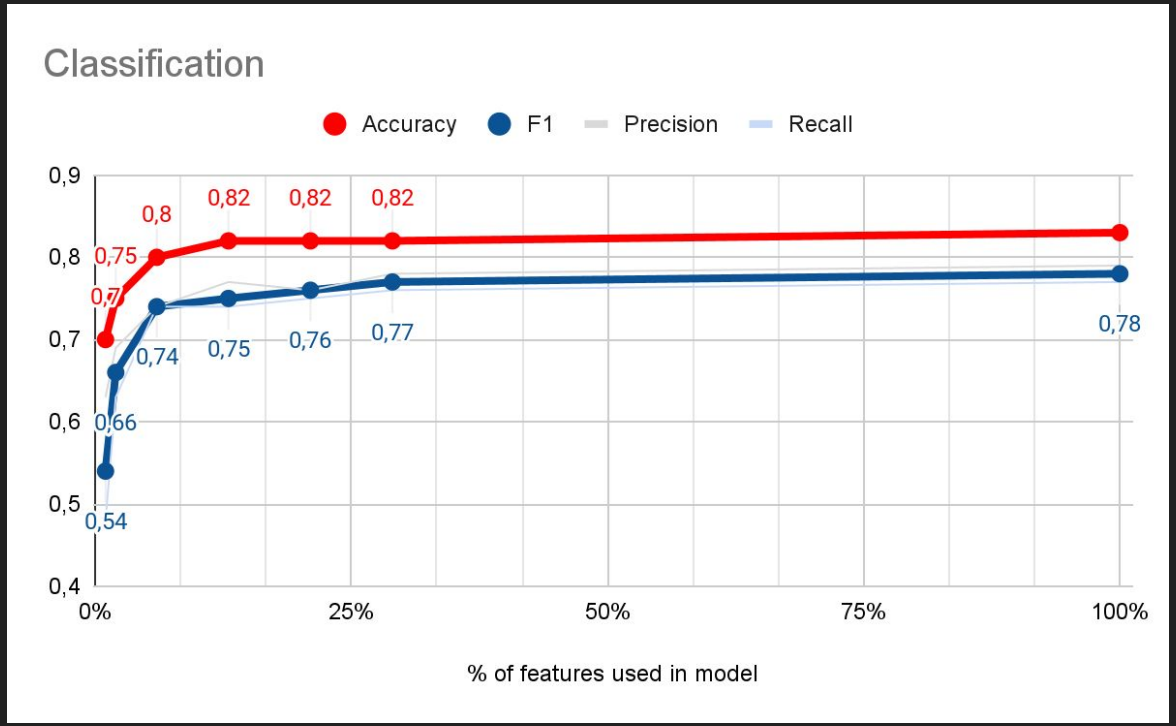
Zastosowane metody

```
Classification model:    accuracy:   F1:
SVC RBF                  0.790385    0.792972    {'C': 10, 'gamma': 0.01}
SVC Poly                 0.768269    0.769157    {'C': 1, 'coef0': 1, 'degree': 2}
SVC Linear               0.702404    0.711422    {'C': 0.1}
Logistic                 0.698077    0.704887    {'C': 0.1}
Random Forest            0.671635    0.648482    {'max_leaf_nodes': 16, 'n_estimators': 500}
Gradient Boosting        0.776442    0.777831    {'learning_rate': 0.1, 'n_estimators': 500}
XGBoosting               0.769231    0.769674    {'learning_rate': 0.01, 'max_depth': 10, 'n_estimators': 100}
K Nearest Neighbors      0.763462    0.769231    {'leaf_size': 50, 'n_neighbors': 5}

Best accuracy
SVC RBF                  0.790385    {'C': 10, 'gamma': 0.01}


Best F1
SVC RBF                  0.792972    {'C': 10, 'gamma': 0.01}
    preccision           0.8098933074684772
    recall               0.7767441860465116
```

