

IMPLEMENTASI SUPPORT VECTOR MACHINE DALAM PREDIKSI HARGA RUMAH

I Wayan Pio Pratama

Diploma III Teknologi Informasi, Politeknik eLBajo Commodus, Labuan Bajo, Indonesia,
pio.pratama@poltekelbajo.ac.id

Abstract

This study aims to predict house prices in the United States based on features such as Avg. Area Income, Avg. Area House Age, Avg. Number of Rooms, and Area Population. The dataset used consists of several thousand entries without missing values. Through initial data exploration, a significant correlation was found between several features and house prices. Outliers were analyzed and considered in the modeling process. The Support Vector Machine (SVM) algorithm with various kernels was applied, where the Radial Basis Function (RBF) kernel showed the best performance, explaining about 70.78% of the variation in house prices. The results of this study highlight the potential of the SVM algorithm in property price prediction and provide insights for further property analysis.

Keywords: *Support Vector Machine, House Prices, Outliers, Regression Analysis, Multicollinearity.*

Abstrak

Penelitian ini bertujuan untuk memprediksi harga rumah di Amerika Serikat berdasarkan fitur-fitur seperti *Avg. Area Income*, *Avg. Area House Age*, *Avg. Number of Rooms*, dan *Area Population*. Dataset yang digunakan terdiri dari beberapa ribu entri tanpa ada *missing value*. Melalui eksplorasi data awal, ditemukan adanya korelasi yang signifikan antara sejumlah fitur dengan harga rumah. *Outlier* dianalisis dan dipertimbangkan dalam proses pemodelan. Algoritma Support Vector Machine (SVM) dengan berbagai kernel diaplikasikan, di mana kernel Radial Basis Function (RBF) menunjukkan performa terbaik dan berhasil menjelaskan sebesar 70,78% variasi harga rumah. Hasil penelitian ini menunjukkan potensi algoritma SVM dalam prediksi harga properti dan memberikan wawasan untuk analisis properti lebih lanjut.

Kata kunci: *Support Vector Machine, Harga Rumah, Outlier, Analisis Regresi, Multikolinearitas.*

PENDAHULUAN

Properti merupakan salah satu jenis investasi yang memiliki nilai *return* yang tinggi. Investasi di sektor properti menunjukkan ketahanannya terhadap berbagai krisis baik ekonomi maupun kesehatan. Sejalan dengan riset yang dilakukan oleh Bank Tanah Negara (BTN), harga properti di masa pandemi covid-19 nyatanya mengalami kenaikan sebesar 5,24% secara *year on year* per Maret 2021 (Bank BTN, 2021). Selain menguntungkan dari aspek perolehan *capital gain*, properti juga dapat memperoleh *cash flow* dari hasil sewa menyewa (Sariasih et al., 2021). Hal tersebut menyebabkan banyak orang tetap memilih properti sebagai investasi favorit selain surat berharga, saham, emas, dan deposito. Salah satu jenis investasi properti yang paling umum adalah perumahan. Dengan meningkatnya jumlah populasi dan keuntungan dari sisi investasi menyebabkan rumah masih menjadi kebutuhan yang penting dimiliki oleh banyak orang terutama generasi milenial (Kementerian PUPR, 2019). Tingginya permintaan, pembangunan infrastruktur, kawasan strategis, dan laju inflasi merupakan

beberapa faktor kenapa harga rumah menunjukkan peningkatan setiap tahunnya secara umum (SMF IT, 2017).

Kebutuhan yang tinggi akan rumah, menyebabkan banyak orang kesulitan menentukan atau mengestimasi harga dari sebuah rumah. Beberapa orang kerap terjebak karena seringnya terjadi permainan harga yang dibuat oleh spekulan, sehingga terjadi efek *bubble* terhadap harga properti (Ruhlessin, 2022). Harga rumah yang tidak wajar menyebabkan tujuan awal investasi tidak tercapai dengan baik. Walaupun pemerintah telah membuat serangkaian aturan seperti Nilai Jual Objek Pajak (NJOP) sebagai dasar penentuan harga rumah, realitanya harga rumah masih sangat bergantung dari mekanisme pasar yang terbentuk dari spekulan. Beberapa aspek yang menjadi penentu harga dari sebuah rumah adalah luas, jumlah ruangan, umur rumah, jumlah kamar tidur, dan jumlah populasi di sekitar perumahan.

Sulitnya menentukan harga suatu rumah berdasarkan kriteria yang ada, menyebabkan harga rumah di suatu lingkungan tertentu memiliki harga yang tidak wajar. Salah satu cara untuk mengatasi hal tersebut adalah dengan memanfaatkan data historis dari penjualan rumah. Konsep data historis penjualan rumah dapat membantu dalam melakukan prediksi harga rumah dengan lebih baik. Permasalahannya adalah tiap lokasi perumahan memiliki karakteristik yang berbeda dan bahkan tidak ada pada data historis. Lebih lanjut untuk memahami data historis bukan sesuatu yang mudah untuk dilakukan terlebih bagi orang awam. Perlunya pemahaman pengolahan, analisis, dan interpretasi data yang tepat agar dapat memprediksi harga suatu rumah dengan baik. Untuk mengatasi hal tersebut, penggunaan algoritma *Machine Learning* (ML) dapat dimanfaatkan. ML menggunakan data historis untuk membentuk sebuah model dan mampu mengenali pola rumit untuk dapat memprediksi harga rumah kendati belum pernah dipelajari.

