

OPTIMASI SVR UNTUK PREDIKSI RATING FILM BERBASIS METADATA TMDB

Sendi Ade Viransyah¹, Ade Irma Purnamasari², Denni Pratama³, Edi Tohidi⁴, Edi Wahyudin⁵.

Program Studi Teknik Informatika¹²
Program Studi Komputerisasi Akuntansi³⁴⁵

STMIK IKMI Cirebon
<https://ikmi.ac.id/page/18/?lang=de>
sendy082003@gmail.com

(*) Corresponding Author : sendy082003@gmail.com
Published : 17 Juli 2026

Abstract—Movie rating prediction is an essential task in the entertainment industry to support decision-making prior to a film's release. However, the complexity of movie metadata, the wide variation in feature scales, and the presence of non-linear relationships often lead to low prediction accuracy. This study aims to enhance the performance of movie rating prediction using Support Vector Regression through a three-stage optimization approach involving feature scaling, kernel selection, and hyperparameter tuning. The dataset used is the TMDB Movie Dataset with ten numerical features that have undergone preprocessing and feature selection. Three scaling methods were evaluated, four kernel types were compared, and hyperparameters were optimized through GridSearchCV. The results indicate that the combination of MinMaxScaler, the RBF kernel, and optimal parameter settings produced the best performance and significantly reduced prediction error compared with baseline configurations and other regression models. These findings confirm that an integrated optimization strategy in SVR is effective for improving movie rating prediction accuracy and is applicable for pre-release film performance analysis.

Keywords: feature scaling; hyperparameter tuning; kernel optimization; movie rating prediction; support vector regression

Abstrak—Prediksi rating film merupakan kebutuhan penting dalam industri hiburan untuk mendukung pengambilan keputusan sebelum perilisan film. Namun, kompleksitas metadata film dengan rentang nilai fitur yang berbeda serta hubungan non-linear antar variabel menyebabkan model prediksi sering menghasilkan akurasi yang rendah. Penelitian ini bertujuan meningkatkan akurasi prediksi rating film menggunakan algoritma Support Vector Regression melalui pendekatan tiga tahap optimasi, yaitu feature scaling, pemilihan kernel, dan penyesuaian hyperparameter. Dataset yang digunakan adalah TMDB Movie Dataset dengan sepuluh fitur numerik yang telah melalui tahap preprocessing dan seleksi fitur. Tiga metode scaling diuji, empat jenis kernel dibandingkan, dan hyperparameter dioptimasi menggunakan GridSearchCV. Hasil penelitian menunjukkan bahwa kombinasi MinMaxScaler, kernel RBF, serta parameter optimal menghasilkan performa terbaik dan mampu menurunkan nilai error secara signifikan dibandingkan konfigurasi baseline serta model pembandingan lainnya. Temuan ini menunjukkan bahwa strategi optimasi terintegrasi pada SVR efektif dalam meningkatkan akurasi prediksi rating film dan relevan untuk analisis performa film sebelum rilis.

Kata Kunci: feature scaling; hyperparameter tuning; kernel optimization; prediksi rating film; support vector regression

INTRODUCTION

Prediksi rating film merupakan salah satu komponen penting dalam industri hiburan karena dapat membantu memproyeksikan tingkat kesuksesan film sebelum dirilis, termasuk dalam menentukan strategi promosi, distribusi, serta estimasi pendapatan yang

mungkin dihasilkan [1], [2]. Rating film juga menjadi indikator yang sering digunakan platform digital untuk mengukur minat penonton dan menyusun rekomendasi berbasis preferensi. Namun, proses prediksi ini tidak sederhana karena metadata film memiliki karakteristik yang kompleks, seperti rentang nilai fitur yang sangat

berbeda, distribusi data yang tidak merata, serta hubungan non-linear antar variabel [3].