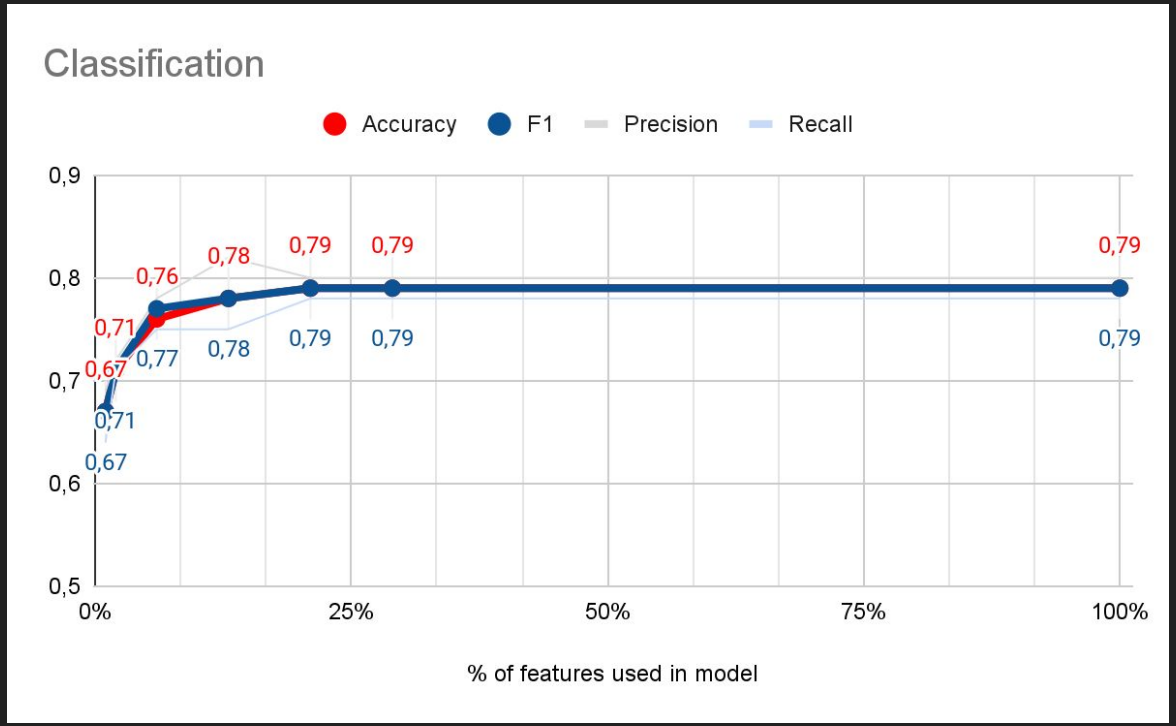
IC50 threshold: 500 (11.55%/88.45% Toxic/Non Toxic)



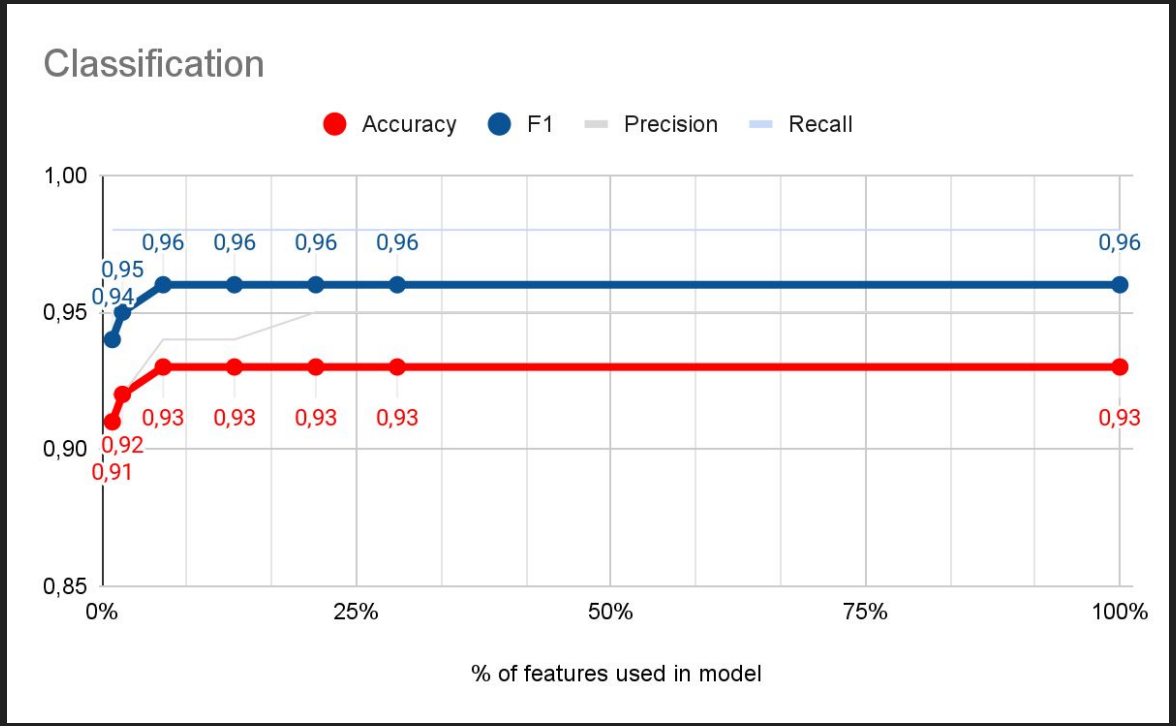
IC50 threshold: 5 000 (37.78%/62.22% Toxic/Non Toxic)



