
pip install auto-prep

Wydział Matematyki
i Nauk Informacyjnych
POLITECHNIKA WARSZAWSKA

AutoPrep

pakiet do automatycznego preprocessingu

Julia Kruk, Paweł Pozorski, Katarzyna Rogalska, Gaspar Sekula

Automatyczne Uczenie Maszynowe 2024/2025

Agenda

- Dla kogo?
- Inne rozwiązania
- Dla jakich danych?
- Jak działa?
- Etapy preprocessingu
- Wykorzystane modele
- Jak korzystać?
- Raport

Dla kogo?

- Programistów Pythona szukających najlepszych metod preprocessingowych
- Użytkowników analizujących każdy krok ML bez manualnego wykonywania
- Programistów chcących korzystać z automatycznych rozwiązań na co dzień
- Osób zainteresowanych poszerzaniem wiedzy o technikach preprocessingu
- Badaczy sprawdzających wpływ preprocessingu na zadania ML
- Programistów ceniących tradycyjne raporty papierowe

Inne rozwiązania

- Auto-sklearn
- TPOT
- H2O.ai
- PyCaret
- MLJAR
- Hyperopt-Sklearn
- Google AutoML Tables

Do jakich danych?

tabelarycznych

regresja i klasyfikacja

Jak działa?

1. **Wykrycie problemu** (klasyfikacja binarna / wieloklasowa / regresja)
2. Stworzenie **pipeline'ów** do preprocessingu
3. **Scorowanie** pipeline'ów prostymi modelami i wybranie najlepszych
4. **Trenowanie modeli, tunowanie hiperparametrów** i wybranie najlepszych
5. Tworzenie **raportu**

Etapy preprocessingu

Obowiązkowe

1. Uzupełnianie braków danych
2. Usuwanie kolumn z samymi unikalnymi wartościami
3. Kodowanie kategorycznych kolumn
4. Skalowanie
5. Usuwanie kolumn z zerową wariancją
6. Usuwanie skorelowanych kolumn

Dodatkowe

7. Selekcja zmiennych
8. Redukcja wymiarów

Wykorzystane modele

Klasyfikacja

- K Neighbors Classifier,
- Logistic Regression,
- Gaussian Naive Bayes,
- Support Vector Machine (Classifier),
- Decision Tree Classifier.

Regresja

- Linear Support Vector Machine (Regressor),
- K Neighbors Regressor,
- Random Forest Regressor,
- Bayesian Ridge,
- Gradient Boosting Regressor,
- Linear Regression.

Jak korzystać?

```python
import logging
from auto_prep.utils import config

config.update(log_level=logging.DEBUG)

import numpy as np

from auto_prep.prep import AutoPrep
from sklearn.datasets import fetch_openml

# Load your dataset
data = fetch_openml(name="titanic", version=1, as_frame=True, parser="auto").frame
data["survived"] = data["survived"].astype(np.uint8)

# Create and run pipeline
pipeline = AutoPrep()

if __name__ == "__main__":
    pipeline.run(data, target_column="survived")
```

Co zawiera raport?

- **Wstęp**: informacje systemowe, informacje o ramce danych (target, braki danych, opis predyktorów)
- **EDA**: rozkład zmiennej celu i predyktorów, macierz korelacji, wykresy skrzynkowe
- **Preprocessing**: opis pipeline'ów, wybór najlepszych z nich, szczegółowy opis najlepszych, statystyki procesu preprocessingu
- **Modele**: lista badanych modeli wraz z siatkami parametrów, najlepsze modele z wartościami metryk
- **Wyjaśnialność**: wykresy Shapleya

Wydział Matematyki
i Nauk Informacyjnych
POLITECHNIKA WARSZAWSKA

AutoPrep

pakiet do automatycznego preprocessingu

Julia Kruk, Paweł Pozorski, Katarzyna Rogalska, Gaspar Sekula

Automatyczne Uczenie Maszynowe 2024/2025