



Implementasi Algoritma Regression Untuk Prediksi Harga Rumah Di Boston

Sukron Ma'mum¹, Reza Pratama¹, Muhammad Nurkholis Ajie Kurniawan¹, Ageng Putra Mulya¹, Alfi Ariyadi Anatasia¹, Maulana Fansyuri^{2*}

^{1,2}Ilmu Komputer, Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia

Email: ¹sukronmamum389@gmail.com, ^{2*}dosen02359@unpam.ac.id

(* : coressponding author)

Abstrak - Prediksi harga rumah merupakan aspek krusial dalam industri properti, khususnya di kota besar seperti Boston yang memiliki dinamika pasar yang kompleks. Penelitian ini mengimplementasikan algoritma regresi linier untuk memprediksi harga rumah berdasarkan berbagai atribut seperti jumlah kamar tidur, luas tanah, usia bangunan, dan kedekatannya dengan pusat kota. Dataset yang digunakan terdiri dari 501 data dengan 14 fitur yang mencakup karakteristik rumah dan lingkungan. Proses penelitian melibatkan tahapan Knowledge Discovery in Databases (KDD) mulai dari pemilihan hingga evaluasi model. Hasil analisis menunjukkan bahwa model regresi linier memberikan akurasi yang baik dengan nilai R-squared sebesar 0,745, mengindikasikan bahwa model dapat menjelaskan 74,5% variasi harga rumah. Penelitian ini memberikan wawasan berharga bagi pelaku industri properti untuk mendukung pengambilan keputusan berbasis data yang lebih efektif. Namun, terdapat ruang untuk peningkatan, terutama dalam mengurangi kesalahan prediksi melalui eksplorasi algoritma lain.

Kata Kunci: Prediksi Harga Rumah, Regresi Linear, Boston Housing.

Abstract - House price prediction is a critical aspect of the real estate industry, particularly in major cities like Boston, where the market dynamics are highly complex. This study implements a multiple linear regression algorithm to predict house prices based on various attributes such as the number of bedrooms, land area, building age, and proximity to the city center. The dataset used consists of 501 entries with 14 features that describe the characteristics of houses and their surrounding environments. The research process follows the Knowledge Discovery in Databases (KDD) framework, from data selection to model evaluation. The analysis results show that the linear regression model achieves good accuracy, with an R-squared value of 0.745, indicating the model explains 74.5% of the variability in house prices. This study offers valuable insights for real estate practitioners to support more effective, data-driven decision-making. However, there is room for improvement, particularly in reducing prediction errors by exploring alternative algorithms.

Keywords: House Price Prediction, Linear Regression, Boston Housing.

1. PENDAHULUAN

Harga sektor properti memegang peranan penting dalam perekonomian, terutama di kawasan perkotaan yang memiliki permintaan properti yang tinggi. Salah satu faktor yang krusial dalam sektor ini adalah penentuan harga rumah. Prediksi harga rumah yang akurat sangat penting bagi pembeli, penjual, investor, dan profesional properti dalam membuat keputusan yang tepat. Berbagai faktor, seperti lokasi, karakteristik properti, dan kondisi pasar, memengaruhi penetapan harga rumah, menjadikannya suatu tugas yang kompleks untuk diprediksi dengan tepat (Hidayah, 2023).

Di kota-kota besar seperti Boston, yang dikenal dengan pasar properti yang dinamis, pemahaman terhadap faktor-faktor tersebut serta hubungan antar variabelnya sangatlah penting. Dalam beberapa tahun terakhir, penerapan algoritma machine learning, khususnya teknik regresi, semakin populer dalam memprediksi harga rumah. Model regresi dapat menganalisis data historis dan mengidentifikasi pola yang dapat membantu dalam meramalkan tren di masa depan (Hidayah, 2023; Hasan, 2023).