IC50 threshold: 10 000 (50.17%/49.83% Toxic/Non Toxic)



IC50 threshold: 50 000 (91.26%/8.74% Toxic/Non Toxic)

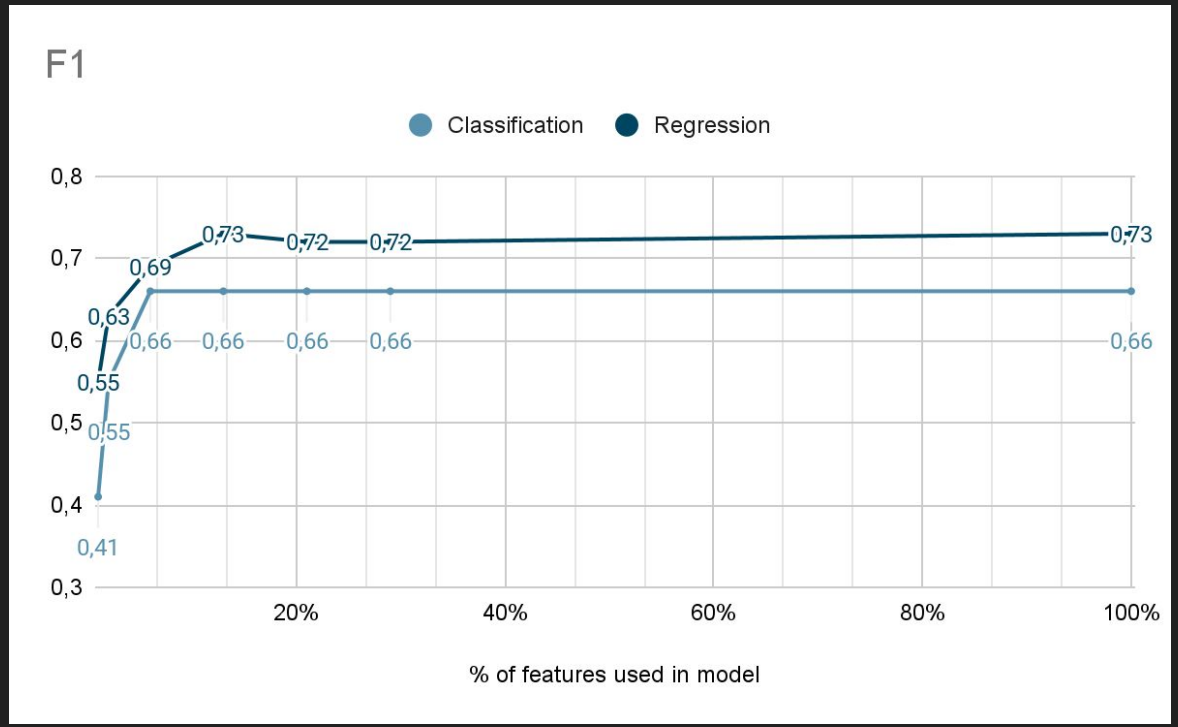


Klasyfikacja vs. Regresja

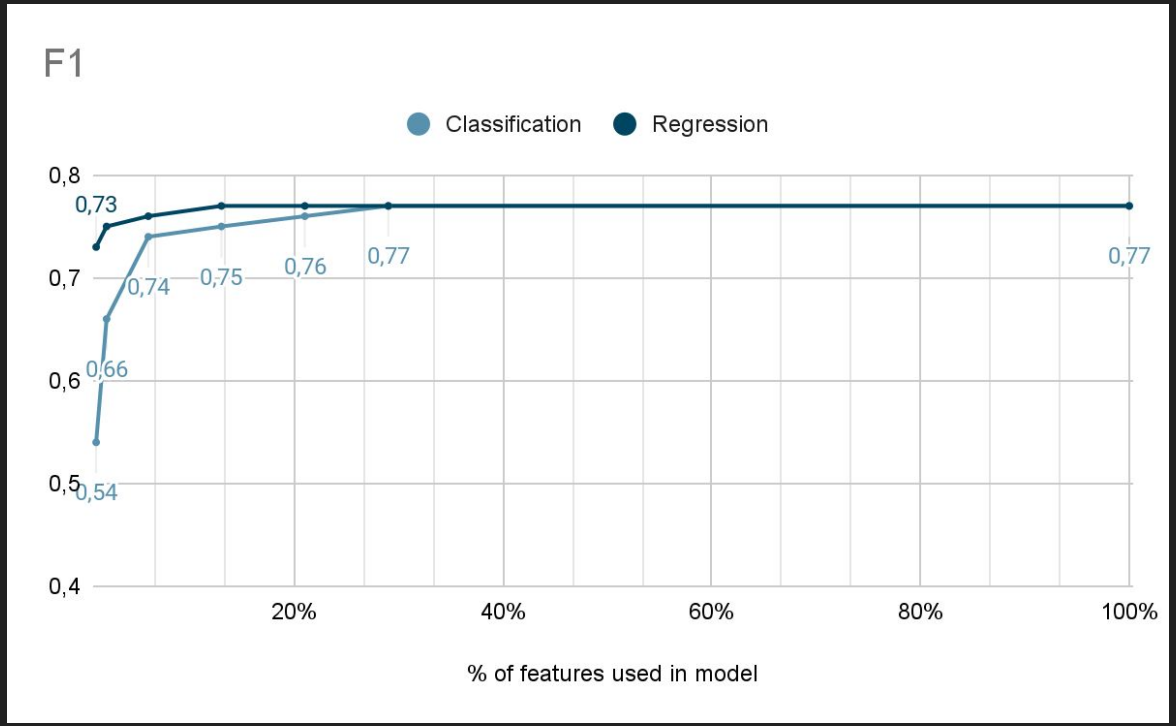
Metody

- Najpierw przeprowadzam regresje najlepszym modelem regresyjnym, a następnie wyniki zamieniam na klasyfikację wg. progu IC50.
