Analisis Pengaruh Data *Preprocessing* terhadap Performa Model *Machine Learning* pada Kasus Regresi *House Price*

Syahmi Sajid[1]

Magister Kecerdasan Artifisial, Universitas Gadjah Mada
Bulaksumur, Caturtunggal, Kec. Depok, Kabupaten Sleman, Daerah Istimewa Yogyakarta 55281
[1]syahmisajid12@gmail.com

Abstrak— **Penelitian ini bertujuan untuk menganalisis dampak penggunaan teknik data *Preprocessing* khususnya dalam penanganan *Outlier* terhadap kinerja model *Machine Learning* pada kasus regresi harga rumah. *Outlier*, sebagai data ekstrem yang signifikan, dapat memengaruhi estimasi parameter model regresi dan menghasilkan prediksi yang tidak akurat. Kami mengkaji berbagai metode penanganan *Outlier*, termasuk penghapusan, transformasi, dan penggantian nilai-nilai ekstrem, serta menganalisis bagaimana penerapannya memengaruhi kualitas prediksi harga rumah. Data *Preprocessing* yang tepat dapat meningkatkan kestabilan dan akurasi model regresi, yang pada gilirannya dapat membantu pemilik properti, investor, dan agen real estate dalam membuat keputusan yang lebih cerdas terkait harga rumah. Hasil penelitian ini memberikan wawasan penting tentang pentingnya penanganan *Outlier* dalam analisis harga rumah menggunakan machine learning, serta memberikan panduan praktis untuk meningkatkan kinerja model dalam konteks ini.**

Kata kunci— **Data *Preprocessing*, *Outlier* Handling, Machine Learning, Linear Regression, *House Price***

I. PENDAHULUAN

Kasus Harga rumah adalah salah satu aspek penting dalam pasar perumahan yang memiliki implikasi signifikan dalam keputusan finansial individu dan organisasi. Kemampuan untuk dengan tepat memprediksi harga rumah memiliki nilai yang sangat besar dalam berbagai konteks, seperti investasi properti, perencanaan keuangan, dan penentuan harga penjualan. Dalam era modern yang didukung oleh teknologi, *Machine Learning* (ML) telah menjadi alat yang penting dalam menganalisis dan memprediksi harga rumah. Penelitian sebelumnya pernah dilakukan menggunakan algoritma pembelajaran mesin sebagai metodologi penelitian untuk mengembangkan model prediksi harga rumah [1]. Banerjee menerapkan teknik pembelajaran mesin untuk memprediksi apakah harga rumah akan naik atau turun [2]. Kang mengusulkan kerangka fusi data untuk mengkaji seberapa baik potensi apresiasi harga rumah dapat diprediksi dengan menggabungkan berbagai sumber data [3].

Namun, proses pengembangan model ML untuk memprediksi harga rumah tidak selalu berjalan mulus. Data yang digunakan untuk melatih model seringkali memiliki karakteristik yang rumit, termasuk *Outlier*. *Outlier* adalah titik data yang jauh dari nilai-nilai lain dalam dataset, yang dapat disebabkan oleh berbagai faktor seperti kesalahan pengukuran, ketidakpastian, atau bahkan situasi khusus yang langka. Ketika *Outlier* tidak dikelola dengan baik, mereka dapat memiliki dampak yang signifikan pada performa model ML dalam kasus regresi harga rumah.

Penanganan *Outlier* menjadi fokus utama dalam penelitian ini. Kami akan menjelajahi bagaimana penggunaan berbagai teknik data *Preprocessing*, khususnya dalam penanganan *Outlier*, dapat memengaruhi performa model ML dalam memprediksi harga rumah dengan akurat. Beberapa teknik yang akan dievaluasi termasuk penghapusan *Outlier*, transformasi data, dan penggantian nilai-nilai *Outlier* dengan nilai yang lebih wajar. Tujuan utama kami adalah untuk memahami dampak dari setiap teknik *Preprocessing* ini dan memberikan panduan praktis tentang pemilihan teknik terbaik untuk kasus regresi harga rumah.

Adanya *Outlier* dalam dataset harga rumah bisa menjadi tantangan besar. *Outlier* dapat memberikan sinyal yang salah kepada model, yang akhirnya mengarah pada prediksi yang tidak akurat. Misalnya, ketika *Outlier* tidak ditangani dengan benar, mereka dapat "menarik" model regresi jauh dari tren sebenarnya dalam data, sehingga prediksi harga rumah menjadi sangat tidak realistis. Sebaliknya, jika *Outlier* dihilangkan secara kasar, kita bisa kehilangan wawasan berharga tentang sifat-sifat unik dari properti yang mahal atau langka.

