1. Cel

Celem ćwiczeń było zastosowanie technik uczenia nadzorowanego do przewidywania wartości mediany cen nieruchomości w Bostonie (kolumna MEDV) na podstawie dostępnych cech opisujących dane socjo-ekonomiczne oraz demograficzne, w szczególności przy użyciu regresji liniowej. Zadanie obejmowało także wprowadzenie regularyzacji (regresji grzbietowej, Lasso oraz ElasticNet), aby poprawić efektywność modelu.

2. Materiały i metody

Dane użyte w zadaniu to zbiór opisujący ceny nieruchomości w Bostonie, składający się z 506 wierszy oraz 13 cech opisowych, m.in. CRIM (wskaźnik przestępczości), RM (średnia liczba pokoi) czy LSTAT (odsetek ubogich mieszkańców).
Cechą przewidywaną była MEDV, reprezentująca medianę wartości nieruchomości w tysiącach dolarów.

CRIM - współczynnik przestępczości w mieście,
ZN - odsetek "dużych działek" - powyżej 2500 m2,
INDUS - odsetek terenów industrialnych w mieście,
CHAS - jeśli teren znajduje się przy rzece Charles -1, w pozostałych przypadkach 0,
NOX - stężenie tlenków azotu,
RM - średnia ilość pomieszczeń w budynku,
AGE - odsetek "starych budynków" - powstałych przed 1940 r.,
DIS - ważona odległość od urzędów pracy w Bostonie,
RAD - wskaźnik dostępności do głównych dróg,
TAX - wartość podatku od nieruchomości liczona od 10 tys. dolarów,
PTRATIO - stosunek liczby uczniów na nauczycieli w mieście,
B - odsetek osób pochodzenia afroamerykańskiego,
LSTAT - odsetek mieszkańców zaliczany do ubogich (odsetek ubóstwa),
MEDV - mediana wartości domów z danego terenu (w tys. dolarów).

W analizie zastosowano:

Cztery zestawy danych:

Zestaw nieprzetworzony – oryginalne dane z pełną liczbą cech.
Zestaw zredukowany – wybrane cechy o wysokiej korelacji z celem (RM, PTRATIO, LSTAT).
Zestaw zestandaryzowany – dane poddane standaryzacji z wykorzystaniem StandardScaler.
Zestaw zredukowany i zestandaryzowany – wybrane cechy z zestawu 2 poddane standaryzacji.

Dla każdego z zestawów danych zbudowano cztery modele regresji:

Regresja liniowa (bez regularyzacji),
Regresja grzbietowa (Ridge),
Regresja metodą Lasso,
Regresja metodą elastycznej siatki (ElasticNet).
Modele trenowano i testowano na danych podzielonych na zbiór treningowy (80%) i testowy (20%).

Do ewaluacji modeli wykorzystano wskaźniki:

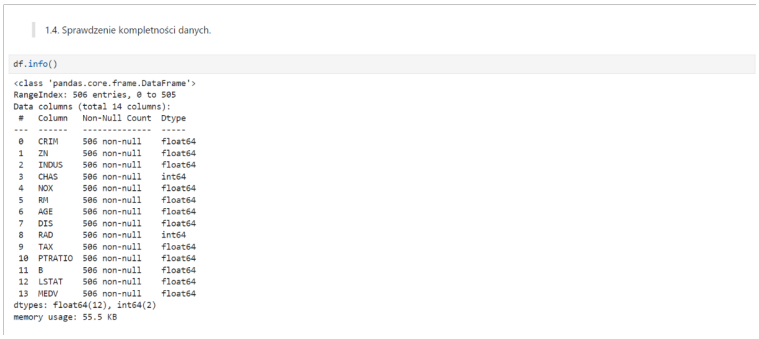
MAE (średni błąd bezwzględny),
MSE (błąd średniokwadratowy),
RMSE (pierwiastek błędu średniokwadratowego),
R^2 (współczynnik determinacji).

3. Wyniki i dyskusja

Zadanie 1: Regresja liniowa na pełnym zbiorze danych

Import danych:

Zaimportowano dane nieruchomości bostońskich i zweryfikowano ich strukturę. Dane były kompletne, składały się wyłącznie z wartości numerycznych.



