JURNAL FASILKOM
Volume 13 No. 02 | Agustus 2023: 104-111

P-ISSN : 2089-3353
E-ISSN : 2808-9162

Perbandingan Random Forest Regression dan Support Vector Regression Pada Prediksi Laju Penguapan

Ferdinandus Edwin Penalun[1], Arief Hermawan[2], Donny Avianto[3]
[123] Program Studi Teknologi Informasi, Magister Teknologi Informasi, Universitas Teknologi Yogyakarta
[1]6220211008.edwin@student.uty.ac.id, [2]ariefdb@uty.ac.id*, [3]donny@uty.ac.id*

Abstract

Predicting evaporation rates has wide-ranging benefits in applications such as water resources management, agriculture, and the environment. However, obtaining complete and accurate data to study evaporation rates is challenging. In addition, the low degree of linearity between evaporation rate data and other meteorological factors in tropical regions can lead to variable prediction results. The purpose of this study is to predict the daily evaporation rate at Yogyakarta Climatology Station by comparing the performance of two machine learning (ML) models, namely random forest regression (RFR) and support vector regression (SVR) using daily meteorological observation data. To improve prediction accuracy, hyperparameter optimization is performed using the gridsearch cross-validation method to find the best hyperparameter combination. Hyperparameter optimization results on training data show that the RFR model produces an RMSE score of -0.67 while the SVR model on the RBF kernel produces a negative RMSE score of -0.57. Further evaluation is carried out on testing data using a combination of hyperparameter optimization results of the RFR model produces an R2 value of 0.79 and an RMSE of 0.56 while the SVR model produces a coefficient of determination (R2) of 0.81 and RMSE of 0.53. Based on the comparison of the two models, it can be concluded that the SVR model has better performance in predicting the daily evaporation rate. The use of prediction techniques with ML models to predict evaporation rates can be a solution to fill the void of meteorological observation data and has significant benefits in agriculture and hydrology. Future research could involve the development of a more effective and efficient water resources monitoring and management information system.

Keywords: rate of evaporation, random forest regression, support vector regression, hyperparameter optimization

Abstrak

Memprediksi laju penguapan memiliki manfaat yang luas dalam berbagai aplikasi seperti manajemen sumber daya air, pertanian, dan lingkungan hidup. Namun untuk mendapatkan data yang lengkap dan akurat dalam mempelajari laju penguapan memiliki tantangan tersendiri. Selain itu, rendahnya tingkat linieritas antara data laju penguapan dan faktor meteorologi lainnya di wilayah tropis dapat menyebabkan hasil prediksi yang bervariasi. Tujuan dari penelitian ini adalah memprediksi laju penguapan harian di Stasiun Klimatologi Yogyakarta dengan membandingkan kinerja dua model *machine learning* (ML) yaitu *random forest regression* (RFR) dan *support vector regression* (SVR) menggunakan data pengamatan meteorologi harian. Untuk meningkatkan akurasi prediksi, dilakukan optimasi *hyperparameter* menggunakan metode *gridsearch cross-validation* untuk mencari kombinasi *hyperparameter* terbaik. Hasil optimasi *hyperparameter* pada data *training* menunjukkan bahwa model RFR menghasilkan skor RMSE sebesar -0,67 sementara model SVR pada kernel RBF menghasilkan skor RMSE negatif sebesar -0,57. Evaluasi lebih lanjut dilakukan pada data *testing* dengan menggunakan kombinasi *hyperparameter* hasil optimasi model RFR menghasilkan nilai R2 sebesar 0,79 dan RMSE sebesar 0,56 sedangkan model SVR menghasilkan koefisien determinasi (R2) sebesar 0,81 dan RMSE sebesar 0,53. Berdasarkan hasil perbandingan kedua model dapat disimpulkan bahwa model SVR memiliki kinerja yang lebih baik dalam memprediksi laju penguapan harian. Penggunaan teknik prediksi dengan model ML untuk memprediksi laju penguapan dapat menjadi solusi untuk mengisi kekosongan data pengamatan meteorologi dan memiliki manfaat yang signifikan dalam bidang pertanian dan hidrologi. Penelitian selanjutnya dapat melibatkan pengembangan sistem informasi pemantauan dan pengelolaan sumber daya air yang lebih efektif dan efisien.

Kata kunci: laju penguapan, *random forest regression*, *support vector regression*, optimasi *hyperparameter*

1. Pendahuluan

Memahami laju penguapan sangat penting perencanaan dan pengelolaan dalam sumber daya air [1]. Mengingat hal ini, prakiraan laju penguapan yang akurat bermanfaat untuk efisiensi dan optimalisasi alokasi air untuk berbagai keperluan. Banyak metode termasuk langsung dan tidak langsung yang diterapkan untuk memperkirakan penguapan dari air permukaan terbuka, khususnya di daerah kering dan semi-kering yang mengandalkan waduk untuk persediaan air minum dan makanan[2]. Penguapan (*evaporation*) adalah peristiwa perubahan wujud air menjadi gas yang terjadi pada permukaan tanah atau permukaan air bergerah ke udara [3]. Laju penguapan banyak digunakan sebagai indeks untuk evapotranspirasi potensial atau referensi tanaman. Informasi dari laju pengupaan digunakan dalam berbagai aplikasi, seperti manajemen sumber daya air, pertanian, dan lingkungan hidup [4]. Mendapatkan data yang lengkap dan akurat untuk mempelajari laju penguapan adalah tugas yang menantang dalam penelitian ini.

Beberapa penelitian yang telah dilakukan sebelumnya membahas tentang penggunaan pendekatan *machine learning* (ML) untuk memperkirakan laju penguapan

P-ISSN : 2089-3353
E-ISSN : 2808-9162

bulanan di dua stasiun pengamatan di Irak. empat pendekatan ML yang digunakan dalam studi ini adalah *machine learning extreme (ELM), machine gradient boosting (GBM), random quantile forest (QRF), dan machine regression process Gaussian (GPR*). Data iklim bulanan digunakan sebagai variabel input untuk mengestimasi laju penguapan. Hasil studi ini menunjukkan bahwa GBM menghasilkan peningkatan evaluasi prediksi dengan nilai lebih baik dibandingkan dengan ELM, GPR, dan QRF. Untuk studi lokasi kedua GBM mengalami penurunan performa nilai evaluasi dibandingkan dengan ELM, GPR, dan QRF [5].

