Boston House Prediction

by: CV B(Fei Fei Li)

Table of Content

Problem and Business Understanding

Problem

- Membantu client dalam memprediksikan harga rumah berdasarkan data yang sudah dikumpulkan (Boston Data)

Business

- Memberikan insight kepada client bahwa dari data yang dimiliki selain memberikan prediksi harga rumah namun juga memberikan masukan business terhadap client beberapa parameter(fitur) yang dapat meningkatkan pembelian atau penjualan rumah (properti)

Insight Data

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	NaN	36.2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
501	0.06263	0.0	11.93	0.0	0.573	6.593	69.1	2.4786	1	273	21.0	391.99	NaN	22.4
502	0.04527	0.0	11.93	0.0	0.573	6.120	76.7	2.2875	1	273	21.0	396.90	9.08	20.6
503	0.06076	0.0	11.93	0.0	0.573	6.976	91.0	2.1675	1	273	21.0	396.90	5.64	23.9
504	0.10959	0.0	11.93	0.0	0.573	6.794	89.3	2.3889	1	273	21.0	393.45	6.48	22.0
505	0.04741	0.0	11.93	0.0	0.573	6.030	NaN	2.5050	1	273	21.0	396.90	7.88	11.9

	Variabel	Deskripsi
0	CRIM	Tingkat kejahatan per kapita berdasarkan kota.
1	ZN	Proporsi lahan tempat tinggal yang dizona untu...
2	INDUS	Proporsi hektar non-ritel per kota.
3	CHAS	Variabel dummy Charles River (1 jika daerah te...
4	NOX	Konsentrasi oksida nitrat (bagian per 10 juta).
5	RM	Jumlah rata-rata kamar per hunian.
6	AGE	Proporsi unit yang ditempati pemilik yang diba...
7	DIS	Jarak tertimbang ke lima pusat ketenagakerjaan...
8	RAD	Indeks aksesibilitas ke jalan raya radial.
9	TAX	Tingkat pajak properti nilai penuh per $10.000.
10	PTRATIO	Rasio murid-guru berdasarkan kota.
11	B	1000(Bk - 0.63)^2 di mana Bk adalah proporsi p...
12	LSTAT	% status sosial ekonomi rendah dari populasi.
13	MEDV	Nilai rata-rata rumah yang ditempati pemilik d...

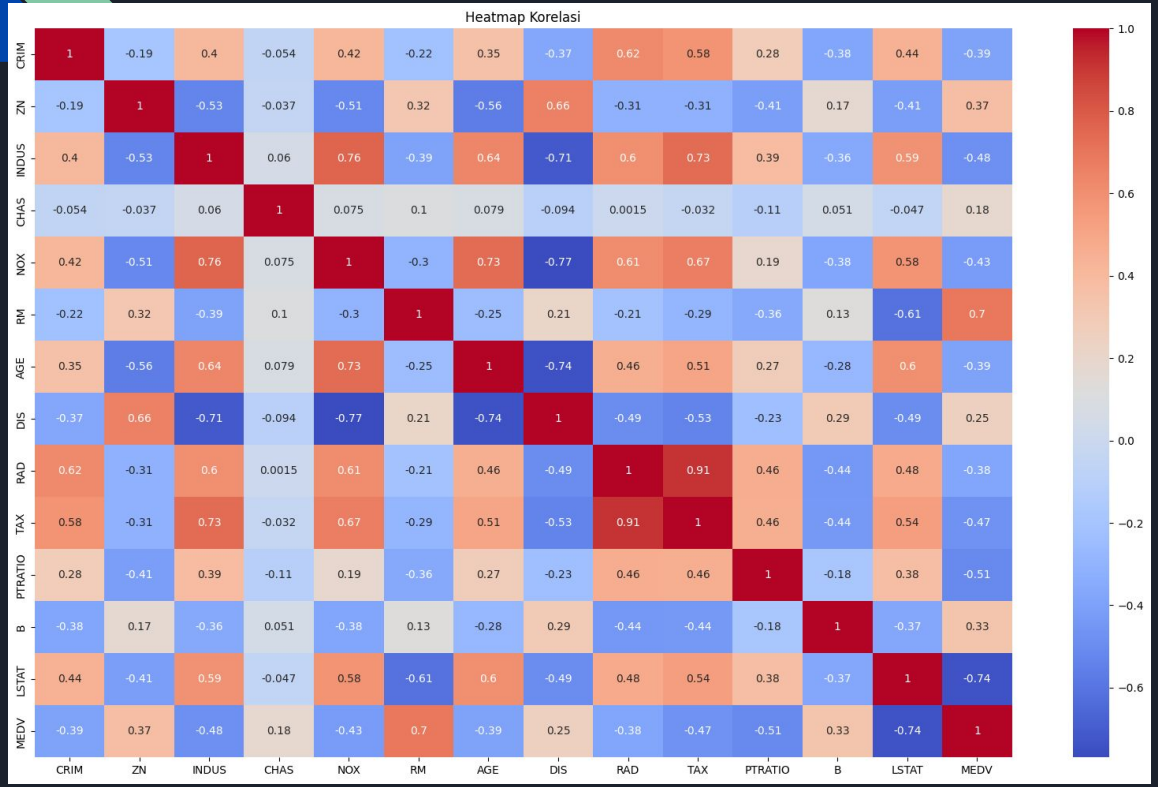
- Dataset terdiri dat 506 baris dan 14 kolom.
- Sebagian besar kolom memiliki tipe data numerik (float64) dan 2 kolom (CHAS, RAD) bertipe integer (int64).
- Terdapat missing values di beberapa kolom: CRIM, ZN, INDUS, CHAS, AGE, dan LSTAT.
- 1 data target (MEDV) dan 13 data fitur.