Berdasarkan uraian tersebut, maka penelitian ini menunjukkan bagaimana mengimplementasikan salah satu algoritma ML yaitu SVM untuk melakukan prediksi harga rumah. Penelitian ini memanfaatkan langkah-langkah data *preprocessing*, pelatihan, dan pengujian untuk mengetahui akurasi dari model yang terbentuk.

TINJAUAN PUSTAKA

Beberapa penelitian untuk melakukan prediksi harga rumah telah dilakukan. Analisis dan prediksi harga rumah bekas menggunakan Random Forest menunjukkan hasil yang menjanjikan (Zhang et al., 2022). Penelitian lain yang dilakukan oleh Quang dkk. melakukan pengujian dengan tiga algoritma ML yaitu *Random Forest*, *XGBoost*, dan *LightGBM* (Quang et al., 2020), melalui penelitiannya dijelaskan bahwa metode *Light GBM* menunjukkan kompleksitas terbaik, sedangkan secara akurasi metode berbasis *Random Forest* menunjukkan hasil yang lebih baik.

Riset lain dilakukan oleh Fondy dkk. menunjukkan hasil yang baik ketika SVM digunakan untuk melakukan prediksi harga laptop (Fondy et al., 2019). Prediksi harga rumah juga dilakukan pada penelitian Lumbanraja dkk., hasilnya diketahui bahwa akurasi yang diperoleh berkisar antara 86,60% sampai 95,99%

menggunakan SVM (Lumbanraja et al., 2021). Pengembangan yang dapat dilakukan dari penelitian tersebut adalah penggunaan data yang lebih banyak.

Tingkat akurasi dari beberapa penelitian sebelumnya menunjukkan hasil yang menjanjikan. Penggunaan berbagai algoritma seperti *Random Forest*, *XGBoost*, *LighGBM*, dan SVM juga sangat kompetitif. Walaupun pada beberapa penelitian menunjukkan *Random Forest* dapat mencapai tingkat akurasi yang sedikit lebih tinggi dibandingkan SVM (Zhang et al., 2022), namun penggunaan algoritma berbasis *tree* memiliki beberapa kelemahan seperti lamanya proses pelatihan model dan sulitnya menemukan *hyperparameter* yang optimal (Azizah, 2017). Sebaliknya, SVM lebih mudah untuk diinterpretasikan, kompleksitas yang lebih rendah, penanganan yang lebih baik terhadap data tidak seimbang, dan lebih mudah menyusun *hyperparameter* yang optimal (El Morr et al., 2022).

METODE

Data Collection

Data Collection merupakan cara untuk memperoleh data yang digunakan dalam penelitian. Penelitian ini memanfaatkan data sekunder yang diperoleh melalui sumber lain. Sumber data yang digunakan adalah Kaggle yang dapat diakses melalui dataset (Prabhu, 2023). Data tersebut merupakan data mengenai perumahan dari 5000 orang yang tinggal di Amerika. Penelitian ini memanfaatkan seluruh data yang tersedia dalam proses pembelajaran maupun pengujian sistem.

Data Description

Data Description merupakan proses identifikasi informasi asli data yang digunakan dalam penelitian ini. Informasi ini penting untuk diketahui agar data yang digunakan dapat dipahami dengan lebih jelas. Data rumah yang diperoleh adalah 5000. Data tersebut berisikan 4 kolom yang terdiri dari *Avg. Area Income*, *Avg. Area House Age*, *Avg. Area Number of Rooms*, *Avg. Area Number of Bedrooms*, *Area Population*, *Price*, dan *Address*. Kolom *Avg. Area Income* merupakan rata-rata pendapatan orang-orang yang tinggal di suatu daerah. *Avg. Area House Age* merupakan rata-rata umur rumah yang ada di suatu daerah. *Avg. Area Number of Rooms* menunjukkan jumlah kamar di setiap rumah pada dataset tersebut. Kolom berikutnya adalah *Avg. Area Number of Bedrooms* merupakan rata-rata jumlah kamar tidur di suatu rumah. *Area Population* menunjukkan jumlah orang yang tinggal di suatu wilayah. *Price* adalah harga rumah dan *Address* menunjukkan alamat rumah pada dataset.

Data Preprocessing

Data Preprocessing merupakan tahap yang penting untuk dilakukan karena dampak yang diperoleh dapat meningkatkan performa algoritma ML. Tahap ini dapat terdiri dari beberapa hal seperti *missing value*, deteksi *outlier*, penanganan *outlier*, analisis korelasi, dan normalisasi data.

Data Splitting

Penelitian ini menggunakan algoritma SVM untuk membuat sebuah sistem yang dapat memprediksi harga dari rumah. Penelitian ini membagi dataset menjadi

data *training* dan *testing*. Data *training* yang digunakan dalam penelitian ini adalah 80% dari keseluruhan data, sedangkan 20% sisanya menjadi data *testing*.

