Katedra Informatyki
Wydział Informatyki i Telekomunikacji
Politechnika Krakowska

Wykorzystanie biblioteki DEAP w problemie optymalizacji parametrów klasyfikatorów oraz selekcji cech

Jakub Zygmunt
Błażej zieliński

Kraków 2022

Spis treści

Przedstawienie zbioru danych

Zbiór opisuje ryzyko kredytów zaciąganych przez wybranych obywateli południowych niemiec w latach 1973 - 1975. Baza danych zawiera 1000 rekordów, z których każdy podzielony jest na 20 kolumn

status konta

czas trwania kredytu

historia kredytowa

przeznaczenie

wielkość kredytu

oszczędności

okres zatrudnienia

wielkość raty

status związku

współkredytobiorcy

czas zameldowania

własności osobiste

inne planowane raty

mieszkanie

ilość kredytóœ

praca

podopieczni

telefon

pracownik obcokrajowy

ryzyko kredytowe

Przedstawienie zbioru danych

	status	duration	credit_history	purpose	amount	savings	employment_duration	installment_rate	personal_status_sex	other_debtors	present_reside...	property	age	other_installment_plans	housing	number_credits	job	people_liable	telephone	foreign_worker	credit_risk
1	1	18	4	2	1049	1	2	4	2	1	4	2	21	3	1	1	3	2	1	2	1
2	1	9	4	0	2799	1	3	2	3	1	2	1	36	3	1	2	3	1	1	2	1
3	2	12	2	9	841	2	4	2	2	1	4	1	23	3	1	1	2	2	1	2	1
4	1	12	4	0	2122	1	3	3	3	1	2	1	39	3	1	2	2	1	1	1	1
5	1	12	4	0	2171	1	3	4	3	1	4	2	38	1	2	2	2	2	1	1	1
6	1	10	4	0	2241	1	2	1	3	1	3	1	48	3	1	2	2	1	1	1	1

Zestawienie klasyfikatorów dla parametrów domyślnych

SVC	Logistic regression	Decision Tree	K-nearest neighbors	Random forest
74,9%	74,2%	65,8%	70,7%	73,3%

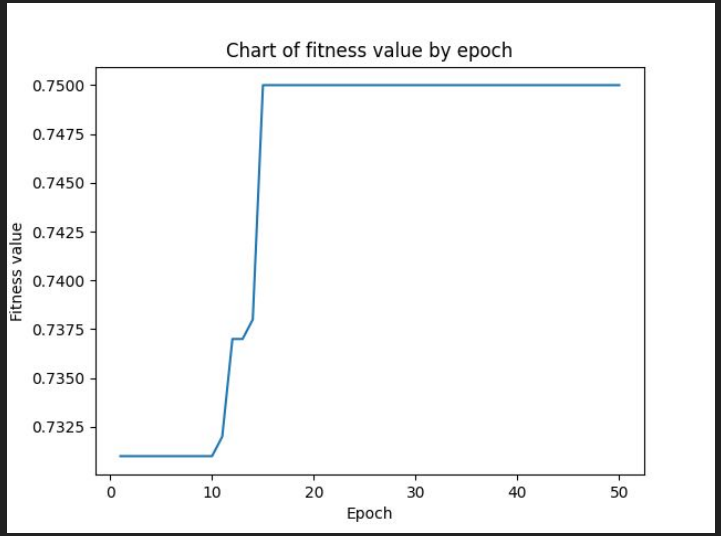
W przypadku Decision tree oraz Random forest wyniki są bardzo mało powtarzalne w porównaniu z resztą klasyfikatorów.

Przedstawienie wyników dla genetycznej optymalizacji parametrów

Klasyfikator maszyny wektorów pomocniczych (SVC)

Optymalizowane parametry:

- kernel: Określa typ jądra, który ma być użyty w algorytmie
  - wartości: ‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’
- C: Parametr regularyzacji
  - wartości: 0.1-5
- degree: Stopień wielomianu, używany dla typu jądra “poly”
  - wartości: 0.1-5
- gamma: Współczynnik kernela dla ‘rbf’, ‘poly’ i ‘sigmoid’.
  - wartości: 0.001-2
- coefficient: Współczynnik niezależny dla różnych funkcji jądra. Ma znaczenie tylko w „poly” i „sigmoid”.
  - wartości: 0.01-1