Insight Data

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
count	486.000000	486.000000	486.000000	486.000000	506.000000	506.000000	486.000000	506.000000	506.000000	506.000000	506.000000	506.000000	486.000000	506.000000
mean	3.611874	11.211934	11.083992	0.069959	0.554695	6.284634	68.518519	3.795043	9.549407	408.237154	18.455534	356.674032	12.715432	22.532806
std	8.720192	23.388876	6.835896	0.255340	0.115878	0.702617	27.999513	2.105710	8.707259	168.537116	2.164946	91.294864	7.155871	9.197104
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	5.000000
25%	0.081900	0.000000	5.190000	0.000000	0.449000	5.885500	45.175000	2.100175	4.000000	279.000000	17.400000	375.377500	7.125000	17.025000
50%	0.253715	0.000000	9.690000	0.000000	0.538000	6.208500	76.800000	3.207450	5.000000	330.000000	19.050000	391.440000	11.430000	21.200000
75%	3.560263	12.500000	18.100000	0.000000	0.624000	6.623500	93.975000	5.188425	24.000000	666.000000	20.200000	396.225000	16.955000	25.000000
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	50.000000

- Distribusi data beberapa fitur seperti CRIM, ZN, TAX, LSTAT cenderung *skewed* (condong), sedangkan RM, DIS, PTRATIO relatif simetris.
- Beberapa fiitur seperti CRIM, ZN, dll memiliki variasi tinggi, menunjukkan sebaran data yang besar dan potensi nilai ekstrem.
- Terdapat outlier pada fitur CRIM, ZN, RM, B, LSTAT, MEDV, ditunjukkan oleh jarak jauh antara nilai maksimum dan kuartil atas.