Penelitian ini bertujuan untuk mengeksplorasi penggunaan algoritma regresi dalam memprediksi harga rumah di Boston, dengan fokus pada berbagai atribut properti seperti jumlah kamar tidur, luas tanah, usia bangunan, dan kedekatannya dengan area utama.

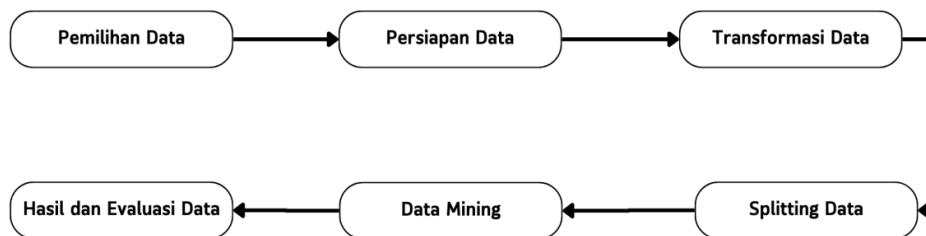
2. METODE

2.1 Pengumpulan Data

Penelitian dalam jurnal ini menggunakan dataset yang diambil dari Kaggle.com yaitu Boston Housing Dataset (Arun, Boston Housing dataset, 2023), yang berisi atribut-atribut seperti tingkat kriminalitas (crim), proporsi tanah untuk zona lot besar (zn), proporsi area bisnis non-ritel (indus), kedekatan dengan Sungai Charles (chas), konsentrasi nitrogen oksida (nox), rata-rata jumlah kamar per rumah (rm), proporsi rumah yang dibangun sebelum 1940 (age), jarak ke pusat pekerjaan (dis), akses ke jalan raya radial (rad), tingkat pajak properti (tax), rasio murid-guru (ptratio), proporsi penduduk keturunan Afrika-Amerika (b), dan persentase populasi dengan status sosial rendah (lstat), dengan label target berupa harga rumah (PRICE). Dataset ini terdiri dari 503 baris data 14 kolom fitur yang masing-masing merepresentasikan berbagai karakteristik lingkungan dan properti rumah.

2.2. Teknik Penelitian

Teknik penelitian menggunakan pendekatan Knowledge Discovery in Databases (KDD) dan memanfaatkan perangkat lunak RapidMiner untuk mempermudah proses pengolahan dan analisis data. Dibawah ini merupakan tahap tahap KDD :



Gambar 1. Tahapan Penelitian

Proses KDD :

- Pemilihan data : Data yang digunakan dipilih berdasarkan relevansinya dengan tujuan penelitian, yaitu memprediksi harga rumah. Data yang dipilih mencakup 503 baris data dan 14 kolom atribut .
- Persiapan data : Data akan diseleksi, dibersihkan dan diproses agar siap untuk analisis. Proses ini meliputi penghapusan data yang tidak lengkap atau duplikat.
- Transformasi Data : Menransformasi data ke format yang sesuai untuk jenis data ke dalam format yang sesuai untuk diproses data mining. Proses transformasi melibatkan modifikasi peran atribut yang akan digunakan sebagai identitas dan label (Khoiriyah, 2024)
- Splitting Data : Split data adalah proses untuk membagi antara data latih dan data uji (A. Yasin, 2022). Data training merupakan data yang akan dipakai dalam pembelajaran, sedangkan data testing merupakan data yang akan digunakan dalam perhitungan penelitian, atau sebagai pengujian kebenaran atau keakurasi hasil (Muningsih, 2020).
- Data Mining : Pemilihan metode algoritma yang sesuai dengan data yang telah dikumpulkan. Di penelitian ini, digunakan algoritma Regresi Linear untuk memprediksi harga rumah. Rumus yang digunakan adalah

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

Y = prediksi nilai variabel dependen

X = variabel independent, m

β_0 = konstanta

β_n = Bobot (koefisien) regresi untuk variabel independent (Fitri, 2023)



Penggunaan regresi linier dalam konteks ini memungkinkan peneliti untuk memahami faktor-faktor yang mempengaruhi harga rumah dan memberikan prediksi yang akurat berdasarkan data yang tersedia. Metode ini juga sering digunakan dalam penelitian serupa karena kesederhanaannya dan kemampuannya untuk memberikan hasil yang dapat diandalkan dalam analisis harga properti (Mu'tashim, 2021).