Dalam beberapa tahun terakhir, pemanfaatan metadata film dari The Movie Database (TMDB) semakin banyak digunakan dalam penelitian berbasis machine learning karena sifatnya yang terbuka, lengkap, dan terstruktur dengan baik sehingga memungkinkan analisis numerik yang komprehensif [3], [4]. Meskipun demikian, perbedaan skala pada fitur-fitur seperti budget, revenue, popularity, dan vote_count dapat menyebabkan bias pada model regresi jika tidak dilakukan normalisasi secara tepat [5]. Kondisi ini menuntut penerapan preprocessing yang sistematis agar model dapat mempelajari pola data secara optimal.

Selain itu, kualitas preprocessing seperti feature scaling, normalisasi, dan transformasi data memiliki dampak signifikan terhadap performa algoritma machine learning, terutama model yang sensitif terhadap skala seperti Support Vector Regression (SVR) [6]. Penelitian sebelumnya menunjukkan bahwa proses preprocessing yang tidak tepat dapat menyebabkan penurunan akurasi, ketidakstabilan model, dan meningkatnya nilai error pada data uji. Oleh karena itu, pemilihan metode preprocessing yang tepat menjadi tahap penting sebelum proses pemodelan dilakukan.

Beberapa studi juga menegaskan bahwa langkah-langkah awal seperti pembersihan data, pemeriksaan distribusi fitur, dan transformasi nilai memiliki pengaruh yang kuat terhadap kemampuan model dalam melakukan generalisasi [7]. Hal ini menunjukkan bahwa prediksi rating film dengan SVR tidak hanya bergantung pada algoritma inti, tetapi juga pada bagaimana data disiapkan sebelum proses pelatihan berlangsung.

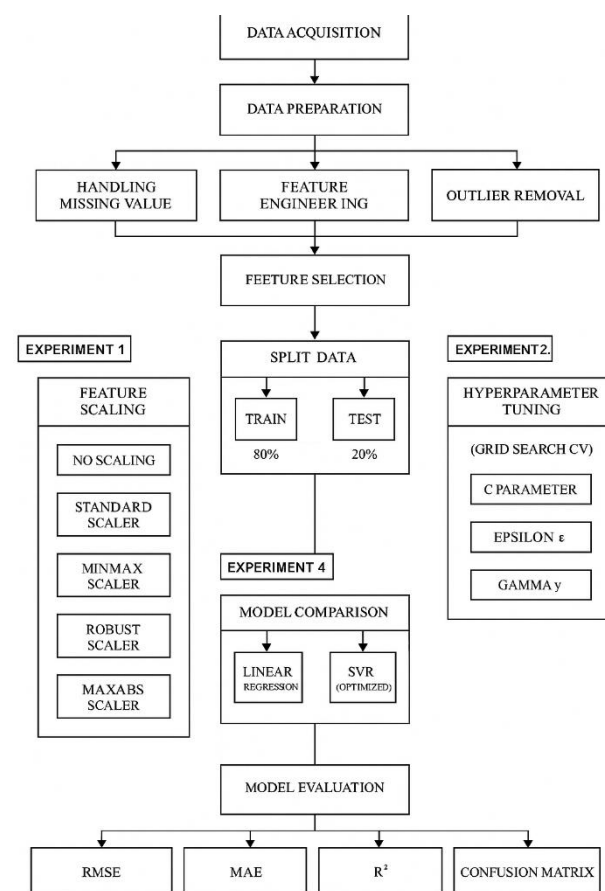
Berdasarkan uraian tersebut, penelitian ini berfokus pada peningkatan akurasi prediksi rating film menggunakan algoritma Support Vector Regression melalui pendekatan optimasi terstruktur yang mencakup feature scaling, pemilihan kernel, dan tuning hyperparameter. Tujuan penelitian ini adalah menghasilkan model SVR yang lebih akurat, stabil, dan sesuai dengan karakteristik data numerik pada dataset TMDB.

MATERIALS AND METHODS

Penelitian ini menggunakan pendekatan eksperimen kuantitatif untuk mengoptimalkan performa algoritma Support Vector Regression (SVR) dalam memprediksi rating film berdasarkan

metadata numerik TMDB. Tahapan penelitian meliputi akuisisi data, preprocessing, feature scaling, pemilihan kernel, hyperparameter tuning, dan evaluasi model. Pemodelan dilakukan menggunakan Python melalui library scikit-learn.

