AutoML - projekt 2

Stanisław Kurzątkowski, Jan Kwiecień, Filip Mieszkowski

29 stycznia 2026

Krok 1: Generowanie Kandydatów

Metodologia eksperymentu:

- **Zakres:** 6 zbiorów danych $\times$ 5 typów modeli (LogReg, SVC, SGD, RF, ExtraTrees).
- **Algorytm:** `RandomizedSearchCV`.

Wynik etapu

Zbiór surowych wyników (plików CSV) dla każdej kombinacji zbiór-model.

Krok 2: Filtracja Jakościowa (Delta)

Selekcja najlepszych kandydatów w obrębie danego zbioru danych:

1. Znalezienie modelu o najwyższym *mean balanced accuracy*.
2. **Reguła Delta ($\delta = 0.02$):**
   - Akceptujemy modele gorsze od Lidera o max 2 p.p.
   - $Score_{model} \geq Score_{best} - 0.02$
3. Jeśli delta jest zbyt restrykcyjna, pobieramy 3 najlepsze konfiguracje.

Krok 3: Dywersyfikacja (Kubełkowanie)

Aby uniknąć duplikatów, grupujemy modele w „kubełki” i wybieramy po 1 reprezentancie z każdego:

Typ Modelu	**Kryteria Kubełków**
Drzewa (RF, ET)	Głębość: < 10, $10 - 20$, > 20 min_samples_leaf: 1, > 1
Liniowe (LR, SVC)	Parametr C: < 0.1, $0.1 - 1.0$, > 1.0 Wagi klas: *balanced* / *none*
SGD	Alpha: $< 10^{-5}$, $< 10^{-4}$, $> 10^{-3}$ Penalty: *l1, l2, elasticnet*

Limit: Maksymalnie 5 modeli na jeden zbiór danych.

Krok 4: Budowa Finalnego Portfolio

Synteza wyników ze wszystkich zbiorów do jednej listy (np. 50 modeli).

Strategia selekcji:

1. **Uniwersalność:** Priorytet dla modeli skutecznych na wielu zbiorach danych jednocześnie.
2. **Gwarancja Typu:** Wybór po 6 najlepszych modeli z każdej rodziny (łącznie 30 miejsc).
3. **Dopełnianie (Round-Robin):**
   - Pozostałe miejsca (15) zapełniamy „karuzelowo".
   - Kolejno: najlepszy LogReg $\rightarrow$ najlepszy SVM $\rightarrow$ najlepszy RF...

Fingerprint zbioru danych

Proste statystyki:

- **liczba próbek** (logarytm dziesiętny)
- **liczba cech** (logarytm dziesiętny)
- **liczba cech kategorycznych** (logarytm dziesiętny)
- **odsetek brakujących obserwacji**

Landmarkery:

- **drzewo decyzyjne** z jednym węzłem
- **naiwny bayes**
- **KNN** z $k = 1$

Mary informacyjne:

- **średnia informacja wzajemna**
- **entropia kolumny celu**
- **pca**95 - część wszystkich cech wyjaśniająca 95% wariancji

Krok 1: Meta-Learning i Selekcja Modeli

Miara Podobieństwa Zbiorów

System porównuje wektory cech (fingerprinty) za pomocą **odległości euklidesowej** po uprzedniej normalizacji (Z-score).
Waga historycznego zbioru jest odwrotnością tej odległości ($w \propto \frac{1}{d}$).

Tworzenie Shortlisty (12 kandydatów):

- Wybieramy modele, które historycznie osiągały najlepsze wyniki na sąsiadujących zbiorach.
- Weryfikacja: Modele te nie są przyjmowane "na wiarę"– są one **faktycznie trenowane** i oceniane na nowym zbiorze danych.

Krok 2: Architektura Zespołu (Ensemble)

Baza zespołu to **5 najlepszych modeli** wyłonionych z shortlisty.

Kandydaci do Stackingu (Meta-Learners)

- **Regresja Logistyczna** ($C \in \{0.1, 1.0\}$): Klasyczny standard. Silna regularyzacja neutralizuje silną korelację modeli bazowych.
- **Random Forest** (`max_depth=3`): Wychwytuje nieliniowe relacje w miejscach, gdzie konkretne modele się mylą.
- **MLPClassifier** (10 neuronów): Lekka, jednowarstwowa sieć neuronowa.

Ostateczna Decyzja Systemu

Na zbiorze walidacyjnym wyłaniany jest absolutny zwycięzca spośród 3 strategii: **1. Najlepszy pojedynczy model** | **2. Uśrednianie (Average)** | **3. Stacker**

Dziękujemy za uwagę

Pytania?

Jan Kwiecień, Filip Mieszkowski, Stanisław Kurzątkowski