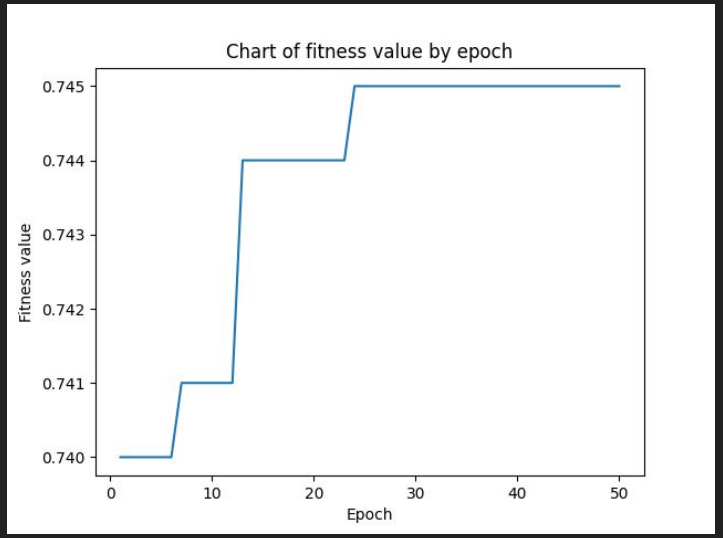


Najlepszy osobnik, rezultat 75%
"['rbf', 4.425212919619 68, 3.6830481801353385, 0.089503005175547 14, 0.006342219191 96]"

Klasyfikator regresji logistycznej

Optymalizowane parametry:

- solver: algorytm używany do rozwiązania problemu
  - wartości: ‘newton-cg’, ‘lbfgs’, ‘liblinear’, ‘sag’, ‘saga’
- C: Parametr regularyzacji
  - wartości: 0.1-5
- fit_intercept: Określa, czy do funkcji decyzyjnej należy dodać stałą (tzw. stronniczość lub przecięcie).
  - wartości: 0-1
- max_iter: maksymalna ilość iteracji algorytmu
  - wartości: 100-1000

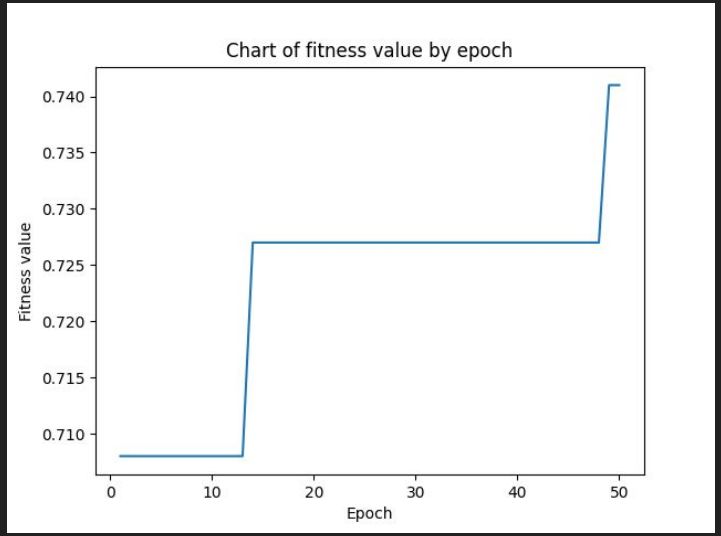


Najlepszy osobnik, rezultat 74,5%
"['liblinear', 0.5675191825122563, 0, 614]"

Klasyfikator drzewa decyzyjnego

Optymalizowane parametry:

- criterion: Funkcja pomiaru jakości podziału
  - wartości: “gini”, “entropy”, “log_loss”
- splitter: Strategia używana przy wyborze podziału każdego węzła
  - wartości: “best”, “random”
- max_depth: Maksymalna głębokość drzewa
  - wartości: 2-8
- min_samples_split: Minimalna liczba próbek wymagana do podziału węzła wewnętrznego:
  - wartości: 0.01-1
- min_samples_leaf: Minimalna liczba próbek, które muszą znajdować się w węźle liścia
  - wartości: 0.01-5