- f. Hasil dan Evaluasi Data : Tahap untuk menilai seberapa baik model yang telah dibangun dalam menganalisis atau memprediksi data.

3. ANALISA DAN HASIL

3.1 Hasil Pemilihan Data

Berikut merupakan data yang sudah di dipilih berdasarkan relevansi dengan prediksi harga rumah. Terdiri dari 503 baris data dan 14 atribut kolom.

crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lstat	medv
0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24
0.02731	0	7.07	0	0.469	6.421	78.9	49.671	2	242	17.8	396.9	9.14	21.6
0.02729	0	7.07	0	0.469	7.185	61.1	49.671	2	242	17.8	392.83	4.03	34.7
0.03237	0	2.18	0	0.458	6.998	45.8	60.622	3	222	18.7	394.63	2.94	33.4
0.06905	0	2.18	0	0.458	7.147	54.2	60.622	3	222	18.7	396.9	5.33	36.2
0.02985	0	2.18	0	0.458	6.43	58.7	60.622	3	222	18.7	394.12	5.21	28.7
0.08829	12.5	7.87	0	0.524	6.012	66.6	55.605	5	311	15.2	395.6	12.43	22.9
0.14455	12.5	7.87	0	0.524	6.172	96.1	59.505	5	311	15.2	396.9	19.15	27.1
0.21124	12.5	7.87	0	0.524	5.631	100	60.821	5	311	15.2	386.63	29.93	16.5

Gambar 2. Dataset

Penjelasan tiap kolom :

1. crim : tingkat kriminalitas per kapita
2. zn : persentase lahan zonasi untuk perumahan
3. indus : persentase lahan untuk industri non-manufaktur
4. chas : keberadaan lokasi dekat sungai (1 jika dekat, 0 jika tidak)
5. nox : konsentrasi nitrogen oksida (polusi udara)
6. rm : rata-rata jumlah kamar per rumah
7. age : persentase rumah yang dibangun sebelum 1940
8. dis : jarak rata-rata ke pusat pekerjaan (seperti pusat kota)
9. rad : indeks aksesibilitas ke jalan raya utama
10. tax : tarif pajak properti per 10.000 dolar
11. ptratio : rasio jumlah siswa terhadap guru di sekolah
12. b : persentase penduduk berkulit hitam
13. lstat : persentase populasi berstatus sosial ekonomi rendah
14. medv : median nilai rumah (harga rumah)

3.2 Hasil Persiapan Data

Pada persiapan data, peneliti menemukan 2 data yang kosong pada bagian kolom rm dan akhirnya data tersebut dihapus. Dari hasil seleksi data tersebut, data yang akan digunakan berkurang menjadi 501 data.

3.3 Hasil Transformasi Data

Data yang sudah diseleksi selanjutnya akan di atur dengan jenis format yang sesuai. Serta memberikan kategori atribut atau label pada bagian masing-masing kolom.