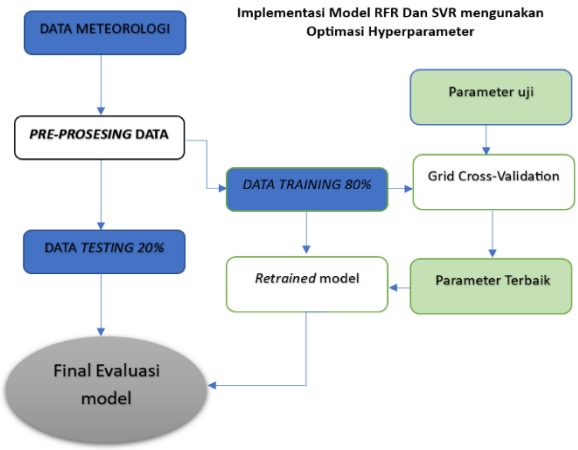
Studi lainnya mengunakan model *hybrid* yang mengintegrasikan algoritma *firefly* (FA) dengan model *Artificial Neural Network* (ANN) untuk mengestimasi penguapan harian dari permukaan air pada dua stasiun cuaca di utara Iran. Variabel input ditentukan dengan menggunakan uji *gamma*, kinerja dari model *hybrid* yang diusulkan dibandingkan dengan *model multilayer perceptron (MLP), neural network feature map self-organizing (SOMNN), dan machine support vector (SVM)*. Model dievaluasi menggunakan koefisien korelasi, RMSE, dan koefisien *nash-sutcliffe* (NS). Hasil penelitian menunjukkan bahwa model SOMNN lebih baik dalam mengestimasi laju penguapan dari permukaan air di stasiun cuaca Anzali dibandingkan dengan model lainnya sedangkan model *hybrid* mengestimasi lebih baik pada stasiun cuaca Astara. Hasil ini menunjukkan bahwa mengintegrasikan algoritma FA dengan model MLP yang sederhana dapat menghasilkan model *hybrid* yang lebih baik untuk mengestimasi laju penguapan [6].

Penelitian lainnya di wilayah Irak yang memiliki iklim gersang dan semi-gersang. empat model ML berbeda *Classification and Regression Trees* (CART), *cascade-correlation neural networks* (CCNN), *gene expression programming* (GEP), dan *support vector mechine* (SVM) dikembangkan dan dibangun dengan menggunakan kombinasi masukan yang berbeda dari variabel meteorologi. Hasil penelitian menunjukkan bahwa prediksi terbaik dicapai dengan memasukkan parameter input lama sinar matahari, kecepatan angin, kelembaban relatif, curah hujan, dan suhu minimum dan suhu maksimum. SVM menunjukkan performa terbaik dengan kecepatan angin, curah hujan, dan kelembaban relatif sebagai input di stasiun I menghasilkan koefisien determinasi (R^2) sebesar 0.92 dan dengan semua variabel sebagai input di Stasiun II menghasilkan akurasi R^2 sebesar 0.97. Seluruh model ML memperlihatkan hasil yang baik dalam memprediksi penguapan di lokasi yang diteliti[7].

Dari hasil kajian sebelumnya pendekatan ML dalam memprediksi laju penguapan cukup sulit karena sangat dipengaruhi kondisi iklim suatu wilayah serta pengaruh banyaknya parameter input yang mempengaruhi variabel output. Penelitian dalam pengerjaannya akan melakukan perbandingan kinerja model *machine learning* mengunakan metode *random forest regression (RFR) dan support vector regression (SVR)* menggunakan data hasil pengamatan meteorologi Stasiun Klimatologi Yogyakarta. Penelitian ini akan cukup berbeda dimana lokasi penelitian dilakukan di wilayah yang memiliki iklim tropis yang lembab. Proses lain yang akan dilakukan yaitu optimasi *hyperparameter* untuk mencari kombinasi *hyperparameter* terbaik menggunakan metode *gridsearch cross validation* (*gridsearchCV*) yang secara teori dapat meningkatkan kinerja akurasi model ML. Penelitian ini diharapkan membantu dalam melakukan prakiraan laju penguapan di wilayah yang memiliki pengamatan meteorologi yang kurang dalam hal ketersediaan data dan untuk pengelolaan sumber daya air lainnya.

2. Metode Penelitian

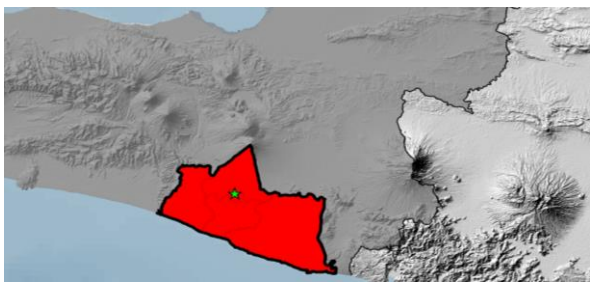
Kerangka konsep metode untuk mengulas penelitian ini terdiri dari 4 tahap yaitu pengumpulan data, *pre-processing* data membagi data menjadi data *training* dan data *testing*, implementasi model RFR dan SVR menggunakan optimasi *hyperparameter* kemudian kedua model ini akan dilakukan evaluasi model yang dapat dilihat pada gambar 1.



Gambar 1. Alur Metode Penelitian

2.1. Pengumpulan Data Meteorologi

Ketersediaan data meteorologi memerlukan jaringan stasiun pengamatan yang baik di permukaan, udara atas, dan di laut, serta sistem pendukung lainnya yang memudahkan pengumpulan, pencatatan, pengolahan,



Gambar 2. Peta Lokasi Penelitian

pengarsipan, dan operasi manajemen data lainnya [8]. Lokasi penelitian dilakukan di Stasiun Klimatologi

P-ISSN : 2089-3353
E-ISSN : 2808-9162

Yogyakarta yang terletak pada 7.73 LS dan 110.35 BT di Indonesia terlihat pada gambar 2.