SVM

Klasifikasi dan regresi adalah dua hal yang dapat didekati penyelesaiannya dengan metode SVM (Huang & Putra Purnoma, 2022). SVM memanfaatkan konsep yang disebut dengan *hyperplane* untuk menemukan pemisah antar data (Alita et al., 2020). Dalam perkembangannya metode ini dapat digunakan untuk memisahkan lebih dari dua kelas dan juga regresi. Dalam kasus regresi, SVM sering juga disebut dengan *Support Vector Regression* (SVR). SVR memberikan fleksibilitas untuk mendefinisikan error yang dapat diterima pada model yang dibuat, sehingga diperoleh *hyperplane* yang pas dengan data. Berbeda dengan kasus klasifikasi dimana SVM berusaha menyelesaikan permasalahan untuk mencari nilai minimum dari *Ordinary Least Squares* (OLS), pada kasus regresi fungsi objektif dan konstrain ditunjukkan pada persamaan (1) dan (2).

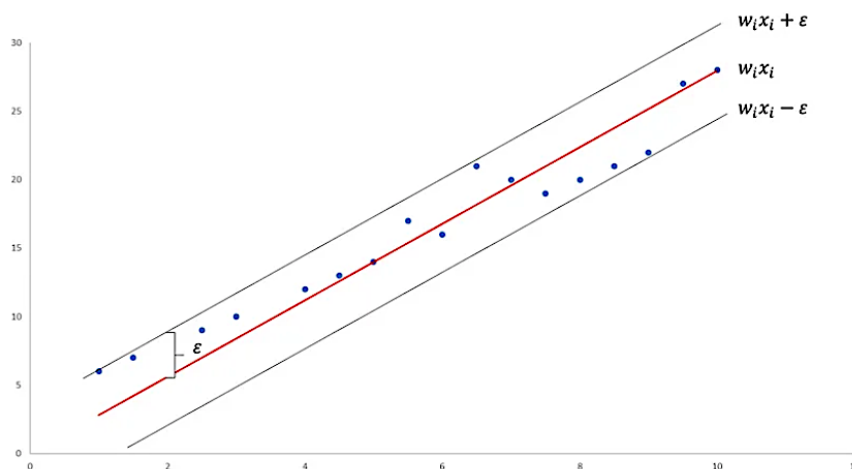
Fungsi Objektif:

$$\text{MIN} \frac{1}{2} \|w\|^2 \quad (1)$$

Konstrain:

$$|y_i - w_i x_i| \leq \varepsilon \quad (2)$$

Berikut adalah contoh pemanfaatan SVR dapat diilustrasikan dengan Gambar 1.



Gambar 1. Ilustrasi SVR
Sumber: Olah Data, 2023

Root Mean Square Error (RMSE)

Metode RMSE merupakan metode kuadratik untuk mengukur skor dengan cara menghitung rata-rata magnitudo dari error. RMSE mengindikasikan beberapa hal yaitu kesesuaian absolut antara model dan data. Selain itu, RMSE menunjukkan rata-rata error prediksi dari model. Penelitian ini memanfaatkan metode RMSE untuk mengukur keakuratan model SVM yang terbentuk untuk memprediksi harga rumah. Semakin rendah nilai dari RMSE menunjukkan model yang dibuat mampu melakukan prediksi dengan baik (EMOTIVEPRO, 2019). Perhitungan RMSE dapat dilakukan dengan persamaan (3).

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2} \quad (3)$$

Dimana,

RMSE : Rata-rata kuadrat akar dari error

N : Jumlah data

$\hat{y}_i$: Target

y_i : Prediksi

HASIL DAN DISKUSI

Eksplorasi Awal

Dalam tahap awal penelitian ini, dataset properti yang berasal dari Amerika Serikat telah berhasil diimpor ke dalam platform Python untuk dilakukan analisis lebih mendalam. Dataset ini memberikan gambaran rinci mengenai harga rumah serta berbagai fitur yang berkaitan dengannya. Sebagai langkah awal, lima entri pertama dari dataset (Tabel 1) diobservasi untuk memberikan pemahaman awal mengenai struktur serta jenis informasi yang disajikan oleh dataset.

Tabel 1. Contoh Data

No	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price
1	79545,458574	5,682861	7,009188	4,09	23086,800503	1,059034e ⁶
2	79248,642455	6,002900	6,730821	3,09	40173,072174	1,505891e ⁶
3	61287,067179	5,865890	8,512727	5,13	36882,159400	1,058988e ⁶
4	63345,240046	7,188236	5,586729	3,26	34310,242831	1,260617e ⁶
5	59982,197226	5,040555	7,839388	4,23	26354,109472	6,309435e ⁶

Sumber: Olah Data, 2023

Selanjutnya, proses verifikasi dilakukan terhadap dataset untuk memastikan tidak adanya nilai yang hilang, dan hasilnya menunjukkan bahwa dataset ini telah lengkap dan siap untuk analisis lebih lanjut. Selain itu, eksplorasi statistik dasar dilakukan untuk memperoleh ringkasan mengenai distribusi dan karakteristik data, meliputi nilai rata-rata, median, standar deviasi, dan metrik lainnya yang relevan seperti diperlihatkan pada Tabel 2.