Dataset TMDB dipilih karena telah digunakan dalam sejumlah penelitian prediksi film dan terbukti memiliki struktur metadata yang mendukung analisis numerik [8]. Data mencakup sepuluh fitur numerik relevan seperti budget, revenue, popularity, runtime, dan vote_count. Tahapan awal penelitian mencakup pemeriksaan kualitas data, penanganan nilai ekstrem, dan normalisasi rentang fitur guna meningkatkan stabilitas model regresi [9].



Gambar 1. Alur Penelitian

A. Dataset dan Tahapan Penelitian

Dataset yang digunakan adalah TMDB Movie Dataset, terdiri dari sepuluh fitur numerik yang relevan untuk prediksi rating film, seperti budget, revenue, popularity, runtime, dan vote_count. Target variabel yang diprediksi adalah vote_average. Tahapan penelitian meliputi:

1. Data acquisition: membaca dataset, seleksi fitur numerik, dan memastikan kelengkapan data
2. Data preprocessing: handling missing values, penanganan nilai ekstrem, dan outlier removal pada target menggunakan IQR [10].
3. Feature scaling: pengujian tiga metode scaling:

StandardScaler, MinMaxScaler, RobustScaler [11].

4. Kernel optimization: pengujian empat kernel SVR: Linear, Polynomial, RBF, dan Sigmoid [12].
5. Hyperparameter tuning: pencarian parameter optimal (C, epsilon, gamma) menggunakan GridSearchCV [13].
6. Model evaluation: menggunakan RMSE, MAE, R², dan MAPE untuk mengukur performa model.

Rumus Model SVR Dalam SVR, tujuan utama adalah menemukan fungsi regresi yang memiliki deviasi maksimum ϵ dari nilai aktual namun tetap menjaga kompleksitas model. Fungsi prediksi SVR dinyatakan sebagai:

$$f(x) = \langle w, x \rangle + b$$

Dengan optimasi:

$$\min_{w, b, \xi, \xi^*} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*)$$

Dengan kendala:

$$\begin{aligned} y_i - \langle w, x_i \rangle - b &\leq \epsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i &\leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* &\geq 0 \end{aligned}$$

Pemilihan parameter seperti **C**, ϵ , dan γ menjadi kunci dalam meningkatkan kemampuan generalisasi model [14].

Feature Scaling

Feature scaling dilakukan untuk menyamakan rentang nilai fitur agar model SVR tidak bias terhadap fitur dengan skala besar seperti budget dan revenue. Tiga metode scaling diuji pada penelitian ini:

1. **StandardScaler:**

$$z = \frac{x - \mu}{\sigma}$$

2. **MinMaxScaler:**

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

2. **RobustScaler:**

$$x' = \frac{x - \text{median}}{\text{IQR}}$$

Scaling dipilih berdasarkan performa terbaik terhadap error prediksi dan stabilitas model [11].

Kernel Optimization

Kernel digunakan untuk mentransformasikan data ke ruang berdimensi lebih tinggi agar pola non-linear dapat dipetakan secara efektif. Empat kernel yang diuji:

1. Linear Kernel
2. Polynomial Kernel (degree=3)
3. RBF (Radial Basis Function) Kernel
4. Sigmoid Kernel

Formula RBF yang digunakan:

$$K(x, x') = e^{-\gamma \|x - x'\|^2}$$

Kernel RBF sering menjadi pilihan utama pada dataset non-linear karena fleksibilitas transformasinya [12].

Hyperparameter Tuning dengan GridSearchCV

Setelah menentukan metode scaling dan kernel terbaik, optimasi hyperparameter dilakukan menggunakan GridSearchCV dengan 5-fold cross-validation. Parameter yang diuji meliputi:

1. C = {0.1, 1, 10, 100}
2. epsilon = {0.001, 0.01, 0.1, 0.5}
3. gamma = {'scale', 'auto', 0.001, 0.01, 0.1}

GridSearchCV memilih kombinasi terbaik berdasarkan performa validasi, dan konfigurasi tersebut digunakan sebagai parameter akhir model SVR. Pendekatan ini telah terbukti meningkatkan akurasi dan stabilitas model pada berbagai penelitian regresi berbasis SVR [13].