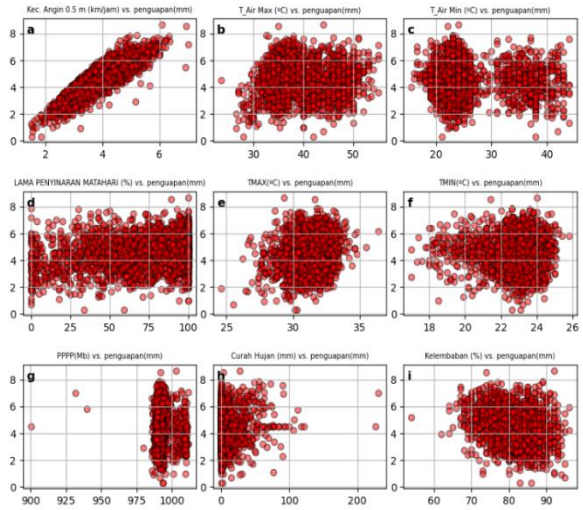
Data penelitian yang digunakan memiliki panjang data 1826 baris yang terdiri dari 10 unsur hasil pengamatan cuaca dari tahun 2017 - 2021 pada skala harian. Sampel data dapat dilihat pada tabel 1.

Tabel 1. Sampel Dataset Dari Stasiun Klimatologi Yogyakarta

DATA PARAMETER METEOROLOGI HARIAN DARI TAHUN 2017 - 2021

waktu	WS	T_Air_ max	T_Air Min	SD	T_udara _max	T_udara _min	P	RR	RH	E
01/01/2017	4,8	40,4	23,5	100,0	30,8	23,3	1001,2	0	89	4,9
01/04/2018	4,0	33,0	24,0	81,0	31,8	23,4	988,9	1	78	4,2
02/02/2019	3,9	37,5	24,5	71,0	31,6	23,6	991,6	16	82	4,9
24/05/2020	3,8	37,0	26,0	100,0	32,1	23,9	992,4	0	76	5,3
31/12/2021	3,5	48,0	41,0	16,0	29,2	22,3	989,2	46	90	3,6

Hasil analisis linieritas pada gambar 3 menunjukan Parameter laju penguapan dan kecepatan angin 0,5 meter memiliki hubungan linieritas yang tinggi dibandingkan parameter lain dimana dari sebaran datanya menunjukan linearitas yang rendah atau dapat dikatakan *non linier*. Proses ini akan membantu penelitian dalam menetukan metode yang cocok untuk pengolahan data untuk menghasilkan akurasi prediksi yang lebih baik.



Gambar 3. Hubungan Linieritas Varibel ependen dan vaVriabel Independen

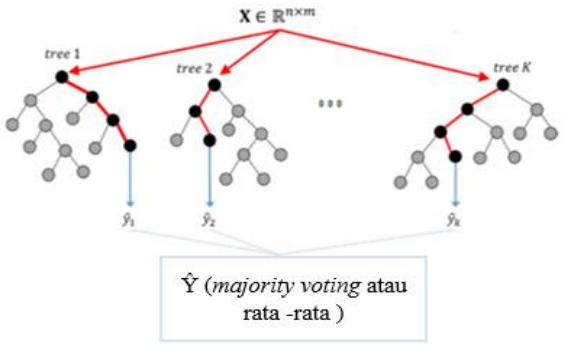
2.2. *Pre-processing* Data

Pre-processing dapat didefinisikan sebagai proses pembersihan dan transformasi data. *Pre-processing* mencakup *memuat library*, memuat dataset, menentukan variabel independen dan variabel dependen, menangani nilai *null* dan membagi set data menjadi set pelatihan dan pengujian [9]. Pada model SVR dilakukan standarisasi menggunakan *standardscaler* dengan menskalakan data berkisar di sekitar 0 dengan standar deviasi 1. Rata-rata dan standar deviasi dihitung untuk setiap komponen kemudian diskalakan [10]. Proses pemisahan dataset dipilih data laju penguapan (E) sebagai variabel dependen dan variabel suhu air panci maksimum dan minimum (T_air max dan T_air min), kecepatan angin ketinggian 0,5meter (WS), suhu udara maksimum (T_max), suhu udara minimum (T_min), lama penyinaran matahari (SD), tekanan udara (P), curah hujan (RR) dan kelembaban rata - rata udara (RH) sebagai variabel independen.

2.3. Implementasi Model RFR dan SVR Menggunakan Optimasi *Hyperparameter*

Machine learning adalah sebuah proses pembelajaran mesin mengunakan data-data yang diinputkan ke mesin sehingga mesin tersebut mampu berpikiran dan berperilaku seperti manusia [11]. Pembentukan model ML *random forest* (RF) Dikembangkan pada 1990-an, *random forest* telah dikenal karena memiliki kemampuannya yang canggih dalam klasifikasi atau regresi, dan kemampuan untuk menangani variabel kategori atau kontinyu, serta menangani data yang hilang [12]. Model Random Forest memiliki popularitas yang pesat dalam industri karena kinerjanya yang baik dalam berbagai aplikasi. Beberapa aplikasi populer termasuk pemodelan prediktif untuk diagnosis kesalahan dan analisis akar penyebab, deteksi titik perubahan dalam data, serta pemodelan diagnostik untuk perbaikan informasi dari model. Random Forest mampu menangani data yang kompleks, non-linear, dan memiliki banyak fitur, sehingga membuatnya cocok untuk berbagai masalah prediksi dan pemodelan di berbagai industri[12].

Tahap pertama metode ini dengan pemilihan *subset training set* sehingga berlanjut membentuk *nodes* pada *tree*. Pencarian *node* akan terus dilakukan hingga semua *nodes* pada setiap *tree* terbentuk. Pada pendekatan RF untuk kasus regresi atau biasa disebut *random forest regression* (RFR), dari gambar 3 menjelaskan pengambilan sampel acak (*bootstrap*) pada data *training* untuk membangun beberapa *decision tree* yang independen. Kemudian hasil prediksi dari decision tree ŷ1, ŷ2, ŷ3 sampai ŷk digabungkan lalu dilakukan *voting* atau rata-rata untuk menghasilkan nilai prediksi akhir seperti yang terlihat pada gambar 4 [13].



