Katarzyna Kanicka, 331388

WSI Sprawozdanie 4 – Regresja i klasyfikacja

Treść polecenia:

1. Zaimplementować algorytm regresji logistycznej.
2. Sprawdzić jakość działania algorytmu dla klasyfikacji na zbiorze danych Census Income https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic.
3. Policzyć wynik dla przynajmniej 3 różnych sposobów przygotowania danych, na przykład usuwając niektóre kolumny, dodając normalizację wartości.

Instrukcja do używania skryptu

Po pobraniu repozytorium należy:

- przejść do głownego katalogu – lab3

```
$ cd lab4
```

- stworzyć środowisko wirtualne i pobrać requirements.txt

```
$ python3 -m venv .venv
$ source .venv/bin/activate
$ pip install -r requirements.txt
```

- uruchomić plik main.py, w terminalu za pomocą:

```
$ python3 main.py [argumenty]
```

Argumenty:

- **seed** – ziarno
- **learning-rate** – wielkość kroku uczącego
- **iterations** – liczba iteracji
- **normalize** – czy dane mają zostać znormalizowane (z-score normalization)
- **exclude-columns** – liczba kolumn do wykluczenia z trenowania

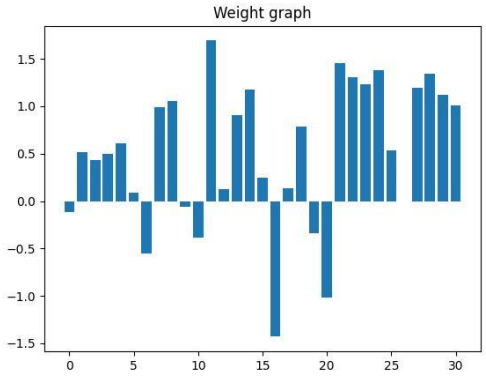
Wartości korelacji dla każdego x

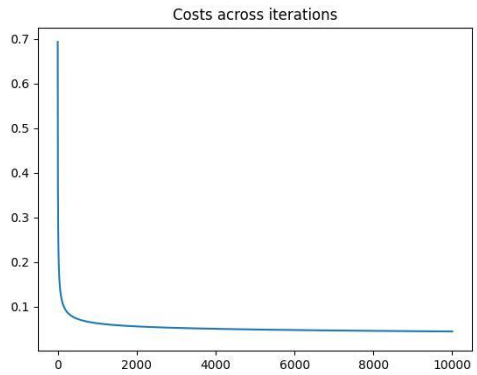
x	correlation
concave_points3	0.793566
perimeter3	0.782914
concave_points1	0.776614
radius3	0.776454
perimeter1	0.742636
area3	0.733825
radius1	0.730029
area1	0.708984
concavity1	0.696360
concavity3	0.659610
compactness1	0.596534
compactness3	0.590998
radius2	0.567134
perimeter2	0.556141
area2	0.548236
texture3	0.456903
smoothness3	0.421465
symmetry3	0.416294
texture1	0.415185
concave_points2	0.408042
smoothness1	0.358560
symmetry1	0.330499
fractal_dimension3	0.323872
compactness2	0.292999
concavity2	0.253730
fractal_dimension2	0.077972
symmetry2	-0.006522
texture2	-0.008303
fractal_dimension1	-0.012838
smoothness2	-0.067016

Wpływ parametrów algorytmu

- learning_rate = 0.06, iterations = 10000

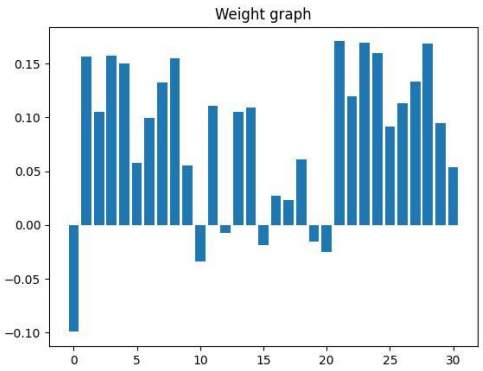
	accuracy	F1	auroc
0	0.972028	0.963636	0.970455