1.4. Sprawdzenie kompletności danych.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 506 entries, 0 to 505
Data columns (total 14 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   CRIM     506 non-null    float64
 1   ZN       506 non-null    float64
 2   INDUS    506 non-null    float64
 3   CHAS     506 non-null    int64  
 4   NOX      506 non-null    float64
 5   RM       506 non-null    float64
 6   AGE      506 non-null    float64
 7   DIS      506 non-null    float64
 8   RAD      506 non-null    int64  
 9   TAX      506 non-null    float64
 10  PTRATIO  506 non-null    float64
 11  B        506 non-null    float64
 12  LSTAT    506 non-null    float64
 13  MEDV     506 non-null    float64
dtypes: float64(12), int64(2)
memory usage: 55.5 KB
```

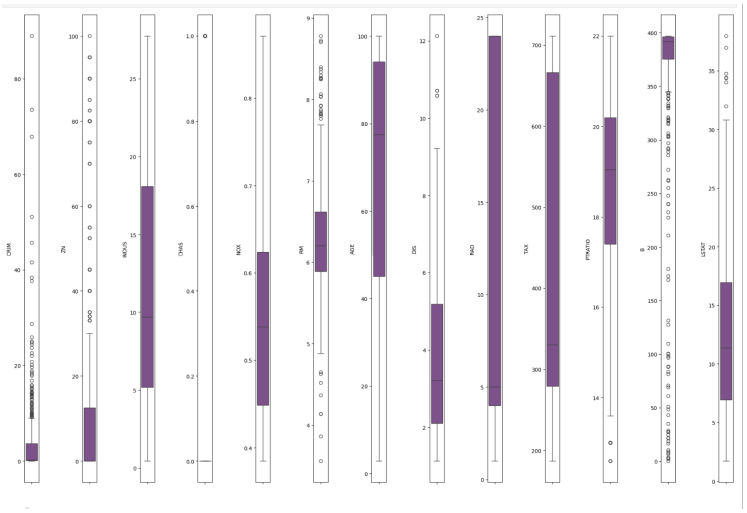
Podstawowa analiza danych:

a) *Przeanalizowano rozkład danych za pomocą statystyk opisowych oraz wykresów pudełkowych.*

1.3. Sprawdzenie podstawowych statystyk.

```
df.describe()
```

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000
mean	3.613524	11.363636	11.136779	0.069170	0.554695	6.284634	68.574901	3.795043	9.549407	408.237154	18.455534	356.674032	12.653063
std	8.601545	23.322453	6.860353	0.253994	0.115878	0.702617	28.148861	2.105710	8.707259	168.537116	2.164946	91.294864	7.141062
min	0.006320	0.000000	0.460000	0.000000	0.385000	3.561000	2.900000	1.129600	1.000000	187.000000	12.600000	0.320000	1.730000
25%	0.082045	0.000000	5.190000	0.000000	0.449000	5.885500	45.025000	2.100175	4.000000	279.000000	17.400000	375.377500	6.950000
50%	0.256510	0.000000	9.690000	0.000000	0.538000	6.208500	77.500000	3.207450	5.000000	330.000000	19.050000	391.440000	11.360000
75%	3.677083	12.500000	18.100000	0.000000	0.624000	6.623500	94.075000	5.188425	24.000000	666.000000	20.200000	396.225000	16.955000
max	88.976200	100.000000	27.740000	1.000000	0.871000	8.780000	100.000000	12.126500	24.000000	711.000000	22.000000	396.900000	37.970000

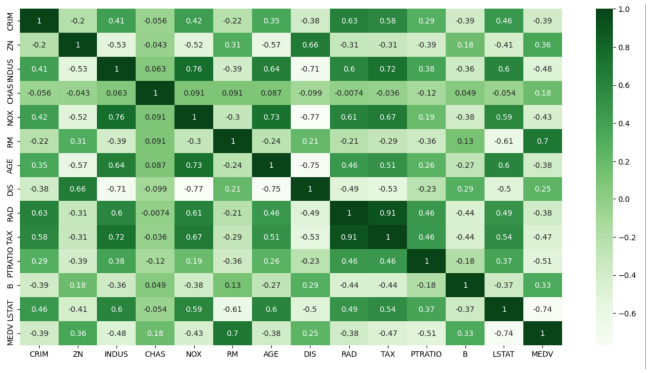


b) *Obliczono procent wartości odstających, gdzie najwyższe wskaźniki odstępstw odnotowano dla zmiennych B (15.217%) i CRIM (13.043%) oraz ZN (13.44%) - może wpływać na wyniki regresji .*