Gambar 4. Representasi Random Forest.

Sedangkan untuk model *Support Vector Machine* (SVM) dilakukan dengan cara menemukan hipotesis

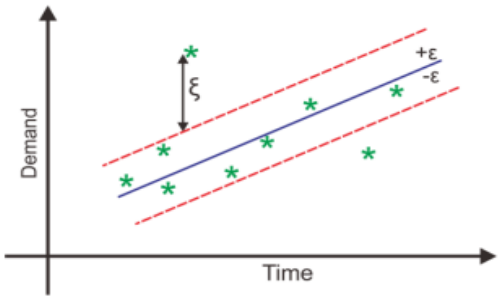
P-ISSN : 2089-3353
E-ISSN : 2808-9162

yang memiliki margin maksimum atau jarak terbesar antara batas keputusan dan sampel terdekat (*hyperplane*), sehingga dapat meminimalkan kesalahan dan meningkatkan kemampuan generalisasi model [14]. Seperti yang terlihat pada gambar 5, SVM memungkinkan batasan keputusan yang sangat kompleks, meskipun data hanya memiliki beberapa fitur. Secara umum akan melakukan pengelompokan dengan mencari *hyperplane* yang memisahkan 2 ikon klasifikasi [10]. Untuk kasus regresi pada SVM disebut *support vector regression* (SVR) dimana tujuan dari SVR adalah untuk menemukan fungsi regresi [15].. Fungsi regresi non linear pada SVR yang dapat dituliskan sebagai berikut :

$$f(x_i) = \sum_{i=1}(\alpha_i - \alpha_i^*)K(x_i, x_j) + b \quad (1)$$

Dengan

(α_i, α_i^*) : pengali lag range,
$K(x_i, x_j)$: Fungsi kernel,
b : konstanta,



Gambar 5. Grafik Representasi SVM

Dalam metode SVR menggunakan beberapa parameter yaitu ε dan C yang berpengaruh dalam menentukan toleransi kesalahan, cLR sebagai penentu kecepatan proses pembelajaran, σ sebagai konstanta yang berpengaruh pesebaran dimensi data, dan λ sebagai penentu ukuran skala dimensi pemetaan kernel SVR. Fungsi kernel digunakan untuk memetakan dimensi data sehingga diharapkan dapat menghasilkan dimensi data yang lebih tinggi dan terstruktur[16]. Pada penelitian ini, menggunakan 3 jenis fungsi kernel nonlinear yaitu kernel *polinomial*, kernel *radial basis function* (RBF) dan kernel *sigmoid*. SVM membangun hyperplane. Berikut adalah pendekatan dari tiga fungsi kernel tersebut:

1. Kernel *polinomial*
$$K(x_i, x_j) = (\gamma x_i^T x_j + 1)^d \quad (2)$$
2. Kernel *Radial Basis Function* (RBF)
$$K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2) \quad (3)$$
3. Kernel *sigmoid*
$$K(x_i, x_j) = \tanh(\gamma(x_i, x_j) + \theta) \quad (4)$$

dengan

x_i, x_j : pasangan dua data dalam data *training*
d : degree parameter yang diperlukan untuk kernel
γ : ukuran similaritas dua vektor (gamma)
θ : Parameter yang menentukan *offset* atau pergeseran dari fungsi sigmoid

2.4. Evaluasi Model

Ukuran kinerja model dievaluasi dengan membandingkan nilai prediksi dengan nilai observasi yang sesuai menggunakan metrik evaluasi statistik seperti *root mean square erro*r (RMSE) dan koefisien determinasi (R^2). Dalam proses penentuan skor *hyperparamete*r menggunakan *gridsearchCV*, RMSE dijadikan sebagai metrik evaluasi dengan nilai negatif. Penggunaan nilai negatif pada RMSE mempermudah dalam mengukur skor, di mana skor yang lebih tinggi menunjukkan kualitas prediksi yang lebih baik dan kesesuaian dengan data observasi.

$$RMSE = \sqrt{\frac{\sum_{j=i}^{n}(y^i - y)^2}{n}} \quad (5)$$

Keterangan:

n = *Error rate*.
i = Jumlah keseluruhan data.
Yi = Nilai keluaran atau output (prediksi).
Y = Nilai aktual atau sebenarnya

koefisien determinasi (R^2). Nilai RMSE digunakan untuk membedakan kinerja model selama periode kalibrasi dan validasi dan digunakan untuk membandingkan kinerja antara model individual dan model prediktif lainnya sedangkan R^2 digunakan untuk menilai kesesuaian antara variabel independen dan dependen [17]. Nilai R^2 yang kecil mendekati 0 artinya variabel independen memiliki kemampuan yang sangat terbatas untuk menjelaskan variabel dependen. Sebaliknya, nilai yang mendekati 1 (1) dan jauh dari 0 (nol) menunjukkan bahwa variabel independen mampu memberikan semua informasi yang dibutuhkan untuk memprediksi variabel dependen [18]

3. Hasil dan Pembahasan

3.1. *Pre-processing* pengisian data *null*

Tahap *pre-processing* untuk mengisi data *null* dengan menggunakan nilai rata-rata di setiap parameter melibatkan beberapa langkah. Tahap awal melibatkan pemeriksaan setiap parameter atau kolom dalam dataset untuk menemukan data yang memiliki nilai *null* atau kosong. Hasil pengecekan menunjukkan adanya 22 data kosong yang ditemukan. Selanjutnya, dilakukan perhitungan nilai rata-rata untuk setiap parameter yang memiliki nilai *null*. Langkah ini melibatkan penjumlahan semua nilai yang valid dalam parameter tersebut, kemudian dibagi dengan jumlah nilai yang valid. Dengan demikian, nilai rata-rata untuk setiap parameter dapat ditentukan.