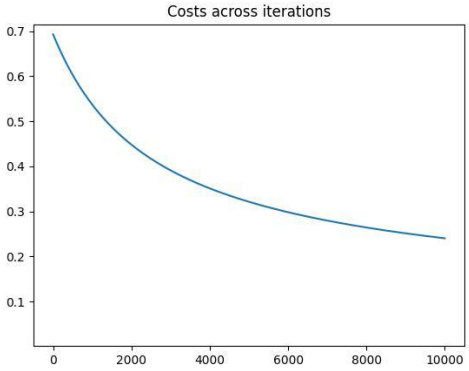




- learning_rate = 0.0001, iterations = 10000

	accuracy	F1	auroc
0	0.937063	0.915888	0.928409

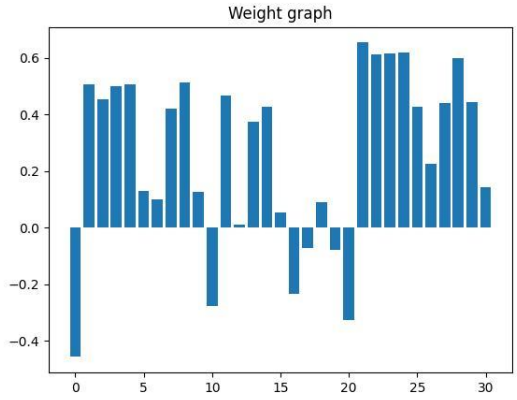


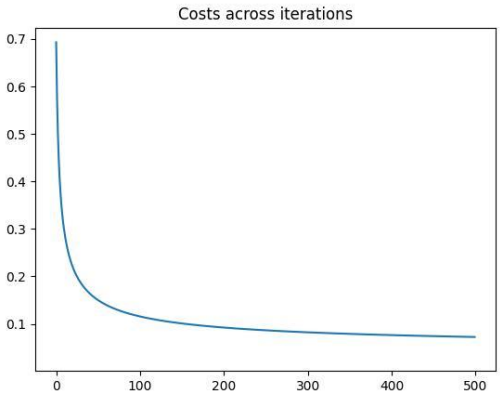


Krok uczący jest za mały, aby osiągnąć dobrą dokładność w określonej liczbie iteracji.

- learning_rate = 0.06, iterations 500

	accuracy	F1	auroc
0	0.958042	0.943396	0.948864



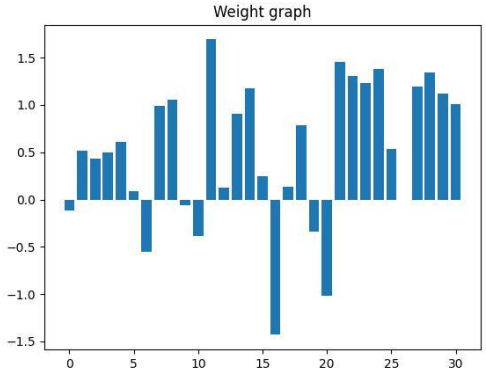


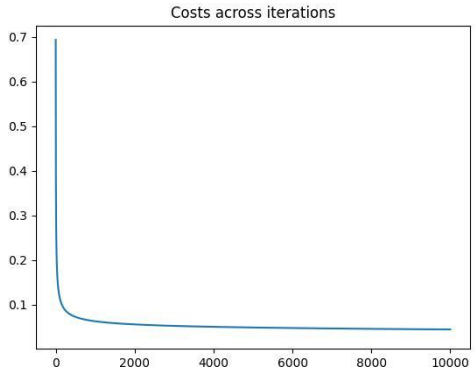
Mała ilość iteracji ma wpływ na jakość modelu regresji logistycznej.