Tabel 2. Hasil Statistika Deskriptif

	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price
Jumlah Data	5000	5000	5000	5000	5000	5000
Rata-Rata	68583	5,9	7	3,9	36163,5	1,2
Std	10658	0,9	1	1	9925,6	3,5
Min	17797,6	2,6	3	2	172,6	1,6
25%	61481,5	5,3	6	3	29403,9	10
50%	68804,3	6	7	4	36199,4	1,2
75%	75783,3	6,6	7,6	4,5	42861,3	1,5
Max	107702,7	9,5	10,7	6,5	69621,7	2,5

Sumber: Olah Data, 2023

Dalam ringkasan statistik dasar dari dataset, beberapa temuan penting dapat diketahui seperti:

Avg. Area Income: Berkisar antara \$17,796 hingga \$107,701 dengan rata-rata sebesar \$68,583;

Avg. Area House Age: Berkisar dari 2.64 tahun hingga 9.52 tahun dengan rata-rata sebesar 5.98 tahun;

Avg. Area Number of Rooms: Berkisar dari 3.24 hingga 10.76 dengan rata-rata sebesar 6.99 ruangan;

Avg. Area Number of Bedrooms: Berkisar dari 2 hingga 6.5 dengan rata-rata sebesar 3.98 kamar tidur;

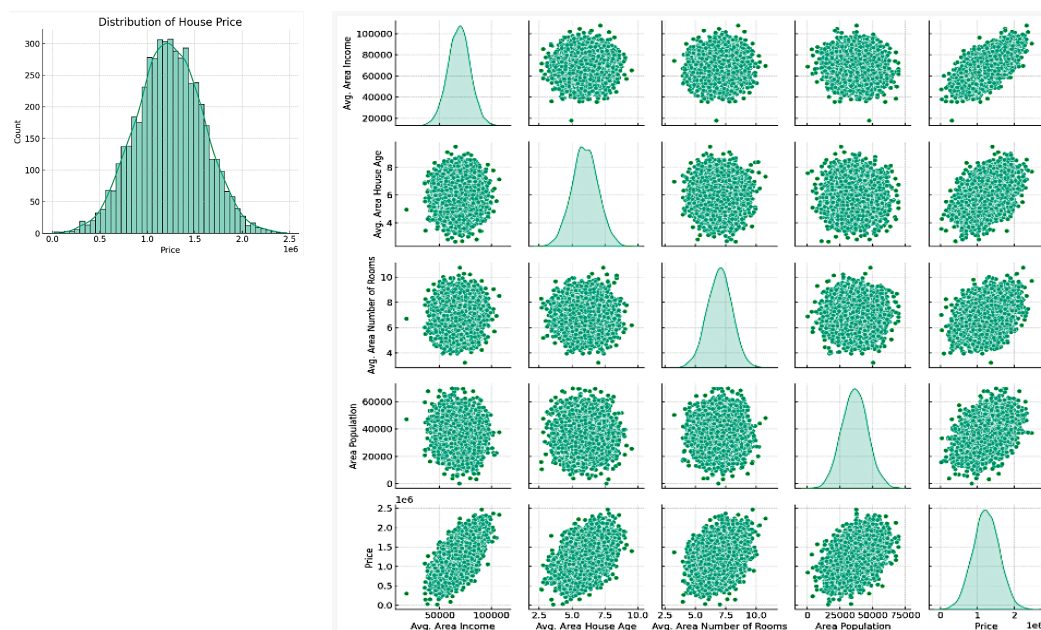
Area Population: Berkisar dari 172.61 hingga 69,621.71 dengan rata-rata sebesar 36,163.52 penduduk;

Price: Berkisar dari \$15,938.66 hingga \$2,469,066 dengan rata-rata sebesar \$1,232,073.

Data Visualisasi

Sebagai bagian dari analisis eksploratif data, berbagai teknik visualisasi diterapkan untuk mendapatkan wawasan mendalam tentang dataset. Pertama, visualisasi distribusi harga rumah, yang merupakan variabel target dalam studi ini, dilakukan. Hal ini memberikan gambaran mengenai bentuk dan karakteristik distribusi harga rumah di Amerika Serikat.

Selanjutnya, untuk memahami bagaimana masing-masing fitur berkaitan dengan harga rumah, hubungan *pairwise* antara fitur-fitur dan variabel target divisualisasikan melalui plot regresi dan scatter. Teknik ini memungkinkan pengamat untuk menilai secara cepat hubungan linear atau non-linear antara variabel-variabel tersebut.