Selain itu, harga rumah sering kali memiliki struktur yang kompleks. Berbagai faktor seperti lokasi, ukuran properti, kondisi fisik, dan aspek-aspek lainnya dapat memengaruhi harga. Oleh karena itu, pengembangan model ML yang kuat dan akurat memerlukan perhatian khusus terhadap karakteristik data dan pemilihan metode *Preprocessing* yang tepat. Dalam konteks ini, analisis pengaruh data *Preprocessing*, khususnya penanganan *Outlier*, menjadi penting untuk memastikan bahwa model yang dihasilkan dapat memberikan prediksi harga rumah yang dapat diandalkan.

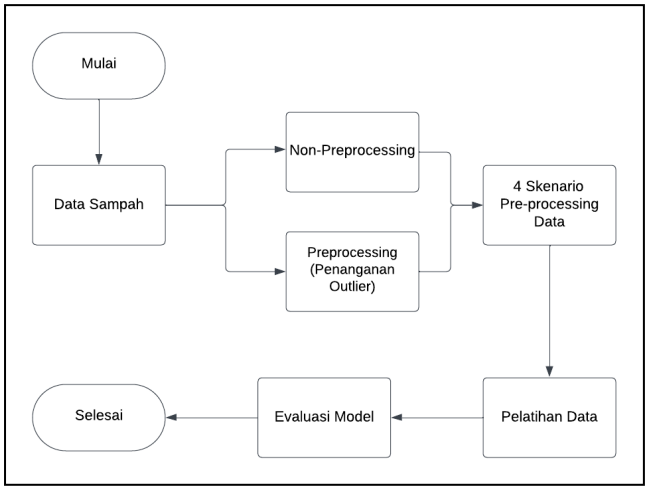
Penelitian sebelumnya telah mengidentifikasi berbagai metode dan teknik untuk menangani *Outlier* dalam berbagai konteks. Namun, masih ada kebutuhan untuk menerapkan dan memahami metode-metode ini secara khusus dalam kasus regresi harga rumah. Pengaruh penggunaan berbagai teknik *Preprocessing* pada performa model ML dalam konteks ini belum sepenuhnya terungkap.

Penelitian ini memiliki beberapa tujuan utama. Pertama, kami akan menganalisis berbagai teknik data *Preprocessing* yang digunakan untuk menangani *Outlier* dalam dataset harga rumah. Kami akan mengevaluasi dampak dari masing-masing teknik ini pada performa model ML. Kedua, kami akan menyediakan panduan praktis kepada para profesional di bidang properti, investor, dan pengembang ML tentang bagaimana memilih dan menerapkan teknik *Preprocessing* yang sesuai untuk meningkatkan akurasi prediksi harga rumah. Ketiga, penelitian ini berkontribusi pada pemahaman yang lebih baik tentang kompleksitas dalam pengembangan model ML untuk kasus regresi harga rumah, dengan menekankan pentingnya penanganan *Outlier*.

Hasil dari penelitian ini akan memberikan wawasan penting kepada para pemangku kepentingan dalam industri perumahan dan machine learning. Diharapkan penelitian ini akan membantu meningkatkan kemampuan prediksi harga rumah, yang pada akhirnya akan mendukung pengambilan keputusan yang lebih baik dalam berbagai konteks yang berkaitan dengan perumahan. Selain itu, penelitian ini dapat menjadi dasar untuk penelitian lanjutan dalam pengembangan model ML yang lebih canggih dalam analisis harga rumah.

II. METODOLOGI

A. Tahapan Penelitian



Gambar 1. Tahapan Penelitian

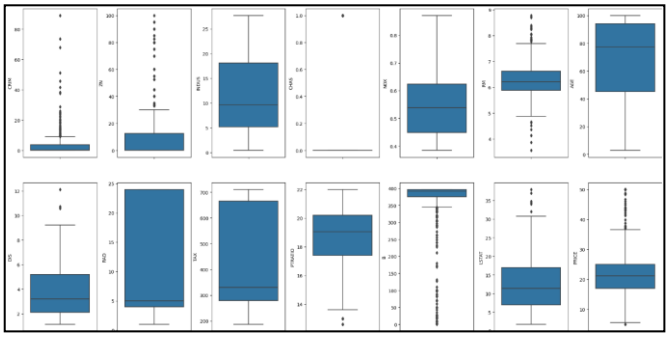
Tahapan penelitian seperti yang ditunjukkan pada Gambar 1 dimulai dari proses pengumpulan dataset house prediction. Tahap berikutnya dilakukan pembagian 2 skenario yaitu *non Preprocessing* dan *Preprocessing* untuk menangani data *Outlier*. Masing-masing hasil dari proses sebelumnya dilakukan kombinasi percobaan terhadap 4 skenario pada proses pre-prosesing data. Setiap rangkaian kombinasi skenario dilakukan proses pelatihan (*training*) menggunakan model *Machine Learning* logistik regresi. Semua hasil model yang terbentuk dilakukan proses pengujian model menggunakan beberapa matriks evaluasi dan semua skenario dibandingkan untuk menentukan skenario terbaik.