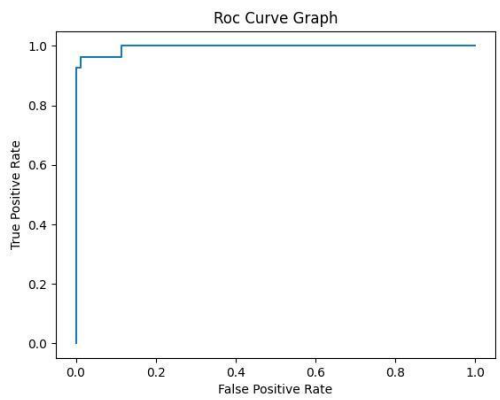
Wpływ normalizacji danych

- z normalizacją danych

	accuracy	F1	auroc
0	0.972028	0.963636	0.970455



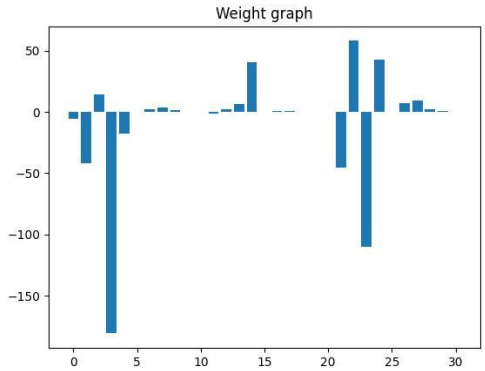


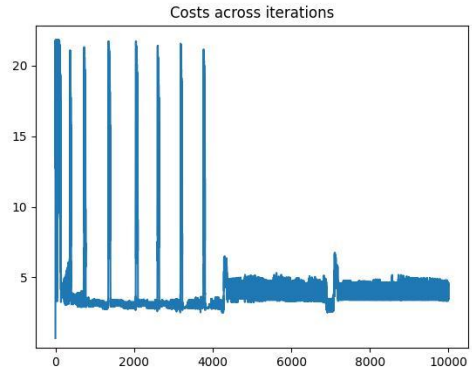


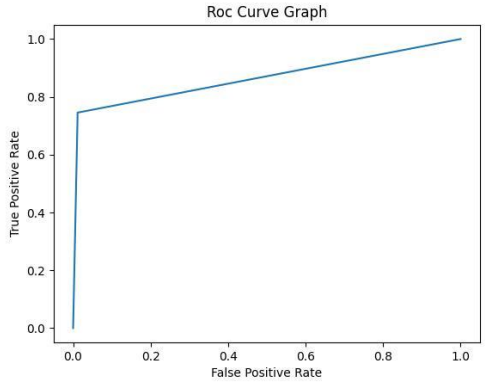
Użyto normalizacji z-score ($z = \frac{x-\mu}{\sigma}$), która przekształca dane tak, aby były w standardowym rozkładzie normalnym. Jest to pomocne, jeśli wartości wejściowe są w różnych skalach/jednostkach. Dzięki tej normalizacji wyniki są duże lepsze.

- bez normalizacji danych

	accuracy	F1	auroc
0	0.888112	0.833333	0.857955





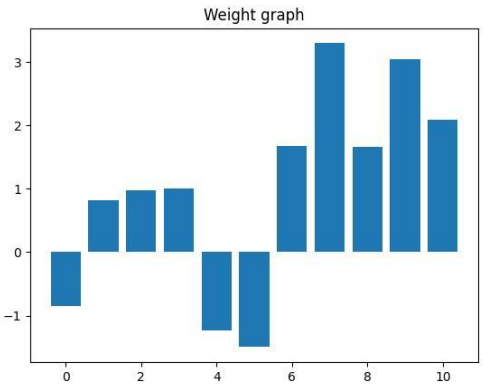


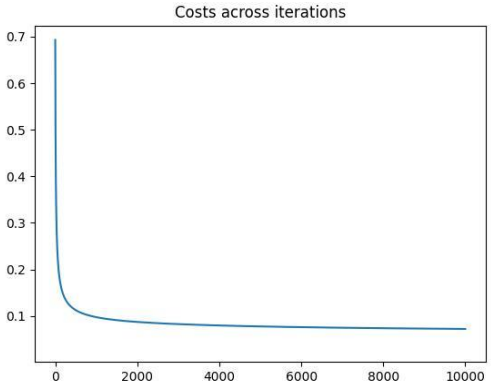
Model jest dużo mniej skuteczny bez normalizacji. Nie sprawdza się dla tych samych parametrów. Wartości wag i kosztów mają duże przeskoki. Dałyby lepsze wyniki dla bardzo dużej liczby iteracji, co znacznie by wydłużyło czas uczenia.

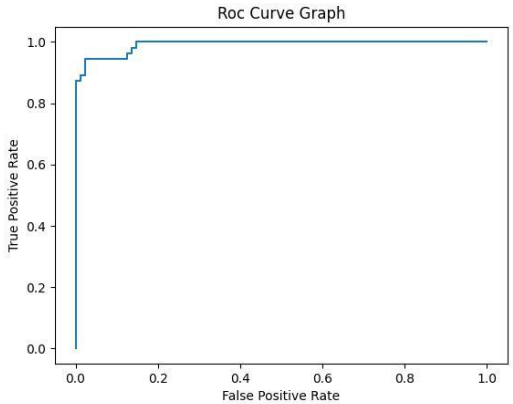
Wpływ usuwanie kolumn (losowo)