Gambar 2. Visualisasi Data

Sumber: Olah Data, 2023

Berdasarkan analisis hubungan *pairwise* pada gambar 2 diatas, diketahui bahwa:

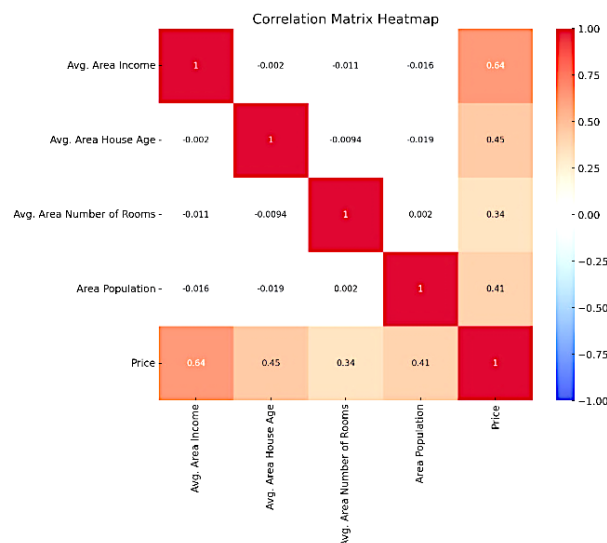
Avg. Area Income: Terdapat korelasi positif yang jelas antara *Avg. Area Income* dengan harga rumah. Hal ini berarti bahwa area dengan pendapatan rata-rata yang lebih tinggi cenderung memiliki harga rumah yang lebih tinggi;

Avg. Area House Age: Hubungan antara usia rumah dengan harga menunjukkan korelasi positif, di mana rumah yang lebih tua cenderung memiliki harga yang lebih tinggi;

Avg. Area Number of Rooms: Ada sebaran tetapi secara umum terdapat korelasi positif antara jumlah ruangan dengan harga rumah;

Area Population: Hubungan antara *Area Population* dengan harga rumah tidak sejelas fitur lainnya, namun ada kecenderungan positif yang ringan; Harga rumah berdistribusi normal membentuk pola kurva lonceng.

Terakhir, untuk mendapatkan pemahaman yang lebih holistik tentang interaksi antar fitur serta potensi adanya multikolinearitas, peta *heatmap* korelasi dibuat (Gambar 3). Korelasi antar fitur dapat memberikan informasi penting dalam pemilihan fitur dan menginformasikan langkah-langkah pemrosesan data selanjutnya.



Gambar 3. Heatmap antar fitur

Sumber: Olah Data, 2023

Dalam *heatmap*, gradien warna biru hingga merah digunakan untuk menandai korelasi negatif hingga positif. Nilai yang mendekati 1 atau -1 mengindikasikan adanya korelasi positif atau negatif yang kuat, sementara nilai yang mendekati 0 menunjukkan korelasi yang lebih lemah.

Dari analisis *heatmap*, beberapa observasi penting dapat dicatat:

Avg. Area Income menunjukkan korelasi positif yang signifikan dengan Harga dengan koefisien korelasi sebesar 0.64;

Avg. Area House Age di Area juga menunjukkan korelasi positif dengan Harga, dengan koefisien sebesar 0.46;

Avg. Area Number of Rooms di *Area* dan *Area Population* menunjukkan korelasi positif dengan Harga, meski koefisiennya lebih rendah dibandingkan fitur lainnya; Fitur-fitur tersebut tidak menunjukkan korelasi yang sangat tinggi satu sama lain, hal ini mengindikasikan bahwa multikolinearitas bukanlah isu signifikan untuk fitur-fitur ini.

Penting untuk diingat bahwa meskipun korelasi memberikan wawasan tentang hubungan linear antara variabel, hal ini tidak selalu menyiratkan hubungan sebab-akibat. Selain itu, hubungan non-linear mungkin tidak sepenuhnya diwakili oleh koefisien korelasi.

Pemahaman mengenai korelasi sangat berguna dalam pengambilan keputusan tentang fitur mana yang akan dimasukkan ke dalam model dan apakah transformasi atau rekayasa fitur mungkin bermanfaat.

Uji Statistika

Tabel 4 menyajikan hasil dari analisis regresi dengan menggunakan metode OLS. Hasil ini mengevaluasi hubungan antara variabel-variabel independen dengan harga rumah sebagai variabel dependen. *R-squared* dan *Adjusted R-squared* menunjukkan bahwa fitur-fitur dalam model menjelaskan sebesar 91,8% variasi dari harga rumah. Proporsi ini sangat tinggi, menandakan bahwa model memiliki kesesuaian yang baik dengan data.

Setiap kenaikan 1-unit pada *Avg. Area Income* dapat meningkatkan harga rumah sebesar \$21,58 dengan asumsi variabel lain tetap. Nilai *p* mendekati 0, menandakan bahwa fitur *Avg. Area Income* memiliki nilai statistik yang signifikan terhadap harga rumah.

Avg. Area House Age menunjukkan untuk setiap kenaikan satu tahun pada *Avg. Area House Age* menyebabkan harga rumah meningkat sebesar \$165.700. Nilai *p* untuk fitur ini juga 0, yang artinya fitur ini memiliki nilai statistik yang signifikan terhadap harga rumah.

Selanjutnya adalah *Avg. Area Number of Rooms*, untuk setiap penambahan satu ruangan, harga rumah meningkat sebesar \$121.600. Fitur ini juga menunjukkan signifikansi statistik dengan nilai *p* sama dengan 0 yang artinya fitur ini signifikan berpengaruh terhadap harga rumah.