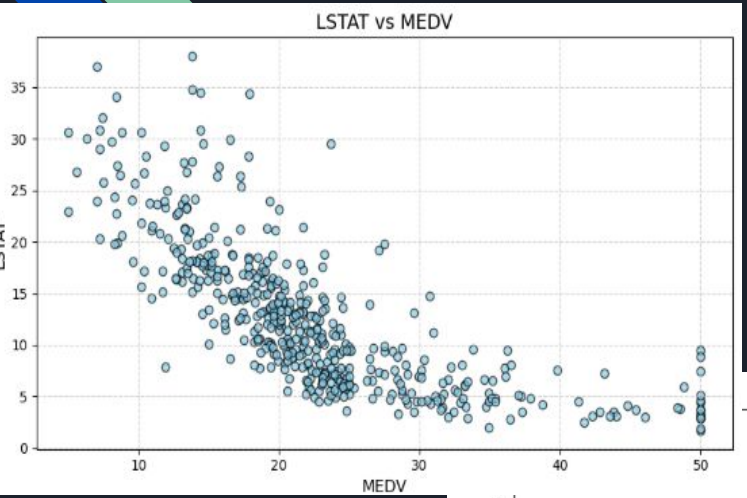
Insight Data



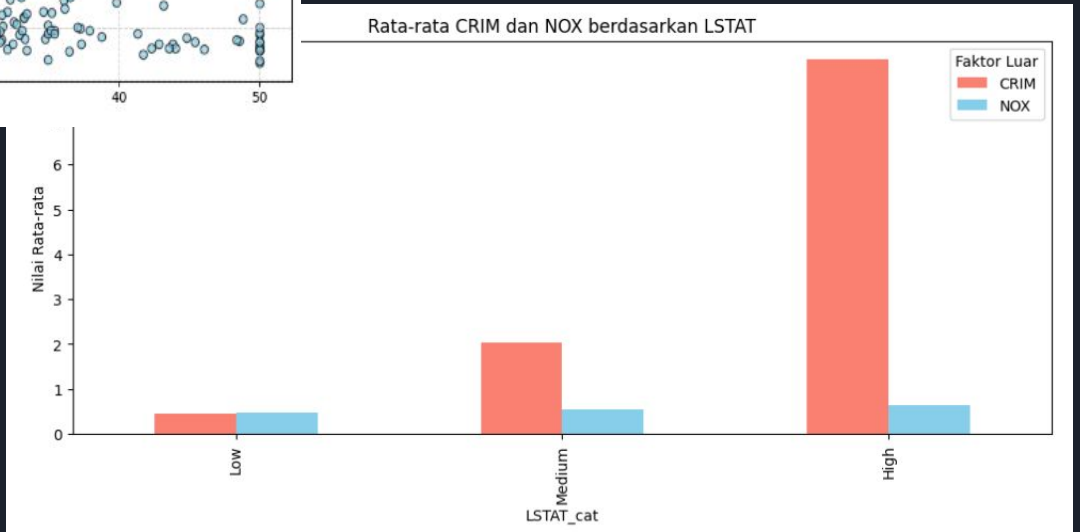
2 korelasi tinggi: 1 korelasi positif pada RM dan 1 korelasi negatif LSTAT.

- Semakin padat penduduk harga rumah di Boston lebih moderat ke rendah.
- Semakin banyak kamar harga rumah semakin tinggi.