Setelah mendapatkan nilai rata-rata untuk setiap parameter, langkah selanjutnya adalah mengisi nilai *null* dalam dataset dengan menggunakan nilai rata-rata yang sesuai dengan parameter tersebut. Setiap nilai *null* digantikan dengan nilai rata-rata yang tepat, sehingga

P-ISSN : 2089-3353
E-ISSN : 2808-9162

dataset menjadi lebih lengkap dan siap untuk menjalani proses analisis lebih lanjut. Proses pengisian data *null* telah dilakukan seperti yang terlihat pada tabel 2, di mana semua data telah terisi. Dengan demikian, proses pembelajaran mesin dapat dilanjutkan menggunakan dataset yang sudah lengkap.

Tabel 2. *Prosessing Filing* Data

List data		*Filing* data *null*	
waktu	0	waktu	0
penguapan(mm)	2	penguapan(mm)	0
Kec. Angin 0.5 m (km/jam)	0	Kec. Angin 0.5 m (km/jam)	0
T_Air Max (ºC)	5	T_Air Max (ºC)	0
T_Air Min (ºC)	3	T_Air Min (ºC)	0
LAMA PENYINARAN MATAHARI (%)	2	LAMA PENYINARAN MATAHARI (%)	0
TMAX(ºC)	2	TMAX(ºC)	0
TMIN(ºC)	2	TMIN(ºC)	0
PPPP(Mb)	2	PPPP(Mb)	0
Curah Hujan (mm)	2	Curah Hujan (mm)	0
Kelembaban (%)	2	Kelembaban (%)	0
dtype: int64		dtype: int64	

3.2. Optimasi *hyperparameter*

Dalam penelitian, optimasi *hyperparameter* merupakan langkah penting dalam mengembangkan model ML yang baik. *Hyperparameter* adalah parameter yang tidak ditentukan oleh model itu sendiri, melainkan harus diatur sebelum proses pembelajaran model dimulai. Salah satu alasan mengapa optimasi *hyperparameter* penting adalah untuk menghindari *overfitting*. *Overfitting* terjadi ketika model terlalu terbiasa dengan data pelatihan dan tidak dapat menggeneralisasi dengan baik pada data uji atau data baru. Jika menggunakan *hyperparameter* yang tidak optimal, model cenderung menjadi terlalu kompleks atau terlalu sederhana, yang dapat mengarah pada *overfitting* atau *underfitting*[19]. Optimasi *hyperparameter* memiliki peran yang sangat penting dalam mengoptimalkan kinerja ML dan mampu meningkatan nilai akurasi model [20].

Kesulitan dalam penelitian ini adalah menentukan *hyperparameter* terbaik yang akan digunakan dalam model. Dalam upaya menyelesaikan tantangan ini, metode yang dapat digunakan adalah *gridSearchCV*, yang merupakan metode optimasi *hyperparameter*. Dengan menggunakan *gridSearchCV*, maka akan dapat mencari kombinasi *hyperparameter* yang menghasilkan performa model terbaik dalam prediksi. [21]. Teknik pencarian *grid* pada *gridSearchCV* bergantung pada percobaan kombinasi *hyperparameter* dalam pengujian, optimasi *hyperparameter gridsearchCV* adalah proses yang rumit dan memakan waktu dan sulit ditafsirkan[22]. Untuk mempercepat proses maka menggunakan fungsi *gridsearchCV* dari library Scikit-Learn yang menggunakan API estimator tipikal. Saat menyesuaikan kumpulan data, semua kemungkinan kombinasi nilai parameter dievaluasi dan kombinasi terbaik disimpan. Pencarian kisi-kisi yang disediakan oleh *gridSearchCV* secara mendalam menghasilkan kandidat dari kisi nilai parameter yang ditentukan dengan param_gridparameter[23].

Optimasi *hyperparameter* pada model RFR dieksplorasi menggunakan 3 lipatan *cross-validation*, banyak iterasi sebesar 100, *n_jobs* bernilai -1, *random_state* = 42. Daftar param_gridparameter selama pencarian *gridSearchCV* dapat dilihat pada tabel 3.

Tabel 3. Daftar Param_gridparameter Optimasi RFR

hyperparameter	Nilai *param_grid parameter*	keterangan
bootstrap	*true, false*	Metode pengambilan sampel titik data.
max_depth	10,20,30,40,50,60,70, 80,90,100,110 dan Nome	Kedalaman maksimum *tree* didefinisikan sebagai jalur terpanjang antara simpul akar dan simpul daun
max_features	*Auto, sqrt*	menyerupai jumlah fitur maksimum yang disediakan untuk setiap *tree* di *random forest*.
min_samples_leaf	1,2,4	menentukan jumlah minimum sampel yang harus ada di simpul daun setelah membelah sebuah simpul
min_sample_split	2,5,10	Parameter yang memberi tahu *decision tree* pada *random forest* jumlah pengamatan minimum yang diperlukan di *node* mana pun untuk membaginya
n_estimator	200,400,600,800, 1000,1200,1400, 1600,1800,2000	Jumlah *tree*

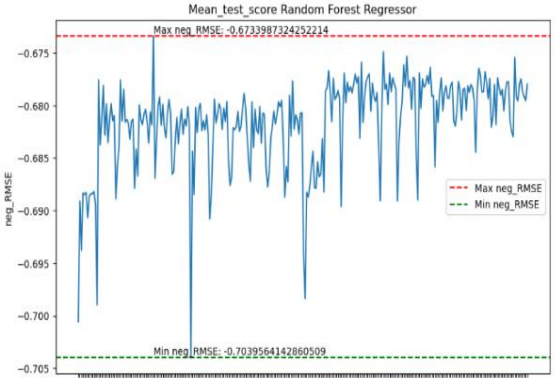
Sedangkan untuk optimasi pada model SVR dilakukan pada 3 fungsi kernel yaitu *polinomial, RBF* dan *sigmoid* menggunakan *hyperparameter* kompleksitas (C) dan *gamma* (γ)[24]. Kombinasi nilai *param_gridparameter* yang digunakan yaitu 0.1, 1 dan 10 yang diuji untuk ketiga kernel. Proses optimasi *hyperparameter* pada penelitian ini melibatkan penggunaan seluruh data pelatihan, yang merupakan 80% dari total data yang tersedia. Seluruh dataset pada data *training* digunakan sebagai data pelatihan untuk melatih dan mengoptimasi model.