B. Dataset

Tabel 1. Dataset House Prediction

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	PRICE
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296.0	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242.0	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222.0	18.7	396.90	5.33	36.2

Dataset yang digunakan dapat dilihat dari tabel 1. Dataset pada penelitian ini bersumber dari Kaggle yang berjudul "Boston *House Prices*" dan dipublish oleh Manimala. Total dataset yang diperoleh yaitu berjumlah 506 data dengan 13 variabel independen (Crim, Zn, Indus, Chas, Nox, Rm, Age, Dis, Rad, Tax, Ptratio, B, Dan Lstat) dan 1 variabel dependen (Price).



Gambar 2. Boxplot

Tabel 1. *Outlier* Masing-masing Variabel

Nama Kolom	*Outlier* %
CRIM	13.04%
ZN	13.44%
INDUS	0.00%
CHAS	100.00%
NOX	0.00%
RM	5.93%
AGE	0.00%
DIS	0.99%
RAD	0.00%
TAX	0.00%
PTRATIO	2.96%
B	15.22%
LSTAT	1.38%
PRICE	7.91%

Pada gambar 2 dan Tabel 1 dapat dilihat adanya *Outlier* pada beberapa variable seperti variable CRIM, ZN, CHAS, RM, DIS, PTRATIO, B, LSTAT, PRICE.

C. Skenario

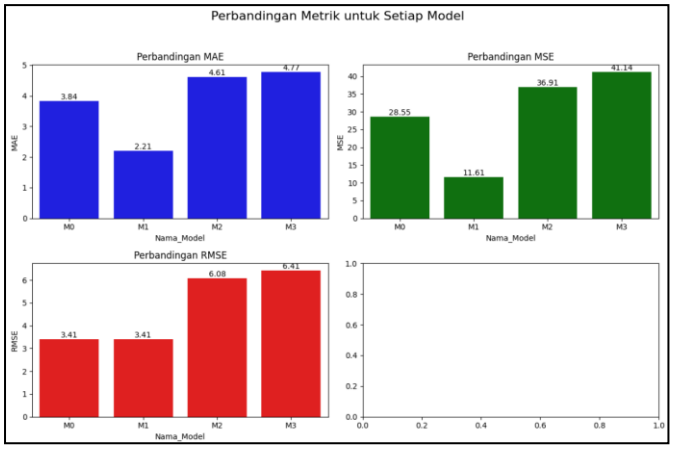
Untuk mendapatkan hasil yang terbaik dalam penelitian ini menggunakan 6 skenario pada proses *Preprocessing*. 6 skenario yang digunakan dalam penelitian ini dapat dilihat pada Tabel 2

Tabel 2. Skenario *Preprocessing*

No	Kode	Skenario
1	M1	Non *Preprocessing*
2	M2	*Outlier* Handling (Menghapus *Outlier*)
3	M3	*Outlier* Handling (Mengganti *Outlier* = Median)
4	M4	*Outlier* Handling (Mengganti *Outlier* = 0)

Tabel 2 berisi 4 Skenario yang akan dilakukan pada penelitan ini, masing masing akan di modelkan dan dievaluasi untuk mencari mana skenario terbaik untuk kasus ini dan melihat pengaruh data *Preprocessing* pada performa model machine learning.

III. HASIL DAN PEMBAHASAN



Gambar 4. Perbandingan Performa Model

Model Terbaik didalam kasus ini adalah Model 1 dengan cara menghapus *Outlier* mendapat semua nilai matriks evalusi terendah dibanding model lain sebesar MAE 2.21, MSE 11.61, RMSE 3.41. M2 (*Outlier* = Median) dengan nilai matrik evaluasi MAE 4.61, MSE 36.91, RMSE 6.08 dan M3 (*Outlier* = 0) dengan nilai matrik evaluasi MAE 4.77, MSE 41.14, RMSE 6.41 mendapat nilai lebih besar dari model 0 (sebelum *Preprocessing*) dengan nilai matrik evaluasi MAE 3.84, MSE 28.55, RMSE 3.41.

IV. KESIMPULAN

Kasus ini didapat model terbaik adalah menghapus *Outlier*, metode mengganti *Outlier* dengan nilai mean dan 0 mengakibatkan performa model *Machine Learning* semakin buruk dibanding sebelum *Preprocessing*

REFERENSI

[1] Byeonghwa Park; Jae Kwon Bae; "Using Machine Learning Algorithms for Housing Price Prediction: The Case of Fairfax County, Virginia Housing Data" EXPERT SYST. APPL., 2015.

[2] Debanjan Banerjee; Suchibrota Dutta; "Predicting The Housing Price Direction Using Machine Learning Techniques", 2017 IEEE INTERNATIONAL CONFERENCE ON POWER, CONTROL, 2017.

[3] Yuhao Kang; Yuhao Kang; Fan Zhang; Wenzhe Peng; Song Gao; Jinmeng Rao; Fábio Duarte; Fábio Duarte; Carlo Ratti; "Understanding House Price Appreciation Using Multi-source Big Geo-data and Machine Learning", LAND USE POLICY, 2020.