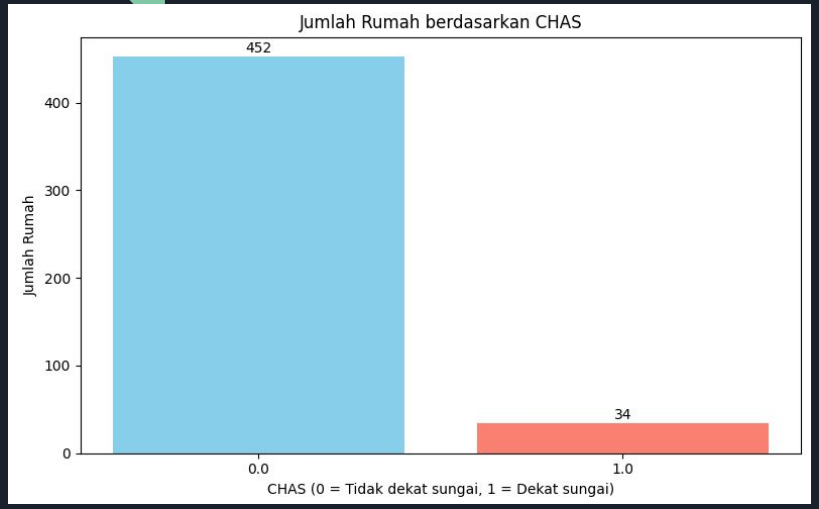
Insight Data

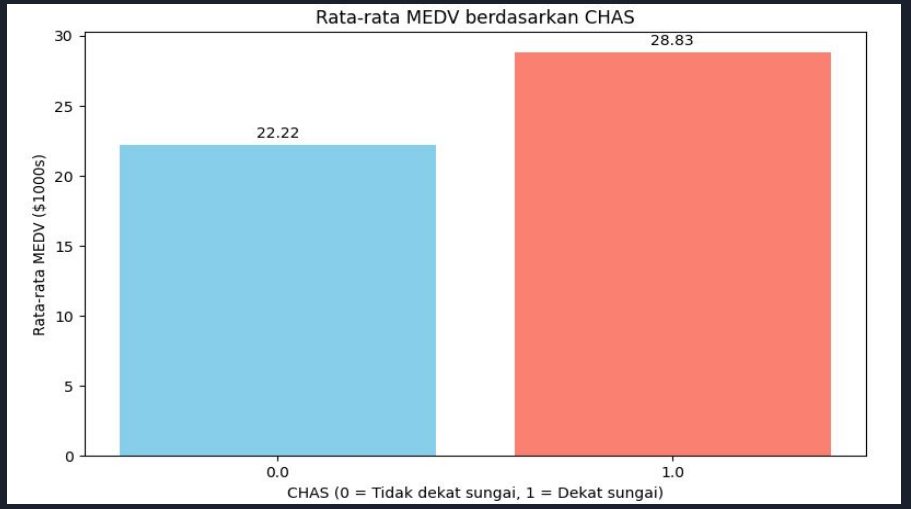


Di Boston rumah akan memiliki selling price cukup tinggi jika berada di populasi masyarakat yang tidak terlalu padat. Selain itu dengan padat penduduk memberikan gambaran semakin tinggi tinkat polutan dan kriminal ratenya.