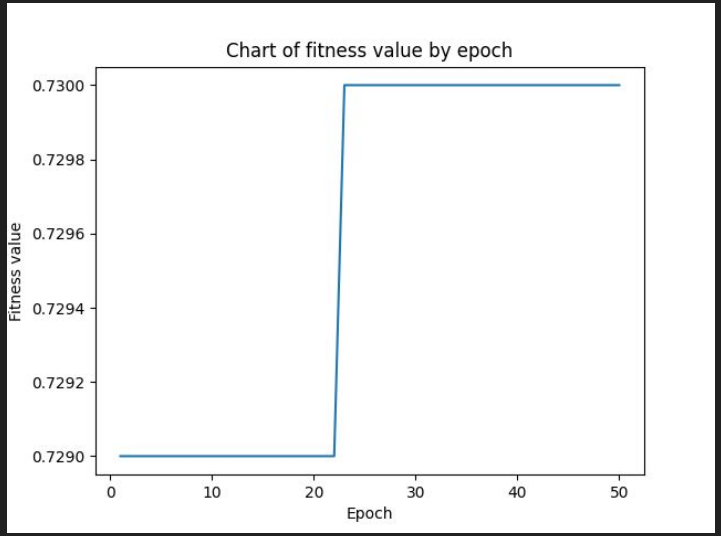


Najlepszy osobnik, rezultat 74,1%
"['log_loss', 'random', 5, 0.14448771756080656, 0.03612751616188254]"

Klasyfikator k - najbliższych sąsiadów

Optymalizowane parametry:

- n_neighbors: Liczba sąsiadów używanych domyślnie w zapytaniach sąsiadów
  - wartości: 1-10
- weights: Funkcja wagi używana w prognozie
  - wartości: ‘uniform’, ‘distance’
- algorithm: Algorytm używany do znajdywania najbliższego sąsiada
  - wartości: ‘auto’, ‘ball_tree’, ‘kd_tree’, ‘brute’
- leaf_size: Rozmiar liścia przekazany do ‘ball_tree’ i ‘kd_tree’
  - wartości: 20-40
- p: Parametr mocy dla metryki Minkowskiego
  - wartości: 2-5

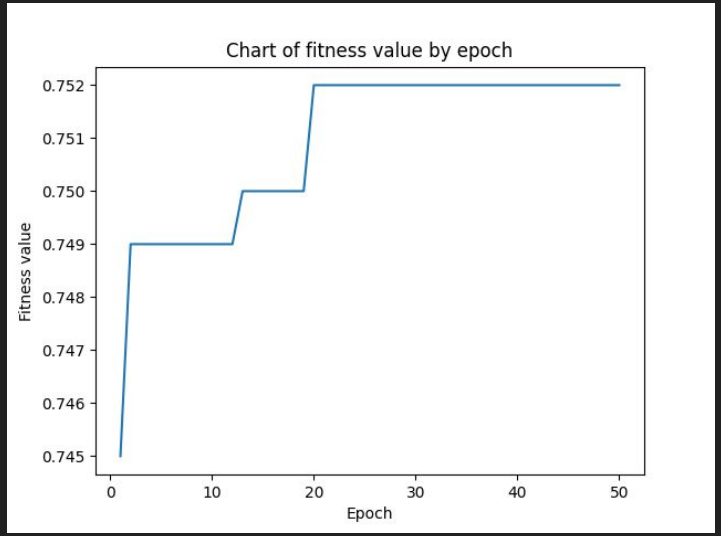


Najlepszy osobnik, rezultat 73%
"[10, 'distance', 'brute', 35, 5]”

Klasyfikator random forest

Optymalizowane parametry:

- criterion: Funkcja pomiaru jakości podziału
  - wartości: “gini”, “entropy”, “log_loss”
- n_estimators: Liczba drzew
  - wartości: 10-50
- max_depth: Maksymalna głębokość drzewa
  - wartości: 2-8
- min_samples_split: Minimalna liczba próbek wymagana do podziału węzła wewnętrznego:
  - wartości: 2-4
- min_samples_leaf: Minimalna liczba próbek, które muszą znajdować się w węźle liścia
  - wartości: 1-4