☒	☒	☒	☒	☒	☒	☒	☒	☒	☒	☒	☒	☒	☒
crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lstat	medv
real	numeric	real	integer	real	numeric	real	real	integer	integer	real	real	real	real
attribute	attribute	attribute	attribute	attribute	attribute	attribute	attribute	attribute	attribute	attribute	attribute	attribute	label
0.006	18	2.310	0	0.538	6.575	65.200	4.090	1	296	15.300	396.900	4.980	24
0.027	0	7.070	0	0.469	6.421	78.900	49.671	2	242	17.800	396.900	9.140	21.600
0.027	0	7.070	0	0.469	7.185	61.100	49.671	2	242	17.800	392.830	4.030	34.700
0.032	0	2.180	0	0.458	6.998	45.800	60.622	3	222	18.700	394.630	2.940	33.400
0.069	0	2.180	0	0.458	7.147	54.200	60.622	3	222	18.700	396.900	5.330	36.200
0.030	0	2.180	0	0.458	6.430	58.700	60.622	3	222	18.700	394.120	5.210	28.700
0.088	12.500	7.870	0	0.524	6.012	66.600	55.605	5	311	15.200	395.600	12.430	22.900
0.145	12.500	7.870	0	0.524	6.172	96.100	59.505	5	311	15.200	396.900	19.150	27.100
0.211	12.500	7.870	0	0.524	5.631	100	60.821	5	311	15.200	386.630	29.930	16.500
0.170	12.500	7.870	0	0.524	6.004	85.900	65.921	5	311	15.200	386.710	17.100	18.900

Gambar 3. Transformasi Data

Dari 14 kolom, dikategorikan 13 atribut meliputi crim, zn, indus, chas, nox, rm, age, dis, rad, tax ptratio, b lstat dan dikategorikan sebagai 1 label yaitu medv (harga rumah).

medv	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lstat
24	0.006	18	2.310	0	0.538	6.575	65.200	4.090	1	296	15.300	396.900	4.980
21.600	0.027	0	7.070	0	0.469	6.421	78.900	49.671	2	242	17.800	396.900	9.140
34.700	0.027	0	7.070	0	0.469	7.185	61.100	49.671	2	242	17.800	392.830	4.030
33.400	0.032	0	2.180	0	0.458	6.998	45.800	60.622	3	222	18.700	394.630	2.940
36.200	0.069	0	2.180	0	0.458	7.147	54.200	60.622	3	222	18.700	396.900	5.330
28.700	0.030	0	2.180	0	0.458	6.430	58.700	60.622	3	222	18.700	394.120	5.210
22.900	0.088	12.500	7.870	0	0.524	6.012	66.600	55.605	5	311	15.200	395.600	12.430
27.100	0.145	12.500	7.870	0	0.524	6.172	96.100	59.505	5	311	15.200	396.900	19.150
16.500	0.211	12.500	7.870	0	0.524	5.631	100	60.821	5	311	15.200	386.630	29.930
18.900	0.170	12.500	7.870	0	0.524	6.004	85.900	65.921	5	311	15.200	386.710	17.100

Gambar 4. Hasil Transformasi Data

3.4 Split Data

Pada rapid miner ,peneliti menggunakan opeator split data untuk membagi dataset training(latih) sebesar 70% dan dataset testing(uji) sebesar 30% secara acak.



Gambar 5. Split Data

3.5 Data Mining

Tahap ini merupakan tahap permodelan, di mana diterapkannya algoritma regresi linear untuk membangun model yang menganalisis hubungan antara fitur-fitur yang ada dan memprediksi nilai target dalam dataset.



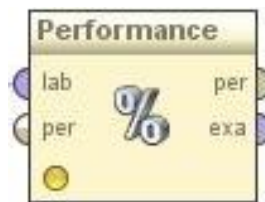
Gambar 6. Linear Regression

Model regresi ini dilatih menggunakan data pelatihan, yang terdiri dari pasangan nilai antara variabel prediktor dan variabel target. Tujuannya adalah untuk menemukan garis lurus yang paling tepat menggambarkan hubungan antara berbagai faktor tersebut.