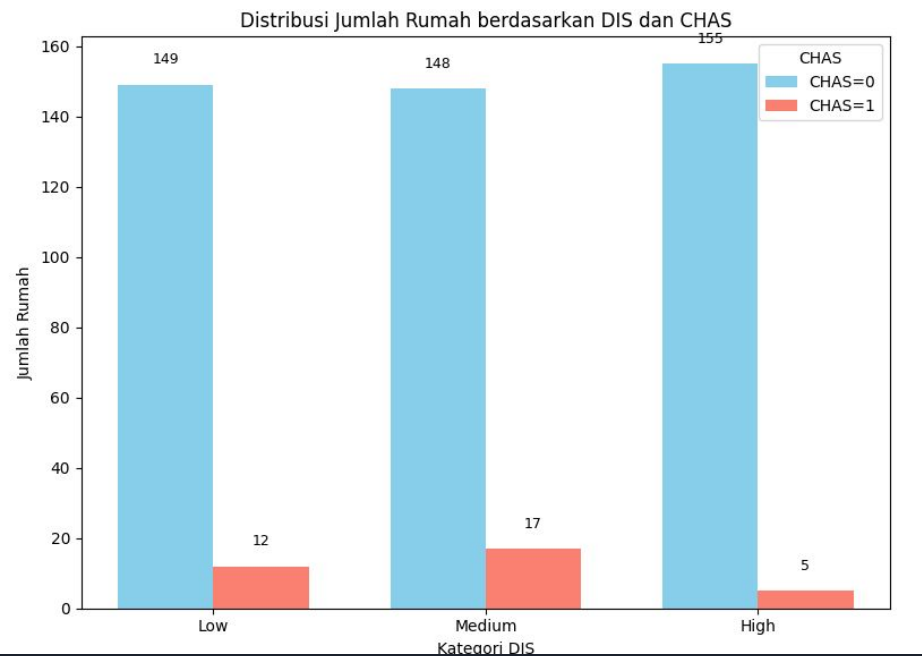
Insight Data





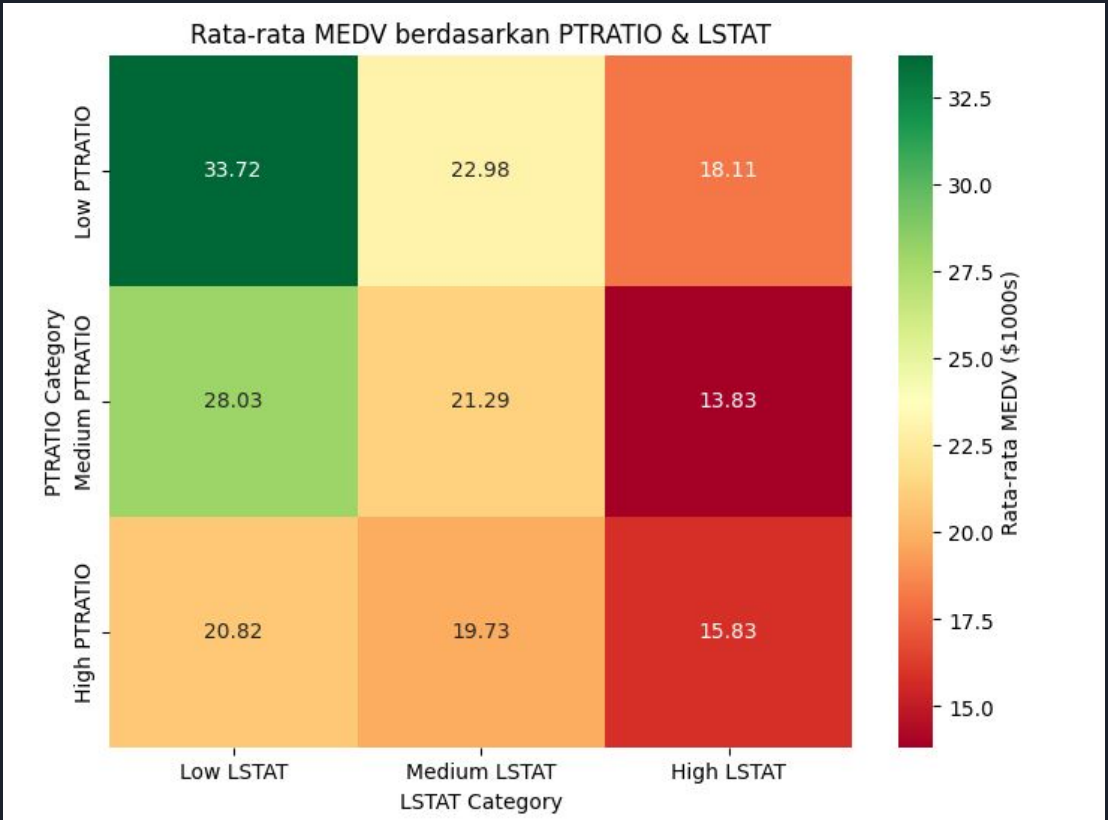
Karakter perumahan di Boston sendiri memberikan rata-rata harga yang cukup mahal jika perumahan tersebut dekat dengan sungai.

Insight Data



Namun jika dilihat dengan membandingkan DIS dengan CHAS. Jumlah rumah dengan CHAS lalu displit menjadi 3 kategori ditance (semakin kecil nilai dekat dengan kota) maka bisa dilihat bahwa karakter masyarakat di Boston lebih ke suburban. Jika dibandingkan dengan Jakarta maka wilayah dengan DIS medium dan high ini seperti wilayah di Depok atau Tangerang.

Insight Data



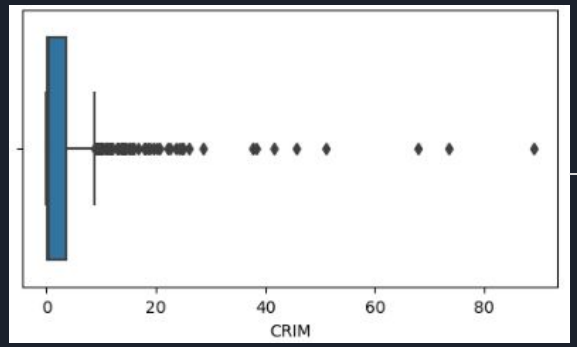
Menjelaskan kualitas dan gaya hidup masyarakat. Perumahan yang bisa memberikan jaminan linkungan baik yaitu tidak terlalu padat penduduk dan memiliki ratio guru berbanding murid balance memberikan harga juga rumah yang cukup tinggi.

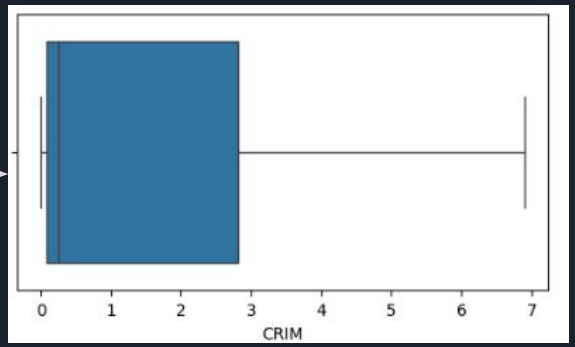
Data Pre-processing

```
Data columns
 #   Column
---  ------
 0   CRIM
 1   ZN
 2   INDUS
 3   CHAS
 4   NOX
 5   RM
 6   AGE
 7   DIS
 8   RAD
 9   TAX
 10  PTRATIO
 11  B
 12  LSTAT
 13  MEDV
```

Total data → 506 row dan 14 kolom (1 dependent target MEDV)

- 5 Kolom memiliki data missing value (CRIM, ZN, INDUS, CHAS, dan AGE) → Mengisi missing value dengan nilai media
- Outlier ditangani dengan cara metode IQR based on winsorize