Area Population, untuk setiap kenaikan satu unit pada *Area Population*, harga rumah meningkat sebesar \$15,20 dengan asumsi variabel lain tetap. Fitur ini juga memiliki signifikansi statistik.

Tabel 3. Uji Signifikansi Statistika

Dep. Variabel	Price		R-squared	0,918		
Model	OLS		Adj. R-squared	0,918		
Method	Least Squares		F-statistic	1,398e ⁴		
No Observations	5000		Prob (F-statistic)	0		
Df Residuals:	4995		Log-Likelihood	-64714		
Df Model:	4		AIC	1,294e ⁴		
Covariance Type:	nonrobust		BIC	1,294e ⁴		
	coef	std error	t	P> t	[0,025	0,975]
Const	2,638e ⁶	1,72e ⁴	-153,726	0,000	-2,67e ⁶	-2,6e ⁶
Avg. Area Income	21,5827	0,134	160,743	0,000	21,320	21,846
Avg. Area House Age	1,657e ⁵	1443,404	114,769	0,000	1,63e ⁵	1,68e ⁵

Avg. Area Number of Rooms	1,216e ⁵	1422,608	85,476	0,000	1,19e ⁵	1,24e ⁵
Area Population	15,1961	0,144	105,388	0,000	14,913	15,479
Omnibus		5,310	Durbin-Watson		2,006	
Prob (Omnibus)		0,070	Jarque-Bera (JB)		4,742	
Skew		0,011	Prob (JB)		0,0934	
Kurtosis		2,851	Cond. No.		9,40e ⁵	

Sumber: Olah Data, 2023

Nilai uji F ($1,398e^4$) menunjukkan nilai yang sangat tinggi, dengan nilai p yang mendekati 0, hal tersebut dapat diinterpretasikan bahwa model secara keseluruhan memiliki signifikansi statistik. Statistik Durbin-Watson mendekati 2, mengindikasikan bahwa tidak ada autokorelasi signifikan pada residu. Uji Omnibus (0,070) dan Jarque-Bera (0,0934) menunjukkan bahwa residu memiliki distribusi yang mendekati normal, meskipun nilai p-nya sedikit lebih tinggi dari level signifikansi umumnya yakni sebesar 0,05. Walaupun kurtosis menunjukkan nilai yang tinggi yakni 2,851 yang mungkin mengindikasikan multikolenieritas. Namun, dari *heatmap* yang telah dianalisis sebelumnya, diketahui bahwa multikolenieritas bukanlah masalah signifikan bagi fitur-fitur yang digunakan dalam penelitian ini.

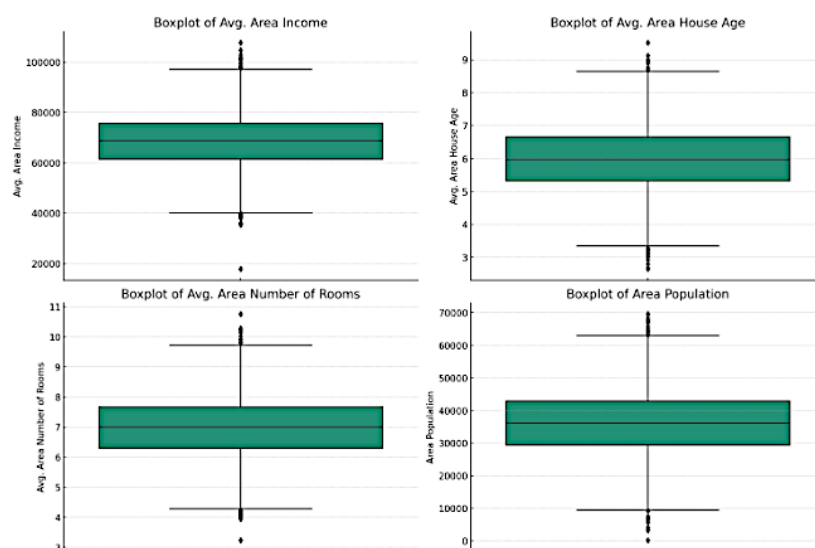
Outlier Detection

Dalam tahap eksplorasi data, penting untuk mendeteksi dan memahami keberadaan *outlier* karena hal ini dapat mempengaruhi performa model regresi. Untuk mendeteksi *outlier*, metode *Interquartile Range* (IQR) diaplikasikan pada setiap fitur. Hasil Deteksi *Outlier* dengan Metode IQR:

Avg. Area Income: Terdapat 32 data yang diidentifikasi sebagai *outlier* potensial;
Avg. Area House Age di Area: Teridentifikasi 25 data sebagai *outlier* potensial;
Avg. Area Number of Rooms di Area: 24 data diidentifikasi sebagai *outlier* potensial;

Area Population: Terdapat 30 data yang merupakan *outlier* potensial.

Visualisasi *Outlier* dengan *Boxplot*:



Gambar 4. *Boxplot* Fitur

Sumber: Olah Data, 2023

Bagian kotak tengah pada setiap plot merepresentasikan IQR, yang mengandung 50% data median. Garis *Whiskers* memanjang hingga 1.5 kali IQR. Poin individu di luar *Whiskers* pada setiap plot adalah *outlier* potensial yang sebelumnya terdeteksi dengan metode IQR.