```
% wartości odstających dla każdej z kolumn:
CRIM    13.043%
ZN      13.439%
INDUS   0.0%
CHAS    6.917%
NOX     0.0%
RM      5.929%
AGE     0.0%
DIS     0.988%
RAD     0.0%
TAX     0.0%
PTRATIO 2.964%
B       15.217%
LSTAT   1.383%
MEDV    7.905%
```

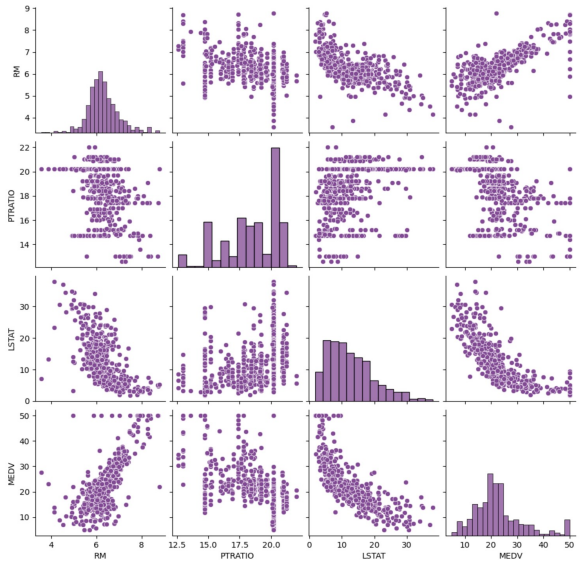
Macierz korelacji:

Analiza korelacji wykazała, że cechy
RM (średnia liczba pokoi) oraz LSTAT (odsetek ubogich mieszkańców)
są silnie skorelowane z cenami nieruchomości (odpowiednio 0.7 i -0.74).



Biorąc pod uwagę wybór najbardziej skorelowanych cech o współczynniku korelacji <- 0.5 lub >0.5 otrzymujemy zestaw cech:

RM, LSTAT oraz PTRATIO (stosunek liczby uczniów na nauczycieli w mieście), odpowiednio (0.7, -0.74 i -0.51)



Model regresji liniowej:

Przeprowadzono podział danych na zbiór treningowy (80%) i testowy (20%), a następnie dopasowano model regresji liniowej.
Zbudowano model regresji liniowej na pełnym zbiorze cech.
Współczynniki dopasowania wykazały, że kluczowe cechy wpływające na MEDV to:

- *RM (+4.018), co oznacza, że wzrost liczby pokoi o jeden podnosi cenę nieruchomości o ok. 4000 dolarów,*
- *NOX (-15.646) pokazuje, że wzrost stężenia tlenków azotu o jednostkę (np. z 0.5 do 1.5) powoduje spadek wartości nieruchomości o około 15 646 dolarów. Wysokie zanieczyszczenie powietrza jest zatem kluczowym czynnikiem obniżającym cenę domów.*