Modelling

- Kami menggunakan 3 model Random Forest, XGBoost, dan Gradient Boosting Regressor.
- Reason:
  - Tidak melakukan feature selection biarkan model menentukan fitur penting
  - Explainability model yang dari feature importance untuk menilai hubungan fitur dengan case boston ini

Model	R2 (Tuned)	MAE	RMSE	MSE
Random Forest	0.7783	2.1434	3.1432	9.8794
Gradient Boosting	0.8244	1.9851	2.8981	8.3988
XGBoost	0.8128	1.9466	2.6978	7.2779

1. evaluasi ini sudah berdasarkan tuning terhadap R2

Selected Model

XGBoost

```
=== XGBoost ===
train Accuracy :  0.999998185833927
test Accuracy :  0.8982777913297229
R² Score  : 0.8983
```

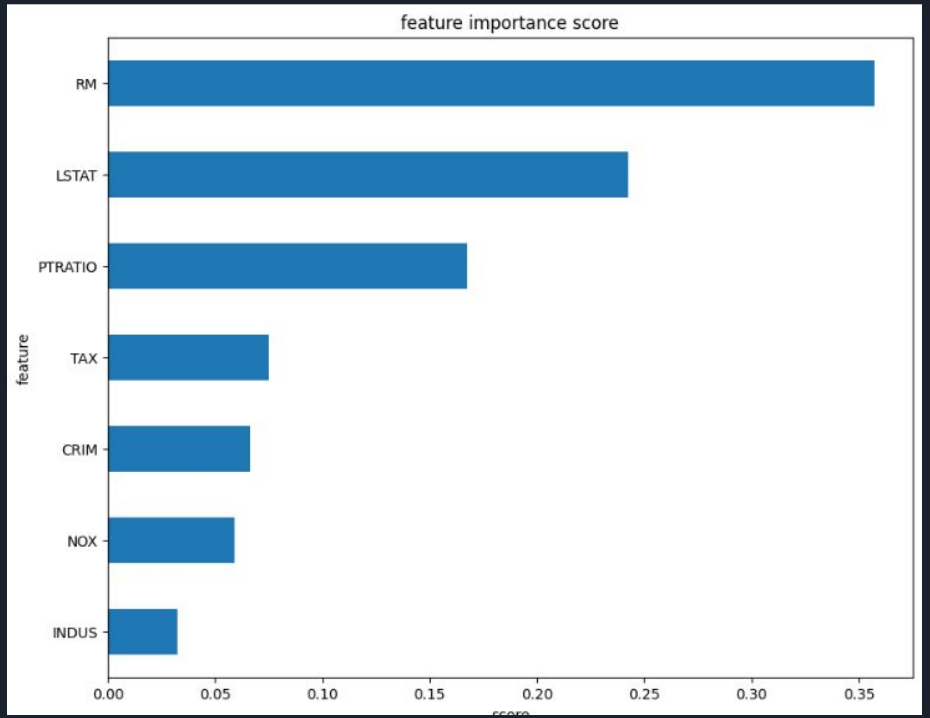
Before
Hyperparameter Tuning

```
Best CV Score (R²): 0.8128
Train score: 0.9702107375716957
Test score :0.9023272024844939
```

After
Hyperparameter Tuning

Selected Model

Feature Importance



Selected Model

Feature Importance memberikan penjelasan 6 fitur teratas yang sangat penting untuk model sehingga dari 6 dijadikan perhatian khusus dalam melakukan prediksi terhadap model.

Akses model dapat dicoba dengan link berikut:

https://78e176ead139098153.gradio.live/

Conclusion

Berdasarkan insight yang didapatkan dan hasil model.

Stakeholder terkait dalam menggunakan prediksi harga Rumah dapat meningkatkan development properti dengan di wilayah yang tidak terlalu padat penduduk namun menjadi tantangan tersendiri bahwa untuk masyarakat di Boston lebih menyukai kemudahan akses ke wilayah industri serta kualitas pendidikan yang memiliki ratio <= 1:18.

Jika dibandingkan casenya dengan properti di Indonesia, perbedaan signifikan terdapat di fitur NOX atau kadar polutan. Masyarakat Boston lebih menyukai properti yang memiliki kadar polutan rendah sehingga perumahan memiliki harga cukup tinggi namun di Indonesia dengan mengambil sample Jabodatabek dengan polutan cukup tinggi memiliki harga properti cukup besar.