- Eksperymenty przeprowadzone na połączeniu wszystkich trzech plików z fingerprintami.
- Wyniki sprawdzam dla różnych progów IC50.
- Tym razem już całkowicie pomijam sieci neuronowe, które otrzymywały słabsze wyniki.

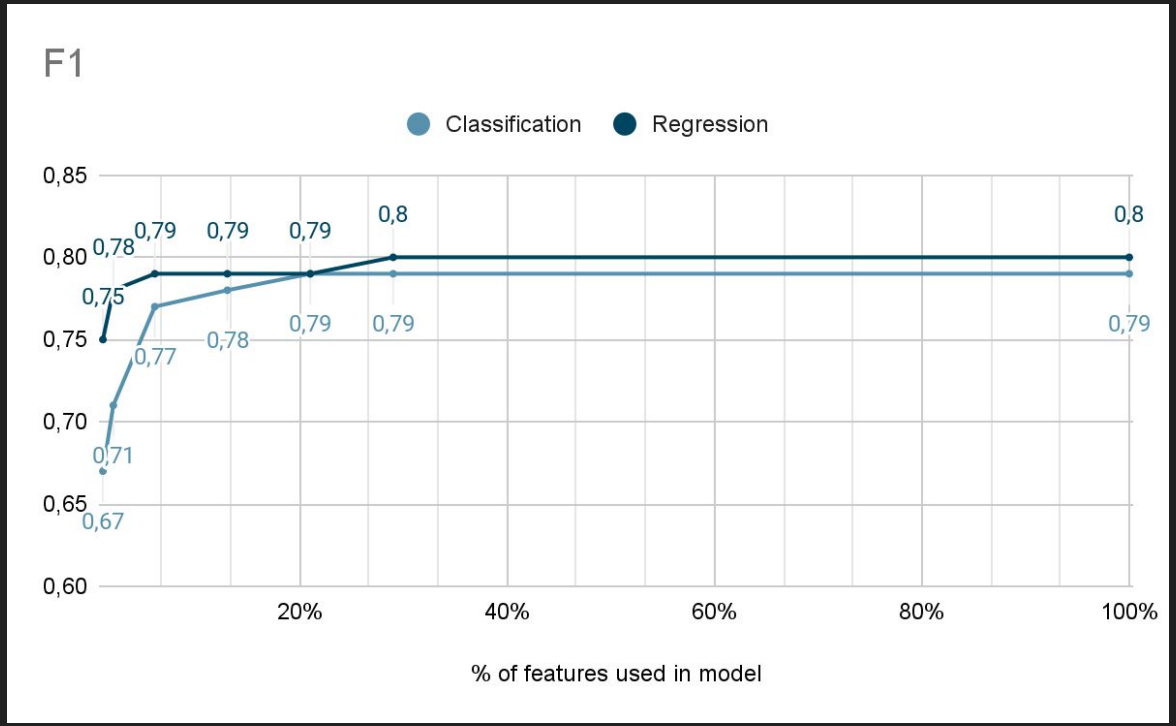
IC50 threshold: 500 (11.55%/88.45% Toxic/Non Toxic)