3.3. Hasil optimasi *hyperparameter* model RFR

Proses optimasi menggunakan data training yang dilakukan secara berulang untuk semua grid bertujuan menghasilkan kombinasi dengan skor terbaik. Melalui proses ini, berbagai kombinasi *hyperparameter* dieksplorasi dan dinilai menggunakan metrik evaluasi yang relevan untuk menemukan kombinasi terbaik yang mengoptimalkan performa model. Setelah melakukan pengujian pada Model RFR, diperoleh nilai skor rata-rata maksimum neg_RMSE sebesar -0.67 dan nilai skor rata-rata minimum neg_RMSE sebesar -0.70. Informasi ini mengindikasikan variasi performa model dalam prediksi dan kesesuaian dengan data observasi. Semakin mendekati nilai 0 pada neg_RMSE, semakin

P-ISSN : 2089-3353
E-ISSN : 2808-9162

baik kualitas prediksi model. Informasi lengkap mengenai hasil rata-rata skor dari keseluruhan proses optimasi *hyperparameter* dengan menggunakan *gridSearchCV* dapat ditemukan pada gambar 6.



Gambar 6. Optimasi *Hyperparameter* Dengan *GridsearchCV* Pada Model RFR

Gambar 6 memberikan gambaran secara komprehensif tentang performa dan hasil evaluasi keseluruhan dari proses optimasi yang dilakukan pada *hyperparameter* model dengan menggunakan metode *gridSearchCV*.

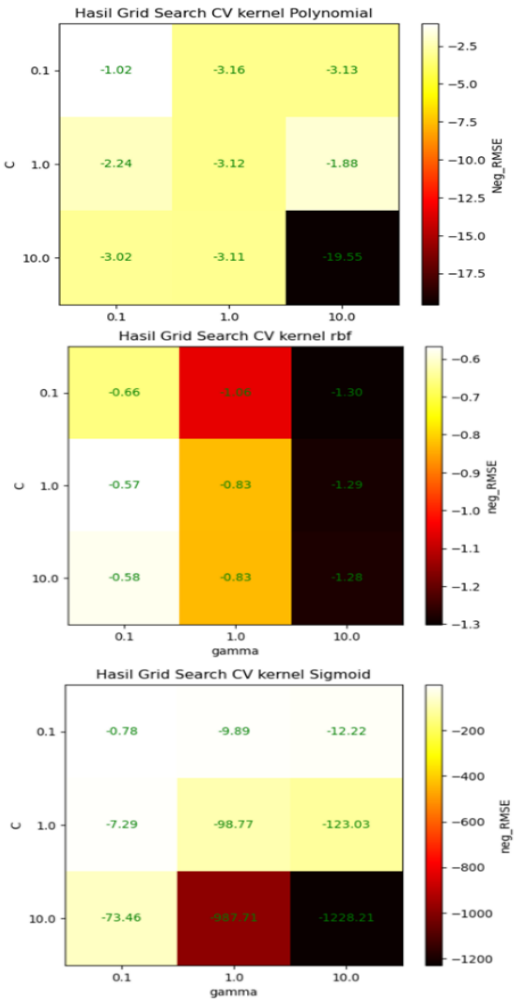
Nilai neg_RMSE maksimum dihasilkan dari kombinasi *hyperparameter* yang dapat dilihat pada tabel 4.

Tabel 4. Grid Pengujian Optimasi RFR

hyperparameter	Hasil Parameter terbaik
bootstrap	false
max_depth	50
max_features	sqrt
min_samples_leaf	2
min_sample_split	5
n_estimator	200

3.4. Hasil proses optimasi *hyperparameter* model SVR

Proses pengujian grid pada model SVR untuk mendapatkan kombinasi *hyperparameter* terbaik dilakukan disetiap kernel secara terpisah. Pada gambar 7 hasil *gridsearchCV* pada kernel *polymonial* menghasilkan nilai skor rata- rata tertinggi dengan nilai neg_RMSE – 1.02 dari kombinasi *hyperparameter* C = 0.1 dan γ = 0.1 dan skor terendah -19.55 dihasilkan oleh *hyperparameter* C = 10 dan γ = 10. Pada kernel RBF menghasilkan nilai skor rata- rata tertinggi dengan nilai neg_RMSE – 0,57 dari kombinasi nilai parameter C = 1 dan γ = 0.1 dan skor terendah -1.30 dihasilkan oleh *hyperparameter* C = 0.1 dan γ = 10. Pada kernel sigmoid menghasilkan nilai skor rata- rata tertinggi dengan nilai neg_RMSE – 0.78 dari kombinasi nilai parameter C = 1 dan γ = 0.1 dan skor terendah -1228.21 dihasilkan oleh *hyperparameter* C = 10 dan γ = 10. Berdasarkan analisis perbandingan tersebut, dapat disimpulkan bahwa hasil pengujian optimasi *hyperparameter* menggunakan kernel RBF lebih baik daripada menggunakan kernel polynomial ataupun sigmoid. Oleh karena itu, langkah selanjutnya adalah melakukan evaluasi model menggunakan data *testing* mengunakan kombinasi hyperparamenter dari kernel RBF yang telah diketahui.



Gambar 7. Hasil Pengujian *GridsearchCV* Model SVR

3.5. Hasil evaluasi model

Proses evaluasi model dilakukan dengan menggunakan 20% data testing dari total dataset. Dalam evaluasi ini, data testing yang belum pernah digunakan sebelumnya digunakan untuk menguji performa model. Hyperparameter terbaik hasil optimasi dijalankan dalam proses evaluasi model. Hasil evaluasi dari model Random Forest Regressor (RFR) dan Support Vector Regressor (SVR) tercantum pada Tabel 5, yang memberikan gambaran tentang performa keduanya.