```
Współczynniki dopasowania:
         Coefficient
CRIM       -0.098991
ZN          0.042505
INDUS       0.016739
CHAS        3.064379
NOX       -15.646253
RM          4.018884
AGE        -0.000832
DIS        -1.446280
RAD         0.267827
TAX        -0.010473
PTRATIO    -0.888652
B           0.008253
LSTAT      -0.549367
```

Metryki ewaluacji modelu:

MAE: 4.00

MSE: 33.27

RMSE: 5.77

R^2: 0.7035

Wyniki sugerują, że model ma przyzwoitą moc predykcyjną, aczkolwiek pozostają pewne obszary poprawy (Model wyjaśnia 70% zmienności w danych, co stanowi przyzwoitą prognozę, jednak wysoki błąd RMSE wskazuje na istotne różnice między przewidywanymi a rzeczywistymi cenami nieruchomości.).

Regularyzacja:

Ridge (α = 0.5):

MAE: 4.01, MSE: 33.60, R^2: 0.7005

Lasso (α = 0.5):

MAE: 4.14, MSE: 35.36, R^2: 0.6848

ElasticNet (α = 0.5, l1_ratio = 0.5):

MAE: 4.18, MSE: 36.04, R^2: 0.6788

Regresje regularyzacyjne nieznacznie pogorszyły wyniki, co sugeruje, że dane nie cierpią znacząco z powodu przetrenowania, a regularyzacja może nie być niezbędna w tym przypadku.
Stworzenie uproszczonego modelu regresji liniowej, który uwzględnia tylko najbardziej skorelowane cechy z zadania 1.8–1.10, czyli RM (średnia liczba pomieszczeń), PTRATIO (stosunek liczby uczniów do nauczycieli) oraz LSTAT (odsetek ubogiej ludności):

> *Podobnie jak we wcześniejszych częściach zadania, dane podzielono na zestaw uczący (80%) i testowy (20%).*
>
> *Trening modelu regresji liniowej oraz trzech metod regularyzacyjnych (Ridge, Lasso, ElasticNet).*
>
> *Ocena modeli za pomocą miar: MAE, MSE, RMSE, R^2.*

Współczynniki dopasowania:

RM: 4.6935 (największy pozytywny wpływ na cenę).

PTRATIO: -0.8691 (negatywny wpływ).

LSTAT: -0.5656 (negatywny wpływ).

Wyniki ewaluacji:

MAE: 4.29

MSE: 39.52

RMSE: 6.29

R^2: 0.6477

Wyniki pokazują, że model oparty na trzech wybranych cechach osiągnął nieco niższą skuteczność niż model bazujący na wszystkich zmiennych (R^2 = 0.6477), co wskazuje, że pomimo

silnej korelacji wybranych cech, model z pełnym zestawem danych lepiej przewiduje wartość nieruchomości.

Regularyzacja:

Model	*MAE*	*MSE*	*RMSE*	*R²*
Ridge	*4.29*	*39.52*	*6.29*	*0.6477*
Lasso	*4.36*	*40.03*	*6.33*	*0.6432*
ElasticNet	*4.44*	*41.06*	*6.41*	*0.6340*

Wyniki dla wszystkich trzech metod regularyzacji są bardzo zbliżone do wyników modelu bazowego, co wskazuje na to, że model nie wykazuje dużej tendencji do nadmiernego dopasowania, ponieważ różnice w wynikach między metodami regularyzacyjnymi a standardową regresją są minimalne.
Ridge utrzymał nieco wyższą wartość R², co czyni go najlepszym rozwiązaniem regularyzacyjnym w tej analizie.

Zadanie 2: Standaryzacja i redukcja cech

Zestaw danych i cechy były takie same jak w zadaniu 1.

- *Dokonano standaryzacji danych przy użyciu metody StandardScaler.*
- *Przeprowadzono analizę regresji liniowej oraz regularyzacji metodami Ridge, Lasso i ElasticNet na zestawie danych przed i po standaryzacji.*
- *Przeprowadzono regresję na wybranych zeskalowanych cechach (RM, PTRATIO, LSTAT)*

Standaryzacja danych:

> *Standaryzacja została przeprowadzona na wszystkich kolumnach zbioru danych z wyjątkiem kolumny celu (MEDV, czyli ceny nieruchomości).*
> *Zmieniono rozkład cech w taki sposób, aby miały one średnią 0 oraz odchylenie standardowe 1,*
> *co pozwala na wyeliminowanie różnic w skali między poszczególnymi cechami, co mogłoby wpływać na wagi w modelu regresji.*
>
> *Cechy przed standaryzacją różniły się skalą, np. zmienna TAX przyjmowała wartości od 187 do 711, podczas gdy ZN wahało się od 0 do 100.*