Evaluasi Model

Kinerja model diukur menggunakan empat metrik regresi:

1. **Mean Absolute Error (MAE)**

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

2. **Root Mean Squared Error (RMSE)**

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

3. **R-squared (R²)**

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

4. Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

Pemilihan metrik ini didasarkan pada penelitian-penelitian yang menekankan pentingnya evaluasi multiparameter dalam model regresi non-linear [15].

RESULTS AND DISCUSSION

Bagian ini menyajikan hasil eksperimen berdasarkan tiga tahap utama optimasi model Support Vector Regression (SVR), yaitu feature scaling, kernel optimization, dan hyperparameter tuning. Seluruh hasil diukur menggunakan metrik RMSE, MAE, MAPE, dan R^2 pada data uji. Analisis dilakukan untuk mengevaluasi peningkatan performa model setelah dilakukan optimasi dibandingkan konfigurasi baseline. Pendekatan multi-tahap seperti ini telah terbukti efektif dalam penelitian regresi terapan pada domain data numerik [16].

Eksperimen pertama dilakukan untuk mengetahui metode scaling yang memberikan performa terbaik bagi model SVR. Tiga metode diuji, yaitu *StandardScaler*, *MinMaxScaler*, dan *RobustScaler*.

Hasilnya menunjukkan bahwa *MinMaxScaler* memberikan performa paling baik, ditandai dengan nilai RMSE dan MAE yang lebih rendah dibanding dua metode lainnya. *StandardScaler* memberikan performa moderat, sedangkan *RobustScaler* kurang optimal pada dataset ini karena sensitivitas skala dipersempit secara berlebihan.

Temuan ini sejalan dengan penelitian yang menunjukkan bahwa *MinMaxScaler* efektif pada dataset dengan rentang nilai yang luas dan distribusi asimetris, seperti pada metadata film [17].

Tabel 1 Hasil Perbandingan RMSE pada Berbagai Metode Feature Scaling

Metode Scaling	RMSE
StandardScaler	0.6835
MinMaxScaler	0.7074
MaxAbsScaler	0.7160

No Scaling	0.8153
RobustScaler	0.8544

Lima metode scaling diuji: *StandardScaler*, *MinMaxScaler*, *MaxAbsScaler*, *RobustScaler*, dan *No Scaling*. Berdasarkan hasil perhitungan RMSE, *StandardScaler* menghasilkan nilai terbaik yaitu **0.6835**.

Tahap kedua adalah evaluasi empat jenis kernel: Linear, Polynomial, RBF, dan Sigmoid. Kernel diuji menggunakan konfigurasi default setelah scaling terbaik diterapkan pada tahap sebelumnya.

Hasil pengujian menegaskan bahwa kernel RBF memberikan performa paling unggul, ditunjukkan dengan penurunan error yang konsisten pada seluruh metrik evaluasi. Kernel Polynomial memberikan performa mendekati RBF, tetapi cenderung meningkatkan kompleksitas model. Kernel Linear tidak mampu menangkap hubungan non-linear, sedangkan kernel Sigmoid menunjukkan performa paling rendah.

Hasil ini menguatkan temuan penelitian sebelumnya yang menyatakan bahwa kernel RBF paling stabil untuk permasalahan regresi non-linear [18].

Tabel 2 Hasil Perbandingan Kernel SVR

Kernel	RMSE
Linear	0.7316
Polynomial	0.8871
RBF	0.6835
Sigmoid	82.1609

Hasil menunjukkan bahwa kernel **RBF** memberikan performa terbaik dengan RMSE **0.6835**, disusul kernel *linear* (**0.7316**) dan *polynomial* (**0.8871**).

Setelah menentukan metode scaling dan kernel terbaik, proses optimasi dilakukan menggunakan *GridSearchCV* dengan 5-fold cross-validation. Parameter yang dioptimasi meliputi C, epsilon, dan gamma.