Dari visualisasi tersebut, dapat diamati bahwa *outlier* potensial tidak memiliki nilai yang ekstrem, menunjukkan kemungkinan bahwa nilai-nilai tersebut adalah data yang sah dan bukan kesalahan. Selanjutnya proses pembatasan (*Capping/Flooring*) dengan mengganti *outlier* dengan nilai *Whisker* atas/bawah dari *boxplot* (yaitu $Q1 - 1.5 \times IQR$ untuk *outlier* bawah dan $Q3 + 1.5 \times IQR$ untuk *outlier* atas) dilakukan.

Data Splitting

Dataset dibagi menjadi set pelatihan dan pengujian dengan rasio 80:20 untuk memastikan model dapat dievaluasi dengan benar pada data yang belum pernah dilihatnya sebelumnya.

Model Training dan Evaluasi

Dengan dataset yang telah diproses, empat variasi model SVM, yang masing-masing menggunakan kernel Linear, Polynomial, RBF, dan Sigmoid, dilatih menggunakan data latih. Evaluasi awal menunjukkan bahwa model SVM dengan kernel RBF menonjol dalam kinerjanya, model SVM dengan kernel RBF dapat menjelaskan 70,78% variansi pada set pengujian. Menariknya, meski kurang konvensional untuk tugas regresi, kernel Sigmoid juga menunjukkan kinerja yang mengesankan dengan koefisien determinasi R^2 0,8079.

Tabel 4. Matrik Evaluasi

Kernel	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	R^2
Linear	\$254.690,59	\$82.768.324.589,89	0,4876
Polynomial	\$279.783,66	\$124.167.293.010,47	-0,0092
RBF	\$141.561,12	\$35.949.612.608,83	0,7078
Sigmoid	\$121.177,74	\$23.629.516.298,89	0,8079

Sumber: Olah Data, 2023

Hasil uji pada Tabel 4 juga menunjukan SVM dengan kernel Linear dan Polynomial gagal menunjukan performa yang baik dalam memprediksi harga rumah berdasarkan fitur-fitur yang telah digunakan.

KESIMPULAN

Dalam penelitian ini, dilakukan analisis mendalam terhadap dataset properti di Amerika Serikat untuk memprediksi harga rumah berdasarkan fitur-fitur yang tersedia. Melalui eksplorasi awal, dapat dipastikan bahwa dataset tidak memiliki nilai yang hilang dan mengandung beberapa fitur penting yang berpotensi mempengaruhi harga rumah. Analisis visual menunjukkan adanya korelasi positif antara sejumlah fitur dengan harga rumah. Selain itu, melalui *heatmap*, ditemukan hubungan korelasi yang signifikan antara fitur-fitur seperti *Avg. Area Income*, *Avg. Area Number of Rooms*, *Area Population*, dan *Avg. Area House Age* dengan harga rumah.

Selanjutnya, dilakukan pendeteksian *outlier* untuk memastikan kualitas data. Meskipun beberapa *outlier* terdeteksi, namun analisis lebih lanjut menunjukkan bahwa nilai-nilai ini mungkin merupakan data point yang sah.

Dalam implementasi model, algoritma SVM dengan beberapa kernel berbeda diuji coba. Hasil menunjukkan bahwa kernel RBF memberikan performa terbaik dalam memprediksi harga rumah, dengan menjelaskan 70,78% variasi data.

Sebagai kesimpulan, meskipun ada beberapa tantangan dalam analisis data, seperti keberadaan *outlier*, namun dengan pendekatan yang tepat dan pemilihan algoritma yang sesuai, model prediksi harga rumah dapat dikembangkan dengan akurasi yang memuaskan. Hasil penelitian ini dapat dijadikan sebagai dasar untuk pengembangan lebih lanjut dalam bidang prediksi properti atau aplikasi praktis dalam industri properti.