>
> *Po standaryzacji każda cecha ma rozkład zbliżony do rozkładu standardowego normalnego.*

Budowa modelu regresji liniowej po standaryzacji:

Po standaryzacji danych, zbudowano model regresji liniowej.
Współczynniki regresji po standaryzacji zmieniły się, ponieważ teraz uwzględniają znormalizowaną wersję danych.

Współczynniki dopasowania dla niektórych cech po standaryzacji były następujące:

RM (średnia liczba pomieszczeń): 2.703861

LSTAT (odsetek ubóstwa): -3.755397

PTRATIO (stosunek liczby uczniów na nauczyciela): -1.907946

DIS (ważona odległość od urzędów pracy w Bostonie): -3.011843

Największy pozytywny wpływ na cenę nieruchomości miała liczba pomieszczeń (RM), natomiast największy negatywny wpływ miała ważona odległość od urzędów pracy w Bostonie (DIS) oraz wskaźnik ubóstwa (LSTAT).

Wyniki ewaluacji modelu po standaryzacji:

MAE: 4.00

MSE: 33.27

RMSE: 5.77

R^2: 0.7035

Standaryzacja nie wpłynęła na wyniki modelu w porównaniu do niestandaryzowanych danych z zadania 1, co może wskazywać, że różnice w skali cech nie miały znaczącego wpływu na proces uczenia się modelu.
Zatem, w tym konkretnym przypadku, standaryzacja nie wniosła istotnej poprawy do modelu, co może sugerować, że model liniowy jest stosunkowo odporny na różnice w skali cech w tym zbiorze danych.

Wyniki Regularyzacji po standaryzacji:

Regresja grzbietowa (Ridge):

MAE: 4.00

MSE: 33.28

RMSE: 5.77

R^2: 0.7033

Ridge osiągnął niemal identyczne wyniki co standardowa regresja liniowa, co sugeruje, że wprowadzenie niewielkiej regularyzacji nie miało dużego wpływu na wyniki.

Regresja Lasso:

MAE: 4.34

MSE: 39.09

RMSE: 6.25

R^2: 0.6516

Lasso wykazało gorsze wyniki niż Ridge, co sugeruje, że ta forma regularyzacji mogła zbyt mocno wpłynąć na model, eliminując część cech, co pogorszyło dopasowanie.
ElasticNet:

MAE: 4.19

MSE: 38.31

RMSE: 6.19

R^2: 0.6585

ElasticNet osiągnął wyniki zbliżone do Lasso, co potwierdza, że w tym przypadku lżejsza regularyzacja (jak Ridge) była bardziej odpowiednia.
ElasticNet łączy w sobie zalety obu metod, ale nie osiągnął lepszych wyników niż Ridge.

Regresja na wybranych cechach (RM, PTRATIO, LSTAT)

Ostateczna analiza dotyczyła wybranych trzech cech: liczba pokoi (RM), wskaźnik PTRATIO oraz odsetek ubóstwa (LSTAT).

Modele po zeskalowaniu tych trzech cech uzyskały następujące wyniki:

Regresja liniowa: MAE: 4.29, MSE: 39.52, RMSE: 6.29, R^2: 0.6477

Ridge: MAE: 4.29, MSE: 39.52, RMSE: 6.29, R^2: 0.6477

Lasso: MAE: 4.36, MSE: 40.26, RMSE: 6.35, R^2: 0.6411

ElasticNet: MAE: 4.44, MSE: 41.43, RMSE: 6.44, R^2: 0.6307

Wyniki dla ograniczonego zestawu cech były gorsze niż dla pełnych danych, ale zgodnie z oczekiwaniami, Ridge osiągnął najlepszą równowagę, minimalizując błąd i zachowując umiarkowane dopasowanie do danych.

4. Podsumowanie

1. *Regresja liniowa okazała się skuteczna w przewidywaniu cen nieruchomości w Bostonie, z R^2 wynoszącym 0.7035, co sugeruje, że model wyjaśnia około 70% wariancji cen.*
2. *Regularyzacja wprowadziła jedynie nieznaczne poprawy w zakresie ochrony modelu przed przetrenowaniem, lecz ogólnie wyniki były nieco gorsze niż przy prostej regresji liniowej.*
3. *Standaryzacja danych miała minimalny wpływ na wyniki modelu regresji liniowej, co wskazuje, że różnice w skali cech nie miały kluczowego wpływu na proces uczenia modelu w tym zbiorze danych.*

5. Bibliografia

https://scikit-learn.org/0.21/documentation.html

https://seaborn.pydata.org/

https://pandas.pydata.org/docs/

https://www.youtube.com/@Analitykdanych/videos