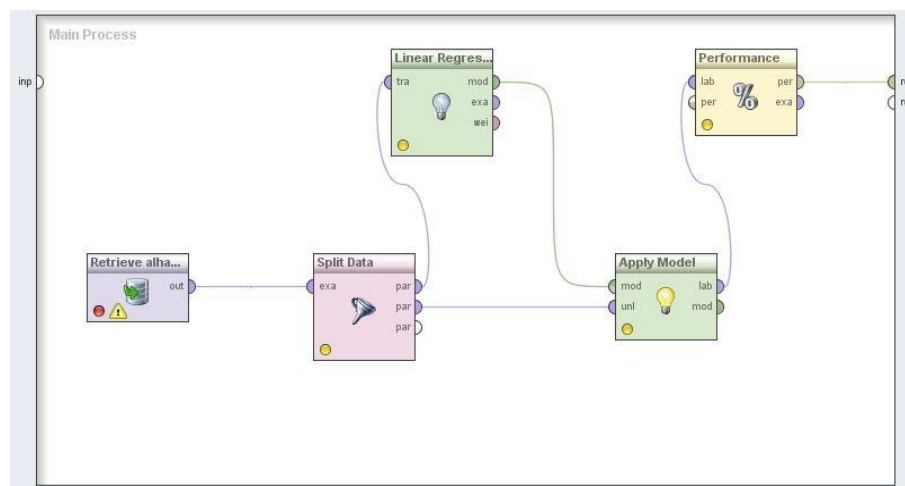
Gambar 7. Apply Model

Apply Model menerapkan model yang telah dilatih (regresi linear) ke data uji untuk menghasilkan prediksi.



Gambar 8. Performance(Regression)

Performance (Regression) mengevaluasi hasil prediksi model dengan membandingkannya dengan nilai asli.



Gambar 10. Design Pada Rapid Miner

3.6 Hasil dan Evaluasi Data

Berikut hasil dari prediksi melalui permodelan Regression Linear :

medv	prediction(...)	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lstat
34.700	30.568	0.027	0	7.070	0	0.469	7.185	61.100	49.671	2	242	17.800	392.830	4.030
22.900	23.280	0.088	12.500	7.870	0	0.524	6.012	66.600	55.605	5	311	15.200	395.600	12.430
16.500	11.581	0.211	12.500	7.870	0	0.524	5.631	100	60.821	5	311	15.200	386.630	29.930
21.700	20.974	0.094	12.500	7.870	0	0.524	5.889	39	54.509	5	311	15.200	390.500	15.710
18.200	19.438	0.638	0	8.140	0	0.538	6.096	84.500	44.619	4	307	21	380.020	10.260
19.900	19.473	0.627	0	8.140	0	0.538	5.834	56.500	44.986	4	307	21	395.620	8.470
17.500	16.716	0.784	0	8.140	0	0.538	5.990	81.700	42.579	4	307	21	386.750	14.670
13.600	11.010	125.179	0	8.140	0	0.538	5.570	98.100	37.979	4	307	21	376.570	21.020
19.600	17.273	0.852	0	8.140	0	0.538	5.965	89.200	40.123	4	307	21	392.530	13.830
15.200	14.680	123.247	0	8.140	0	0.538	6.142	91.700	39.769	4	307	21	396.900	18.720

