|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Magda Jaźwińska | |  | 2024 / 2025 | Programowanie systemów inteligentnych | | | |  | |
|  | **AUTOR** | **ROK AKAD.** |  | | | **SPEC.** | 2 | **GRUPA** |
| Problem przewidywania wartości. Regresja liniowa | | | | |  | | | 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:

1. Wyniki dopasowania modeli:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **MAE** | **MSE** | **RMSE** | **R²** |
| Regresja liniowa | 4.00 | 33.27 | 5.77 | 0.70 |
| Regresja Ridge | 4.01 | 33.60 | 5.80 | 0.70 |
| Regresja Lasso | 4.14 | 35.36 | 5.95 | 0.68 |
| Regresja ElasticNet | 4.18 | 36.04 | 6.00 | 0.68 |

1. Wyniki dopasowania modeli dla zredukowanego zestawu danych oryginalnych:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **MAE** | **MSE** | **RMSE** | **R²** |
| Regresja liniowa | 4.29 | 39.52 | 6.29 | 0.65 |
| Regresja Ridge | 4.29 | 39.52 | 6.29 | 0.65 |
| Regresja Lasso | 4.37 | 40.03 | 6.33 | 0.64 |
| Regresja ElasticNet | 4.44 | 41.06 | 6.41 | 0.63 |

1. Wyniki dopasowania modeli dla danych zredukowanych i zestandaryzowanych:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **MAE** | **MSE** | **RMSE** | **R²** |
| Regresja liniowa | 4.29 | 39.52 | 6.29 | 0.6477 |
| Regresja Ridge | 4.29 | 39.52 | 6.29 | 0.6477 |
| Regresja Lasso | 4.36 | 40.26 | 6.35 | 0.6411 |
| Regresja ElasticNet | 4.44 | 41.43 | 6.44 | 0.6307 |

1. Wyniki dopasowania modeli dla zredukowanych wartości odstających dla MEDV:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **MAE** | **MSE** | **RMSE** | **R²** |
| Regresja liniowa | 4.29 | 39.52 | 6.28 | 0.6477 |
| Regresja Ridge | 4.29 | 39.52 | 6.28 | 0.6477 |
| Regresja Lasso | 4.36 | 40.28 | 6.34 | 0.6409 |
| Regresja ElasticNet | 4.45 | 41.64 | 6.45 | 0.6288 |

Analiza wyników dla czterech modeli regresji (liniowa, Ridge, Lasso, ElasticNet) wskazuje, że najlepsze dopasowanie uzyskano przy zastosowaniu regresji liniowej na pełnym zestawie danych. Wartość współczynnika determinacji **R²** wyniosła 0.70, co oznacza, że model ten w miarę dobrze przewiduje wartości na podstawie dostępnych cech. Z kolei modele z regularyzacją, takie jak Ridge, Lasso i ElasticNet, nie przyniosły znaczącej poprawy w stosunku do prostej regresji. W niektórych przypadkach ich wyniki były nawet minimalnie gorsze, co wskazuje, że regularyzacja nie była konieczna dla tego zestawu danych.

Po redukcji cech do trzech najważniejszych (RM, PTRATIO, LSTAT), obserwujemy pogorszenie dopasowania modeli. Wartości **R²** spadły do około 0.65, co oznacza, że te trzy cechy nie są wystarczające do pełnego przewidywania cen nieruchomości. Pomimo tego, że redukcja cech ułatwia interpretację modelu, traci się dokładność w przewidywaniu wyników.

Standaryzacja danych nie wpłynęła znacząco na jakość dopasowania, co sugeruje, że w tym przypadku standaryzacja nie była kluczowym krokiem w procesie modelowania. Wyniki dla zredukowanych i zestandaryzowanych danych były podobne do tych uzyskanych dla danych bez przekształceń, co potwierdza, że standardyzacja nie miała istotnego wpływu na ostateczne wyniki.

Podsumowując, najlepsze rezultaty osiągnięto stosując regresję liniową na pełnym zestawie danych, bez konieczności stosowania regularyzacji lub standaryzacji.

# 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 (Ridge) oraz liniowa uzyskały zbliżone wyniki, wskazując, że problem nadmiernego dopasowania nie był istotny w analizowanych danych. Modele Lasso i ElasticNet wprowadziły regularyzację, ale nie poprawiły wyników w stosunku do Ridge.
2. **Redukcja Cech**: Redukcja cech (RM, PTRATIO, LSTAT) nieznacznie pogorszyła dopasowanie modeli, szczególnie w zakresie metryki R². Niemniej jednak, modele nadal wykazywały przyzwoitą dokładność, co sugeruje, że uproszczenie modelu nie miało drastycznego wpływu na jego skuteczność.
3. **Standaryzacja Danych**: Standaryzacja cech miała niewielki wpływ na wyniki metryk oceny modeli. W niektórych przypadkach poprawiła stabilność modeli, jednak ogólne dopasowanie pozostawało zbliżone do wyników uzyskanych na danych niestandaryzowanych.

Podsumowując, zastosowane techniki analizy danych i regularyzacji miały umiarkowany wpływ na dokładność prognoz modeli regresyjnych, z największą poprawą widoczną po usunięciu wartości odstających.

# Bibliografia

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