|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Magda Jaźwińska | |  | 2024 / 2025 | Programowanie systemów inteligentnych | | | |  | |
|  | **AUTOR** | **ROK AKAD.** |  | | | **SPEC.** | 2 | **GRUPA** |
| Wykorzystanie regresji liniowej i regularyzacji w przewidywaniu cen nieruchomości | | | | |  | | | 2 | |
| **TEMAT** | | | **NR SPRAWOZDANIA** | |
| 18.10.2024 | Laboratorium maszynowej analizy danych | | | | | dr. inż. Anna Czaderna-Lekka | | | |
| **DATA WYKONANIA** | **PRZEDMIOT** | | | | | **PROWADZĄCY** | | | |

# Cel

Celem analizy jest opracowanie modeli regresji do przewidywania cen nieruchomości w Bostonie na podstawie dostępnych cech. Analiza ma na celu ocenę skuteczności różnych podejść do modelowania, w tym regularyzacji oraz różnych metod przygotowania danych, takich jak standaryzacja i redukcja cech.

# Materiały i metody

W analizie wykorzystany został zbiór danych dotyczący cen nieruchomości w Bostonie, zawierający 14 cech, w tym wartość mediana nieruchomości (MEDV) i 13 cech opisujących nieruchomości. Zbiór danych podzielono na cztery zestawy:

1. **Zestaw nieprzetworzony** - zawiera wszystkie dostępne cechy w ich oryginalnej postaci.
2. **Zestaw zredukowany** - zawiera tylko cechy o istotnych korelacjach z wartością MEDV, co pozwala na uproszczenie modelu.
3. **Zestaw zestandaryzowany** - oryginalne cechy zostaną znormalizowane, co ułatwi porównania i poprawi stabilność algorytmu.
4. **Zestaw zredukowany zestandaryzowany** - zawiera tylko cechy o istotnych korelacjach z wartością MEDV w wersji zestandaryzowanej.

Do analizy danych zastosowane zostały techniki takie jak regresja liniowa, regresja grzbietowa, regresja lasso oraz regresja elastycznej siatki. Przed modelowaniem dane zostały przygotowane poprzez sprawdzenie ich kompletności, typów danych oraz identyfikację wartości odstających przy użyciu metody IQR. Dodatkowo, przed modelowaniem danych przeprowadzono ich standaryzację, aby znormalizować wartości cech.

# Wyniki i dyskusja

Wyniki analizy modeli regresji dla czterech zestawów danych przedstawiono w poniższej tabeli:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **MAE** | **MSE** | **RMSE** | **R²** |
| Regresja liniowa (nieprzetworzone) | 64.85 | 4963.70 | 70.45 | - 43.24 |
| Regresja grzbietowa | 4.29 | 39.52 | 6.29 | 0.65 |
| Regresja Lasso | 4.37 | 40.03 | 6.33 | 0.64 |
| Regresja elastycznej siatki | 4.44 | 41.06 | 6.41 | 0.63 |

Analizując wyniki, widoczna jest znaczna poprawa wskaźników dopasowania po zastosowaniu modeli regularyzacyjnych w porównaniu do regresji liniowej, która osiągnęła bardzo wysokie wartości błędów (MAE, MSE, RMSE) oraz negatywną wartość R².

Zastosowanie redukcji cech oraz regularyzacji (Ridge, Lasso, ElasticNet) doprowadziło do znacznie niższych wartości MAE, MSE i RMSE, co wskazuje na lepszą jakość predykcji. R² w modelach z regularyzacją wynosił ponad 0.63, co sugeruje, że modele te są w stanie wyjaśnić większą część wariancji w danych, w porównaniu do regresji liniowej, co jest korzystne w kontekście zastosowań praktycznych.

# Podsumowanie

W analizie cen nieruchomości w Bostonie zbadano cztery zestawy danych: nieprzetworzone, zredukowane, zestandaryzowane oraz zredukowane zestandaryzowane. Kluczowe wnioski obejmują:

1. **Wpływ Regularyzacji:** Regresja grzbietowa przyniosła najlepsze wyniki, wskazując na jej zdolność do radzenia sobie z nadmiernym dopasowaniem w porównaniu do regresji liniowej na nieprzetworzonych danych.
2. **Redukcja Cecha:** Uproszczenie modelu poprzez redukcję cech (RM, PTRATIO, LSTAT) nie tylko nie pogorszyło wyników, ale w niektórych przypadkach je poprawiło.
3. **Standaryzacja Danych:** Standaryzacja cech pozytywnie wpłynęła na wydajność modeli, eliminując problemy wynikające z różnic w skali.

Podsumowując, zastosowane techniki analizy danych i regularyzacji znacznie poprawiły dokładność prognoz w modelach regresyjnych.

# Bibliografia

1. Materiały z ćwiczeń laboratoryjnych.
2. Dokumentacja bibliotek Python: pandas, matplotlib, numpy, scikit-learn, seaborn.