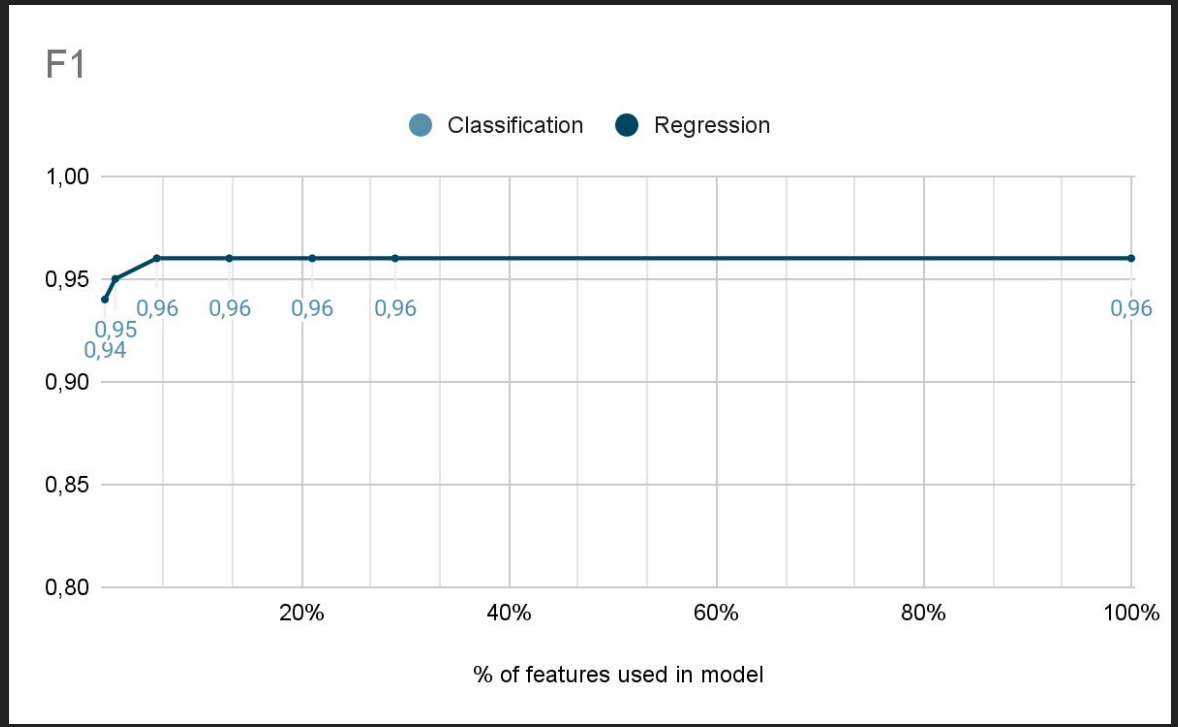
IC50 threshold: 5 000 (37.78%/62.22% Toxic/Non Toxic)



IC50 threshold: 10 000 (50.17%/49.83% Toxic/Non Toxic)



IC50 threshold: 50 000 (91.26%/8.74% Toxic/Non Toxic)



Dziękuję za uwagę