DAFTAR PUSTAKA

- Alghushairy, O., Alsini, R., Soule, T., & Ma, X. (2021). A review of local *outlier* factor algorithms for *outlier* detection in big data streams. *Big Data and Cognitive Computing*, 5(1), 1–24. <https://doi.org/10.3390/bdcc5010001>
- Alita, D., Fernando, Y., & Sulistiani, H. (2020). Implementasi Algoritma Multiclass Svm Pada Opini Publik Berbahasa Indonesia Di Twitter. *Jurnal Tekno Kompak*, 14(2), 86. <https://doi.org/10.33365/jtk.v14i2.792>
- Azizah, N. (2017). Nur Azizah, 2017 Implementasi D An Analisa Waktu Komputasi Pad A Algoritma Rand Om Forest D Engan Parallel Computing D I R Universitas Pendidikan Indonesia | repository.upi.edu | perpustakaan.upi.edu. 1–5.
- Baak, M., Koopman, R., Snoek, H., & Klous, S. (2020). A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics. *Computational Statistics and Data Analysis*, 152, 107043. <https://doi.org/10.1016/j.csda.2020.107043>
- Bank BTN. (2021). Riset: Harga Rumah Naik karena Permintaan Meningkat di Masa Pandemi. <https://www.btn.co.id/en/Conventional/Informasi-yang-Anda-Butuhkan-Saat-Ini/Info/Artikel---Index-Harga-Rumah-HPI>
- El Morr, C., Jammal, M., Ali-Hassan, H., & El-Hallak, W. (2022). Support Vector Machine. *International Series in Operations Research and Management Science*, 334, 385–411. https://doi.org/10.1007/978-3-031-16990-8_13
- EMOTIVEPRO. (2019). Raw EEG. https://emotiv.gitbook.io/emotivpro/data_streams/raw_eeg
- Fondy, H. W., Fajar, M., & Musdar, I. A. (2019). Implementasi Teori Support Vector Machine Untuk Memprediksi Harga Penjualan Laptop Asus. *KHARISMA Tech*, 1–9. <https://jurnal.kharisma.ac.id/kharismatech/article/view/238%0Ahttps://jurnal.kharisma.ac.id/kharismatech/article/download/238/132>
- Huang, K., & Putra Purnoma, E. (2022). Support Vector Machine Algorithm. <https://sis.binus.ac.id/2022/02/14/support-vector-machine-algorithm/>

- Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of Performance of Data Imputation Methods for Numeric Dataset. *Applied Artificial Intelligence*, 33(10), 913–933. <https://doi.org/10.1080/08839514.2019.1637138>
- Kementerian PUPR. (2019). *Kementerian PUPR Siapkan Skema Penyediaan Rumah Bagi Generasi Milenial*. <https://pu.go.id/berita/kementerian-pupr-siapkan-skema-penyediaan-rumah-bagi-generasi-milenial>
- Lumbanraja, F. R., Saputra, R. A., Muludi, K., Hijriani, A., & Junaidi, A. (2021). Implementasi Support Vector Machine Dalam Memprediksi Harga Rumah Pada Perumahan Di Kota Bandar Lampung. *Jurnal Pepadun*, 2(3), 327–335. <https://doi.org/10.23960/pepadun.v2i3.90>
- Nasution, D. A., Khotimah, H. H., & Chamidah, N. (2019). Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN. *Computer Engineering, Science and System Journal*, 4(1), 78. <https://doi.org/10.24114/cess.v4i1.11458>
- Ni Komang Sariasih, Komang Fridagustina Adnantara, & Laras Oktaviani. (2021). Pengaruh Rasio Likuiditas, Solvabilitas, Profitabilitas Dan Aktivitas, Terhadap Kebijakan Dividen Pada Perusahaan Property, Real Estate, and Building. *Journal Research of Accounting*, 3(1), 76–92. <https://doi.org/10.51713/jarac.v3i1.46>
- Pedregosa, Fabian Varoquaux, G., Gramfort, A., & Michel, V. (2011). *Univariate feature selection*. https://scikit-learn.org/stable/modules/feature_selection.html#univariate-feature-selection
- Prabhu, D. (2023). *No Title*. <https://www.kaggle.com/datasets/darshanprabhu09/housing-dataset-of-5000-people-staying-in-usa>
- Quang, T., Minh, N., Hy, D., & Bo, M. (2020). Housing Price Prediction via Improved Machine Learning Techniques. *Procedia Computer Science*, 174(2019), 433–442. <https://doi.org/10.1016/j.procs.2020.06.111>
- Raihan-Al-Masud, M., & Rubaiyat Hossain Mondal, M. (2020). Data-driven diagnosis of spinal abnormalities using feature selection and machine learning algorithms. *PLoS ONE*, 15(2), 1–21. <https://doi.org/10.1371/journal.pone.0228422>
- Raja, P. S., & Thangavel, K. (2020). Missing value imputation using unsupervised machine learning techniques. In *Soft Computing* (Vol. 24, Issue 6). Springer Berlin Heidelberg. <https://doi.org/10.1007/s00500-019-04199-6>
- Ruhleassin, M. F. (2022). *Walau Harga Rumah Terus Naik, Bubble Properti Diprediksi Tak Akan Terjadi*. <https://www.kompas.com/properti/read/2022/02/18/090000421/walau-harga-rumah-terus-naik-bubble-properti-diprediksi-tak-akan?page=all>
- SMF IT. (2017). *Faktor-Faktor Penting yang Menyebabkan Harga Rumah Kian Melambung*. <https://www.hfis-smf.co.id/2017/11/faktor-faktor-penting-yang->

menyebabkan-harga-rumah-kian-melambung/

- Wahanani, H. E., Prami Swari, M. H., & Akbar, F. A. (2020). Case based Reasoning Prediksi Waktu Studi Mahasiswa Menggunakan Metode Euclidean Distance dan Normalisasi Min-Max. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, 7(6), 1279. <https://doi.org/10.25126/jtiik.2020763880>
- Zhang, Y., Huang, J., Zhang, J., Liu, S., & Shorman, S. (2022). Analysis and prediction of second-hand house price based on random forest. *Applied Mathematics and Nonlinear Sciences*, 7(1), 27–42. <https://doi.org/10.2478/amns.2022.1.00052>