Gambar 11. Hasil Linear Regression

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 5.048 +/- 0.000
relative_error: 18.98% +/- 25.84%
squared_error: 25.481 +/- 69.976
correlation: 0.863
spearman_rho: 0.903
kendall_tau: 0.746
```

Gambar 12. Hasil Performance

- Root Mean Squared Error (RMSE):** RMSE model adalah **5.048**, yang menunjukkan kesalahan rata-rata dalam memprediksi harga rumah. Artinya, prediksi model memiliki deviasi sekitar **5.048 unit** (misalnya dalam ribuan, tergantung pada skala dataset Anda) dari harga rumah yang sebenarnya.
- Relative Error:** Kesalahan relatif rata-rata model adalah **18.98%**, yang berarti prediksi model, rata-rata, menyimpang sekitar **19%** dari harga rumah yang sebenarnya
- R-squared (R^2):** Nilai R-squared adalah **0.745**, yang berarti model dapat menjelaskan **74.5%** variabilitas harga rumah, yang dianggap sebagai hasil yang baik untuk banyak tugas regresi.
- Korelasi:** Korelasi antara harga rumah yang diprediksi dan harga yang sebenarnya adalah **0.863**, menunjukkan adanya hubungan linier yang kuat antara prediksi dan harga rumah yang sebenarnya.
- Spearman's Rank Correlation:** Nilai Spearman's rho adalah **0.903**, yang menunjukkan korelasi urutan yang sangat kuat antara harga rumah yang diprediksi dan harga yang sebenarnya, yang menunjukkan bahwa model cukup efektif dalam memprediksi urutan harga rumah.
- Kendall's Tau:** Nilai Kendall's Tau adalah **0.746**, yang lebih lanjut mengonfirmasi hubungan yang kuat antara harga rumah yang diprediksi dan harga yang sebenarnya berdasarkan urutan mereka.

Secara keseluruhan, model ini menunjukkan kinerja yang baik dengan tingkat akurasi yang cukup tinggi, meskipun masih terdapat ruang untuk perbaikan, terutama dalam mengurangi kesalahan prediksi absolut. Model ini efektif dalam memprediksi urutan harga rumah dan dapat digunakan untuk analisis lebih lanjut atau untuk aplikasi praktis dalam prediksi harga properti.

4. KESIMPULAN

Penelitian ini telah berhasil mengimplementasikan algoritma regresi linier dalam memprediksi harga rumah di Boston berdasarkan dataset Boston Housing. Hasil analisis menunjukkan bahwa model regresi linier dapat memprediksi harga rumah yang cukup akurat, dengan nilai R-squared sebesar 0,745 dan kesalahan relatif rata-rata sebesar 18,98%. Hal ini menunjukkan bahwa model dapat menjelaskan sekitar 74,5% variabilitas harga rumah yang dipengaruhi oleh faktor-faktor seperti jumlah kamar, luas tanah, usia bangunan, dan jarak ke pusat kota. Model yang dikembangkan efektif dalam menganalisis hubungan antara variabel input dan output, serta memberikan wawasan berharga bagi para profesional di sektor properti untuk membuat keputusan yang lebih baik. Meski demikian, penelitian ini masih memiliki ruang untuk perbaikan dalam mengurangi kesalahan prediksi absolut melalui eksplorasi algoritma atau teknik lain. Secara keseluruhan, pendekatan ini menunjukkan potensi besar dalam aplikasi prediksi harga properti di pasar nyata.



REFERENCES

- A. Yasin, A. Y. (2022). Efektifitas Algoritma Data Mining dalam Menentukan Pendonor Darah Potensial. *Syntax J. Inform*, 12–22.
- Arun, J. (2023, october 15). *Boston Housing dataset*. Dipetik December 24, 2024, dari Kaggle: <https://www.kaggle.com/datasets/arunjangir245/boston-housing-dataset>
- Fitri, E. (2023). Analisis Perbandingan Metode Regresi Linier, Random Forest Regression dan Gradient Boosted Trees Regression Method untuk Prediksi Harga Rumah. *J. Appl. Comput. Sci. Technol*, 58 - 64.
- Hasan, A. (2023). Penggunaan Metode Regresi Linear dalam Prediksi Harga Rumah pada Dataset Boston. . *RPubs*.
- Hidayah, F. A. (2023). Prediksi Harga Rumah di Boston Menggunakan Metode Linear Regression, SVR, Decision Tree dan Random Forest Regression. *Academia.edu*.
- Khoiriyah, S. &. (2024). Penerapan Algoritma Linear Regression dalam Memprediksi Harga Rumah Menggunakan RapidMiner. *JUSIFOR: Jurnal Sistem Informasi dan Informatika*, 2024.
- Mu'tashim, M. L. (2021). Analisis Prediksi Harga Rumah Sesuai Spesifikasi Menggunakan Multiple Linear Regression. *Inform. J. Ilmu Komput.*.
- Muningsih, E. N. (2020). Comparative Analysis on Dimension Reduction Algorithm of Principal Component Analysis and Singular Value Decomposition for Clustering. *Journal of Physics: Conference Series*, 012101.