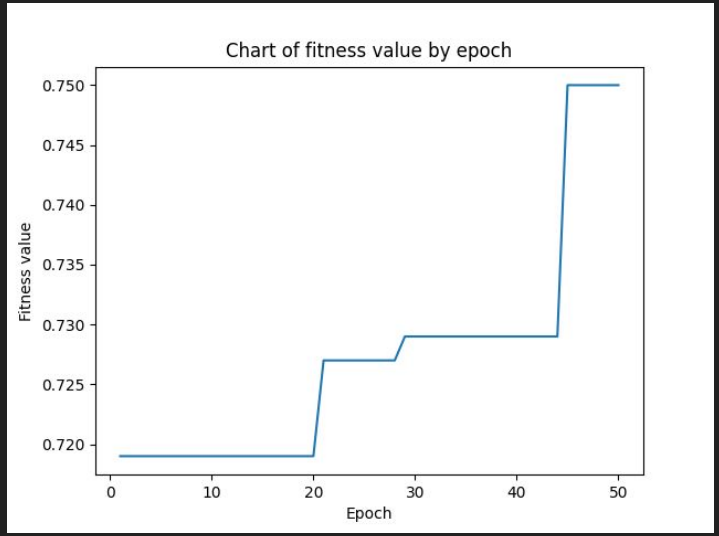
Najlepszy osobnik, rezultat 75,2%
"[45, 'log_loss', 7, 3, 1]"

Przedstawienie wyników dla genetycznej optymalizacji parametrów i selekcji cech

Klasyfikator maszyny wektorów pomocniczych (SVC)

Najlepszy osobnik, rezultat 75%

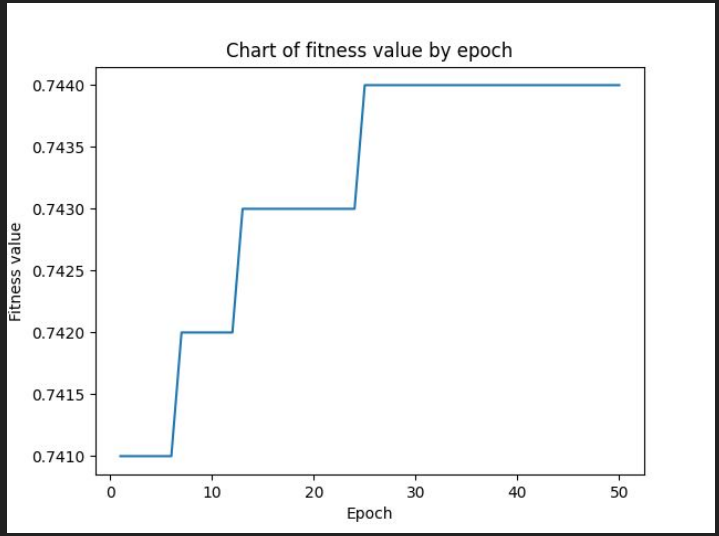
"['rbf', 0.7579598444849946, 4.2259208400124235,
1.0127524587837429, 0.6210313713651605,
1, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1]"



Klasyfikator regresji logistycznej

Najlepszy osobnik, rezultat 74.4%

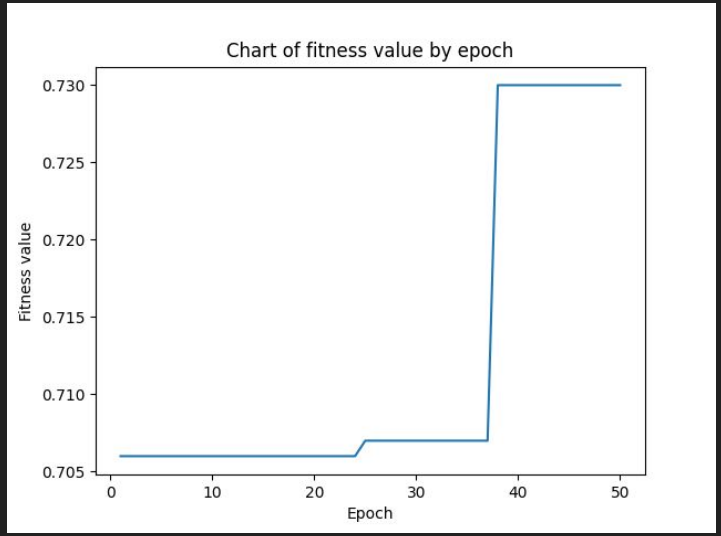
"['lbfgs', 1.0611177611481304, 0, 806,
0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1]"