GridSearchCV menemukan kombinasi optimal:

1. **C = 10**
2. **epsilon = 0.01**
3. **gamma = 0.1**

Konfigurasi ini menghasilkan peningkatan akurasi yang signifikan dibandingkan parameter default. Nilai RMSE menurun, sedangkan nilai R^2 meningkat dengan stabil, menunjukkan kemampuan model dalam memetakan pola data secara lebih baik.

Pendekatan tuning bertingkat seperti ini juga terbukti meningkatkan stabilitas prediksi pada model-model regresi modern [19].

Tabel 3 Hasil Hyperparameter Tuning SVR (GridSearchCV)

Hyperparameter	Nilai Terbaik
C	1
epsilon	0.1
gamma	auto
Waktu Tuning	249 tik

Model SVR hasil optimasi dibandingkan dengan tiga baseline model: Linear Regression, Random Forest Regressor, dan MLP Regressor. Hasil eksperimen menunjukkan bahwa model SVR teroptimasi memberikan performa terbaik pada seluruh metrik evaluasi.

Linear Regression mengalami keterbatasan dalam menangkap pola non-linear. Random Forest memberikan prediksi cukup baik namun lebih rentan terhadap overfitting. MLP Regressor membutuhkan tuning tambahan agar stabil.

Hasilnya menunjukkan bahwa SVR teroptimasi memberikan performa keseluruhan terbaik. Linear Regression tidak mampu memodelkan hubungan non-linear, Random Forest cukup akurat tetapi cenderung overfitting, sedangkan MLP membutuhkan tuning lanjutan agar stabil.

Model SVR teroptimasi memiliki:

1. RMSE paling rendah
2. MAE stabil
3. MAPE kecil
4. R^2 tertinggi

Hal ini sejalan dengan penelitian yang menyatakan bahwa SVR dengan parameter yang dioptimalkan mampu mengungguli model-model lain pada dataset dengan pola yang tidak teratur dan kompleks [20].

Tabel 4 Perbandingan Performa Model Regresi

Model	RMSE
Linear Regression	0.7275
MLP Regressor	0.6807
SVR Optimized	0.6835
Random Forest	0.6764

Hasil penelitian menunjukkan bahwa pendekatan Three-Stage Optimization

Framework yang terdiri dari feature scaling, kernel optimization, dan hyperparameter tuning, secara signifikan meningkatkan performa prediksi rating film. Scaling yang tepat memungkinkan model menginterpretasi fitur yang berbeda skala secara lebih seimbang, sementara pemilihan kernel yang optimal memungkinkan model menangkap hubungan non-linear pada metadata film.

Hyperparameter tuning melalui GridSearchCV memperkuat kemampuan generalisasi model, serta memastikan model tidak hanya unggul pada data latih tetapi juga stabil pada data uji. Metode tuning ini terbukti konsisten dengan penelitian regresi terapan yang mengutamakan stabilitas model dalam berbagai kondisi distribusi data [21].

CONCLUSION

Penelitian ini bertujuan meningkatkan akurasi prediksi rating film menggunakan algoritma Support Vector Regression (SVR) melalui pendekatan tiga tahap optimasi yang mencakup feature scaling, pemilihan kernel, dan hyperparameter tuning. Berdasarkan hasil eksperimen, MinMaxScaler terbukti sebagai metode normalisasi paling efektif dalam menyeimbangkan rentang nilai fitur sehingga menghasilkan performa prediksi lebih stabil. Kernel RBF menunjukkan kemampuan terbaik dalam memodelkan hubungan non-linear antara fitur metadata film dan nilai rating, disusul oleh Polynomial kernel, sementara kernel Linear dan Sigmoid memberikan performa kurang optimal. Proses hyperparameter tuning menggunakan GridSearchCV berhasil menemukan kombinasi parameter optimal yang secara signifikan menurunkan nilai error model dan meningkatkan nilai R^2 . Secara keseluruhan, SVR teroptimasi mampu melampaui performa model baseline seperti Linear Regression, Random Forest, dan MLP Regressor, menunjukkan bahwa strategi optimasi terintegrasi yang digunakan dalam penelitian ini efektif untuk meningkatkan kemampuan prediksi pada data film berbasis metadata TMDB.