- **Obecne kolumny**: 'radius1', 'smoothness1', 'compactness1', 'fractal_dimension1', 'compactness2', 'concave_points2', 'radius3', 'texture3', 'area3', 'fractal_dimension3'

	accuracy	F1	auroc
0	0.951049	0.934579	0.943182





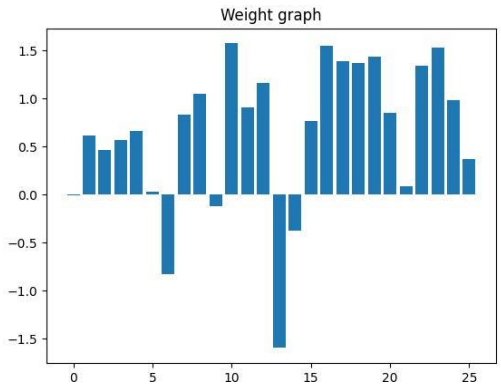


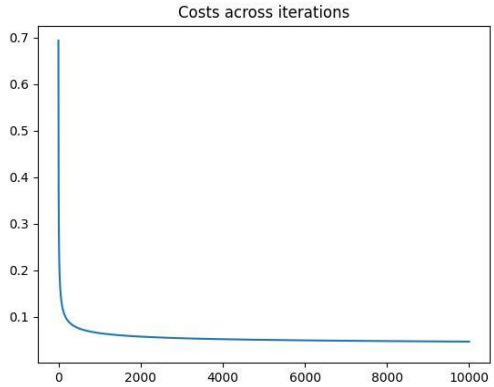
Przy losowym usuwaniu kolumn, nie wiadomo czy usunięta kolumna miała duży, czy mały wpływ na wynik. Dlatego wynik jest trochę gorszy niż dla wszystkich atrybutów.

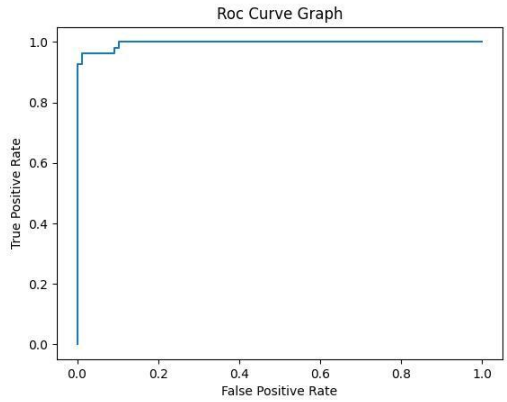
Wpływ usuwania kolumn (z niską korelacją)

- Usunięte zostały kolumny z |korelacja| < 0.25

	accuracy	F1	auroc
0	0.979021	0.972477	0.976136





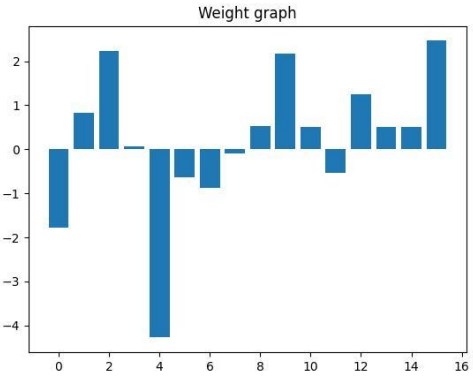


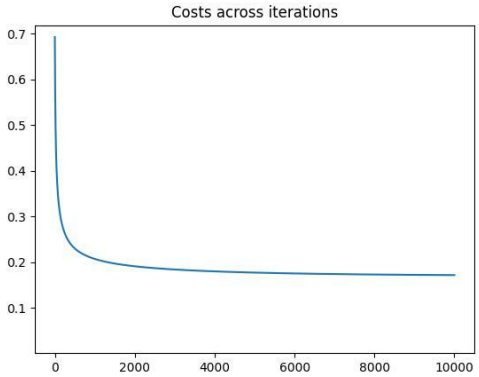
Dzięki usunięciu atrybutów o małej korelacji, wynik się trochę poprawił. Atrybuty, które miały znikomy wpływ na wynik już nie są brane pod uwagę

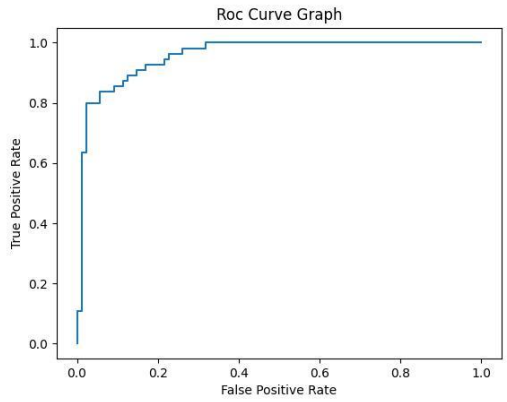
Wpływ usuwania kolumn (z wysoką korelacją)

- Usunięte zostały kolumny z |korelacja| > 0.5

	accuracy	F1	auroc
0	0.888112	0.846154	0.871591







Tutaj mamy sytuację odwrotną. Usuwając kolumny o dużej korelacji, model ma sporo gorszą celność. To znaczy, że usunięte atrybuty miały duży wpływ na jakość modelu.