Klasyfikator drzewa decyzyjnego

Najlepszy osobnik, rezultat 73%

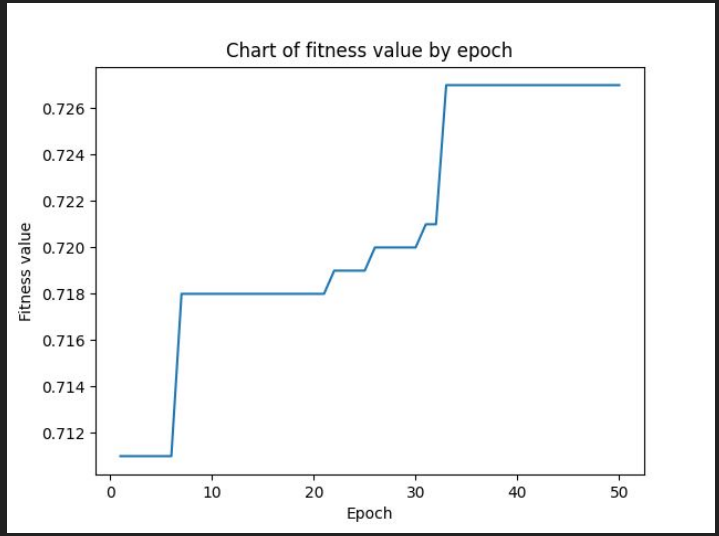
"['log_loss', 'best', 4, 0.114696833534790 82,
0.0545881074345763 44,
1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1]"



Klasyfikator k - najbliższych sąsiadów

Najlepszy osobnik, rezultat 72,7%

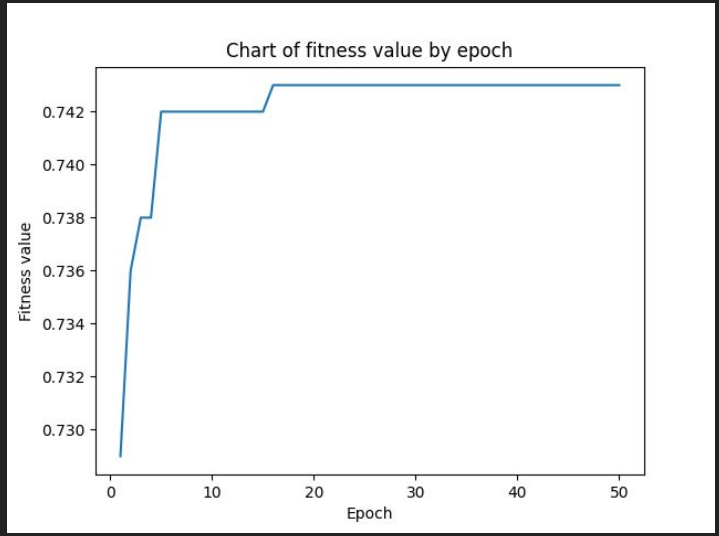
"[9, 'distance', 'ball_tree', 30, 5,
 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 1]"



Klasyfikator random forest

Najlepszy osobnik, rezultat 74,3%

"[15, 'gini', 6, 3, 2,
 0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 0, 1]"



Podsumowanie

Porównanie wyników klasyfikatorów

		Dokładność początkowa	Czas algorytmu	Wielkość populacji	Dokładność po algorytmie genetycznym
SVC	Bez Selekcji	74,9%	82,13s	10	75%
	Selekcja		79,12s		75%
LR	Bez Selekcji	74,2%	54.87s	50	74,5%
	Selekcja		57.77s		74.4%
DT	Bez Selekcji	65,8%	31.34s	50	74,1%
	Selekcja		32.38s		73%
KNN	Bez Selekcji	70,7%	46.77s	10	73%
	Selekcja		41.02s		72,7%
RF	Bez Selekcji	73,3%	41.17s	10	75,2%
	Selekcja		45.81s		74,3%

Wnioski

W przypadku analizowanego zbioru selekcja cech nie miała większego wpływu na otrzymywane wyniki.

Klasyfikatory takie jak SVC, Random forest oraz K-nearest neighbors charakteryzują się znacznie dłuższym (około 5 razy) czasem wykonania klasyfikacji aniżeli w przypadku Logistic regression oraz Decision tree.

Najlepszy wynik po optymalizacji parametrów uzyskał Random forest, bardzo dobrze poradził sobie również SVC oraz Logistic regression.

Optymalizacja parametrów w omawianym zbiorze przynosi znaczące rezultaty dla klasyfikatorów Decision tree, K-nearest neighbors oraz Random forest. w przypadku SVC oraz Logistic regression wyniki poprawiły się o zaledwie 0.1-0.