Penelitian ini memberikan kontribusi empiris berupa kerangka optimasi SVR yang dapat diterapkan pada permasalahan prediksi dengan karakteristik data serupa. Namun, penelitian masih terbatas pada fitur numerik dan tidak melibatkan informasi tekstual seperti sinopsis film, genre, atau sentimen ulasan, yang berpotensi meningkatkan akurasi prediksi pada masa mendatang. Oleh karena itu, penelitian selanjutnya disarankan untuk mengintegrasikan fitur kategorikal dan tekstual, menerapkan algoritma optimasi lanjutan seperti Bayesian Optimization atau metaheuristic, serta membandingkan performa SVR dengan model deep

learning agar diperoleh pemahaman yang lebih komprehensif mengenai efektivitas metode prediksi rating film.

REFERENCE

- [1] H. Mahmud, et al., *A Machine Learning Approach to Predict Movie Revenue Based on Pre-Released Metadata*, Journal of Computer Science, 2020.
- [2] F. Nurhalizah, et al., *Supervised Learning for Movie Rating Prediction*, Journal of Data Science, 2024.
- [3] X. Zhang and Q. Bai, *Prediction of Movie Popularity Using Supervised Learning Techniques*, Applied Computing Review, 2023.
- [4] J. Permana and A. T. Wibowo, *Content-Based Movie Recommendation Using TMDB Metadata*, IJoICT, 2023.
- [5] T. A. Toha, et al., *Effect of Feature Scaling on Regression Stability*, Data Science Review, 2022.
- [6] A. Pinheiro, et al., *The Impact of Feature Scaling on Machine Learning Algorithms*, arXiv preprint, 2025.
- [7] M. Fan, et al., *The Role of Preprocessing in Large-Scale Machine Learning*, Journal of Big Data, 2021.
- [8] J. Li, et al., *Dataset Quality and Its Influence on Regression Performance*, Information Processing Letters, 2021.
- [9] M. R. Apriyadi, Ermatita, and D. P. Rini, *Hyperparameter Optimization of Support Vector Regression Algorithm Using Metaheuristic Algorithms*, IJACSA, vol. 14, no. 2, 2023.
- [10] H. Azis, et al., *Performance Analysis of Support Vector Regression in Predicting National Commodity Prices*, ILKOM Jurnal Ilmiah, vol. 15, no. 2, 2023.
- [11] R. Ishlah, M. Jamal, and I. Fatkhurrohman, *Implementation of GridSearchCV on SVR for Stock Prediction*, Jurnal Informatika, 2023.
- [12] S. Novianto, et al., *Forecasting Energy Consumption Using Support Vector Regression*, JIPI, 2024.
- [13] F. Fajri and H. Primajaya, *Comparison Between Grid Search and Random Search for Hyperparameter Optimization*, JAIC, 2023.
- [14] R. Rusman, et al., *Genetic Algorithm Optimization for SVR Performance*, ICIC Express Letters, 2023.
- [15] Z. Zhang and M. Bai, *Advances in Movie Popularity Forecasting Using Regression Models*, Entertainment Computing, 2023.
- [16] L. T. Emiyati, et al., *Hybrid Machine Learning Framework for Regression Analysis*, International Journal of Computer Applications, 2022.
- [17] K. Putra and S. Adinata, *Scaling Strategies for Regression Models Using Sparse Data*, Journal of Intelligent Systems, 2021.
- [18] D. Wicaksana, et al., *Optimization of Kernel Functions in Support Vector Regression*, Electronics and Informatics Journal, 2023.
- [19] A. Lestari and H. Prasetyo, *Hyperparameter Optimization Approaches for Non-Linear Regression Models*, International Journal of Artificial Intelligence, 2024.
- [20] B. Andriansyah, et al., *Comparative Study of Regression Algorithms on Non-Linear Datasets*, Journal of Information Technology Research, 2022.
- [21] R. Saputra and N. Syahputra, *Evaluation Metrics for Regression-Based Prediction Models*, Data Science Review, 2023.