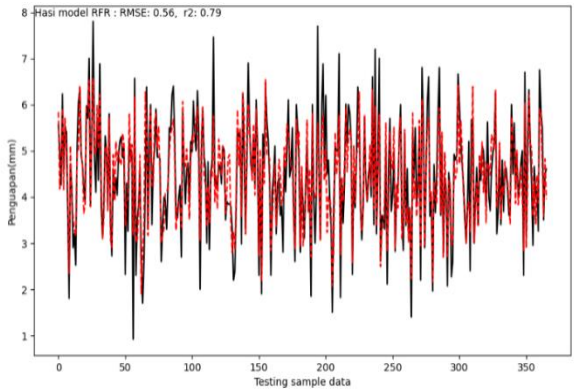
Tabel 5. Hasil Valuasi Eksekusi Model

Model	R^2	RMSE
RFR	0.79	0.56
SVR	0.81	0.53

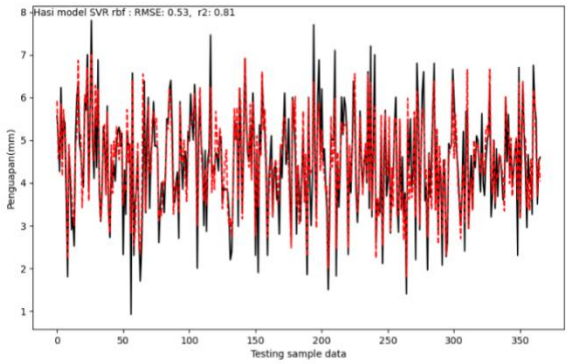
Hasil spesifik prediksi dan nilai observasi dapat ditemukan pada gambar 8 untuk model RFR dan gambar 9 untuk model SVR. Dengan mengacu pada

P-ISSN : 2089-3353
E-ISSN : 2808-9162

kedua gambar tersebut dapat memperoleh pemahaman yang lebih komprehensif tentang performa dan kesesuaian prediksi dari masing-masing model RFR dan SVR.



Gambar 8. Grafik Hasil Prediksi Model RFR



Gambar 9. Grafik Hasil Prediksi Model SVR

Penelitian ini menunjukkan bahwa penggunaan model SVR dan RFR dalam memprediksi laju penguapan mampu memberikan hasil akurasi yang cukup baik dengan teknik optimasi *hyperparameter* untuk meningkatkan kinerja model prediksi. Penelitian ini dilakukan di wilayah beriklim tropis dan lembab menghasilkan akurasi prediksi yang cukup baik, yang dapat memberikan kontribusi penting dalam meningkatkan pemahaman tentang faktor-faktor yang mempengaruhi penguapan di wilayah tersebut.

Hasil penelitian ini selain sebagai rujukan untuk mengisi parameter pengamatan laju penguapan di wilayah pengamatan meteorologi yang kurang lengkap. Selain itu dapat berguna dalam meningkatkan efisiensi dan efektivitas sistem irigasi pada wilayah minim air. Setelah mengetahui laju penguapan yang tepat, sistem irigasi dapat disesuaikan untuk memberikan air yang dibutuhkan oleh tanaman dengan tepat, sehingga dapat menghemat air dan mengoptimalkan produksi tanaman.

Laju penguapan juga merupakan faktor penting dalam siklus hidrologi, terutama dalam menentukan debit sungai dan tingkat kelembaban di lingkungan. Informasi ini dapat digunakan untuk memperkirakan aliran air permukaan dan kelembaban tanah yang dapat berdampak pada manajemen sumber daya air dan hidrologi.

4. Kesimpulan

Prediksi laju penguapan memiliki tingkat kompleksitas yang tinggi karena berkaitan dengan berbagai variabel iklim yang memiliki *non-linearitas* tinggi. Hasil pengujian di stasiun klimatologi Yogyakarta menunjukkan bahwa model SVR memiliki prediksi yang lebih baik, dengan nilai R^2 sebesar 0.81 dan RMSE sebesar 0.53 dibandingkan dengan model RFR yang memiliki nilai R^2 sebesar 0.79 dan RMSE sebesar 0.56 pada saat pengujian. Teknik optimasi *hyperparameter* dengan *gridsearchCV* memudahkan dalam memilih kombinasi *hyperparameter* yang tepat untuk diterapkan pada model dalam meningkatkan kinerja model prediksi. Hasil Penelitian ini dapat memberikan kontribusi signifikan dalam meningkatkan pemahaman tentang kinerja model ML dalam memprediksi laju penguapan dan dapat menjadi dasar untuk pengembangan penelitian selanjutnya.

Penggunaan model ML dalam memprediksi laju penguapan juga merupakan salah satu solusi untuk mengisi kekosongan data pengamatan meteorologi. Kontribusi penting lainnya bermanfaat dalam meningkatkan efisiensi dan efektivitas sistem irigasi serta meningkatkan pemahaman tentang siklus hidrologi di wilayah yang beriklim tropis dan lembab. Oleh karena itu, penelitian ini dapat dijadikan referensi untuk pengembangan lebih lanjut dalam penelitian terkait laju penguapan dan aplikasinya dalam bidang hidrologi dan pertanian.

Daftar Rujukan

[1] L. D. Lowe, J. A. Webb, R. J. Nathan, T. Etchells, and H. M. Malano, 2009. “Evaporation from water supply reservoirs: An assessment of uncertainty,” *J. Hydrol.*, vol. 376, no. 1–2, pp. 261–274, doi: 10.1016/j.jhydrol.2009.07.037.

[2] F. Helfer, C. Lemckert, and H. Zhang, 2012. “Impacts of climate change on temperature and evaporation from a large reservoir in Australia,” *J. Hydrol.*, vol. 475, pp. 365–378, doi: 10.1016/j.jhydrol.2012.10.008.

[3] L. Wang, O. Kisi, B. Hu, M. Bilal, M. Zounemat-Kermani, and H. Li, 2017. “Evaporation modelling using different machine learning techniques,” *Int. J. Climatol.*, vol. 37, pp. 1076–1092, doi: 10.1002/joc.5064.

[4] W. Fang, S. Huang, Q. Huang, G. Huang, E. Meng, and J. Luan, 2018, “Reference evapotranspiration forecasting based on local meteorological and global climate information screened by partial mutual information,” *J. Hydrol.*, vol. 561, no. April, pp. 764–779, doi: 10.1016/j.jhydrol.2018.04.038.

[5] Z. A. Al Sudani and G. S. A. Salem,2022. “Evaporation Rate Prediction Using Advanced Machine Learning Models: A Comparative Study,” *Adv. Meteorol.*, vol. 2022, doi: 10.1155/2022/1433835.

[6] A. Ashrafzadeh, A. Malik, V. Jothiprakash, M. A. Ghorbani, and S. M. Biazar,2020. “Estimation of daily pan evaporation using neural networks and meta-heuristic approaches,” *ISH J. Hydraul. Eng.*, vol. 26, no. 4, pp. 421–429, doi: 10.1080/09715010.2018.1498754.

[7] Z. M. Yaseen *et al.*,2020. “Prediction of evaporation in arid and semi-arid regions: a comparative study using different machine learning models,” *Eng. Appl. Comput. Fluid Mech.*, vol. 14, no. 1, pp. 70–89, doi:

10.1080/19942060.2019.1680576.

[8] R. Muita *et al.*,2021. "Towards Increasing Data Availability for Meteorological Services: Inter-Comparison of Meteorological Data from a Synoptic Weather Station and Two Automatic Weather Stations in Kenya," *Am. J. Clim. Chang.*, vol. 10, no. 03, pp. 300–316, doi: 10.4236/ajcc.2021.103014.

[9] N. Pandey, P. K. Patnaik, and S. Gupta,2020. "Data Pre Processing for Machine Learning Models using Python Libraries," *Int. J. Eng. Adv. Technol.*, vol. 9, no. 4, pp. 1995–1999, doi: 10.35940/ijeat.d9057.049420.

[10] V. N. G. Raju, K. P. Lakshmi, V. M. Jain, A. Kalidindi, and V. Padma, 2020. "Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification," *Proc. 3rd Int. Conf. Smart Syst. Inven. Technol. ICSSIT 2020*, no. Icssit, pp. 729–735, doi: 10.1109/ICSSIT48917.2020.9214160.

[11] O. Simeone, "A brief introduction to machine learning for engineers,2018. " *Found. Trends Signal Process.*, vol. 12, no. 3–4, pp. 200–431, doi: 10.1561/2000000102.

[12] C. Strobl, J. Malley, and G. Tutz, 2009. "An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests," *Psychol. Methods*, vol. 14, no. 4, pp. 323–348, doi: 10.1037/a0016973.

[13] M. Čeh, M. Kilibarda, A. Lisec, and B. Bajat,2018. "Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments," *ISPRS Int. J. Geo-Information*, vol. 7, no. 5, doi: 10.3390/ijgi7050168.

[14] A. Danandeh Mehr, V. Nourani, V. Karimi Khosrowshahi, and M. A. Ghorbani, 2019. "A hybrid support vector regression–firefly model for monthly rainfall forecasting," *Int. J. Environ. Sci. Technol.*, vol. 16, no. 1, pp. 335–346, doi: 10.1007/s13762-018-1674-2.

[15] M. P. Kusuma and A. Kudus, 2022. "Penerapan Metode Support Vector Regression (SVR) pada Data Survival KPR PT. Bank ABC, Tbk.," *Bandung Conf. Ser. Stat.*, vol. 2, no. 2, pp. 167–172, doi: 10.29313/bcss.v2i2.3614.

[16] D. A. Mardhika, B. D. Setiawan, and R. C. Wihandika, 2019. "Penerapan Algoritma Support Vector Regression Pada Peramalan Hasil Panen Padi Studi Kasus Kabupaten Malang," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 10, pp. 9402–9412, [Online]. Available: http://j-ptiik.ub.ac.id

[17] N. D. Maulana, B. D. Setiawan, and C. Dewi,2019. "Implementasi Metode Support Vector Regression (SVR) Dalam Peramalan Penjualan Roti (Studi Kasus : Harum Bakery)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 3, pp. 2986–2995.

[18] A. Saiful,2021. "Prediksi Harga Rumah Menggunakan Web Scrapping dan Machine Learning Dengan Algoritma Linear Regression," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 8, no. 1, pp. 41–50, doi: 10.35957/jatisi.v8i1.701.

[19] C. G. Siji George and B. Sumathi,2020. "Grid search tuning of hyperparameters in random forest classifier for customer feedback sentiment prediction," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 9, pp. 173–178, doi: 10.14569/IJACSA.2020.0110920.

[20] Y. Azhar, G. A. Mahesa, and M. C. Mustaqim,2021. "Prediction of hotel bookings cancellation using hyperparameter optimization on Random Forest algorithm," *J. Teknol. dan Sist. Komput.*, vol. 9, no. 1, pp. 15–21, doi: 10.14710/jtsiskom.2020.13790.

[21] A. Toha, P. Purwono, and W. Gata,2022. "Model Prediksi Kualitas Udara dengan Support Vector Machines dengan Optimasi Hyperparameter GridSearch CV," *Bul. Ilm. Sarj. Tek. Elektro*, vol. 4, no. 1, pp. 12–21, doi: 10.12928/biste.v4i1.6079.

[22] P. Martínez, M. Advisor, : Oriol, and P. Vila, 2017. "Smart optimization of hyper-parameters in Support Vector Machines. Studying model dropout for hyper-parameter optimization in support vector machines," no. October, 2017.

[23] "Tuning the hyper-parameters of an estimator," 2023. https://scikit-learn.org/stable/modules/grid_search.html (accessed May 29, 2023).

[24] K. Cheng, Z. Lu, Y. Wei, Y. Shi, and Y. Zhou,2017. "Mixed kernel function support vector regression for global sensitivity analysis," *Mech. Syst. Signal Process.*, vol. 96, pp. 201–214, doi: 10.1016/j.ymssp.2017.04.014.