ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

PREDIKSI EVAPORASI BERBASIS MESIN: PERBANDINGAN ANN, KNN, RANDOM FOREST DAN REGRESI LINIER

Muchamad Rizqy Nugraha✉, Haidar Amru Rusdan, Marzuki Sinambela
Instrumentasi-MKG, Sekolah Tinggi Meteorologi Klimatologi dan Geofisika, Tangerang, Indonesia
Email: rizqy.nugraha@stmkg.ac.id

DOI: https://doi.org/10.46880/jmika.Vol9No2.pp380-386

ABSTRACT

Evaporation plays a key role in water allocation, yet data limitations are often encountered. This study evaluates four regression models (Linear Regression, K-Nearest Neighbors, Random Forest, and Artificial Neural Network — ANN) to predict evaporation rates at the Banten Climatology Station. Models were assessed using R-squared (R^2) and Root Mean Squared Error (RMSE). The results show that the ANN achieved the best accuracy with RMSE = 0.122 and R^2 = 0.475 (47.5%), followed by Linear Regression (RMSE = 0.123, R^2 = 0.460), K-Nearest Neighbors (RMSE = 0.126, R^2 = 0.437), and Random Forest (RMSE = 0.129, R^2 = 0.406). Other models also provided acceptable predictions, but the ANN stood out as the most accurate and reliable for applications at the Banten Climatology Station. These findings offer valuable insights for water resources management and agricultural planning, highlighting the potential of machine learning techniques to overcome evaporation data limitations.

Keywords: Evaporation, Regression Model, R-squared, Root Mean Squared Error.

ABSTRAK

Evaporasi memainkan peran kunci dalam alokasi air, namun keterbatasan data sering terjadi. Penelitian ini mengevaluasi empat model regresi (Linear Regression, K-Nearest Neighbors, Random Forest, dan Artificial Neural Network - ANN) untuk memprediksi tingkat evaporasi di Stasiun Klimatologi Banten. Model-model dievaluasi menggunakan metrik R-squared (R^2) dan Root Mean Squared Error (RMSE). Hasil dari penelitian ini menunjukkan bahwa model ANN memberikan akurasi terbaik dengan RMSE=0,122 dan R2=0,475 (47,5%). Kemudian disusul oleh linear regression (RMSE=0,123, R2=0,460), K-Nearest Neighbors (RMSE=0,126, R2=0,437), dan Random Forest (RMSE=0,129, R2=0,406). Model-model lainnya juga memberikan prediksi yang layak, namun ANN menonjol sebagai yang paling akurat dan dapat diandalkan untuk aplikasi di Stasiun Klimatologi Banten. Temuan ini memberikan wawasan berharga untuk manajemen sumber daya air dan perencanaan pertanian, menunjukkan potensi teknik pembelajaran mesin dalam mengatasi keterbatasan data evaporasi.

Kata Kunci Evaporasi, Model Regresi, R-squared, Root Mean Squared Error.

PENDAHULUAN

Evaporasi memiliki berperan penting terhadap alokasi air dalam siklus hidrologi, pertanian, dan pengelolaan sumber daya air. Evaporasi merujuk pada perpindahan air ke atmosfer dari tanah atau perairan. Evaporasi dipengaruhi oleh gradien tekanan uap dan ketersediaan energi panas, yang ditentukan oleh data cuaca seperti suhu, kelembaban relatif, radiasi matahari, dan kecepatan angin. (Fan et al., 2018; Vicente-Serrano et al., 2018)

Namun, ketersediaan data evaporasi sering kali terbatas karena berbagai faktor, seperti kerusakan, gangguan pada alat, atau ketiadaan stasiun klimatologi di lokasi tertentu. Oleh karena itu, perlu dibuat suatu model prakiraan evaporasi dengan persamaan tertentu yang melibatkan data iklim lainnya, seperti suhu udara, tekanan udara, radiasi matahari, kelembaban relatif, dan kecepatan angin. Dengan demikian, model pendugaan dapat menjadi alternatif yang efektif untuk mengatasi keterbatasan data yang mungkin terjadi.

Ada beberapa model untuk memprediksi evaporasi, seperti Metode Langbein, Penman, Turc, Thornthwaite, Rohwer, dan Orstom (Singh, 1988). Namun, nilai-nilai dalam metode-metode tersebut spesifik untuk negara tempat rumusnya dikembangkan, sehingga perlu diuji dan disesuaikan dengan kondisi lokal. Teknik-teknik pembelajaran mesin seperti linear regression, jaringan saraf tiruan (ANN), *Random Forest*, mesin vektor pendukung (SVM), sistem inferensi neuro-fuzzy adaptif (ANFIS), mesin pembelajaran ekstrim (ELM), dan pemrograman ekspresi gen (GEP) belakangan ini digunakan untuk menangani berbagai permasalahan hidrologi (Abed et al., 2022). Teknik-teknik ini lebih sederhana, dan dapat memodelkan proses non-linier yang kompleks tanpa masalah. Beberapa penelitian menyatakan bahwa ANN memberikan perkiraan yang lebih baik dibandingkan dengan metode konvensional (Abed et al., 2022).

Penelitian ini bertujuan untuk mengevaluasi prediktabilitas model Linear regression, *Random Forest*, *Artificial Neural Network* (ANN), dan *K-Nearest Neighbors* (KNN) dalam memperkirakan tingkat evaporasi. Data yang digunakan mencakup periode waktu tertentu, dan performa masing-masing model akan dibandingkan untuk menentukan model prediksi terbaik di Stasiun Klimatologi Banten. Hasil dari penelitian ini diharapkan dapat memberikan wawasan yang berharga dalam pemilihan dan penerapan model prediksi evaporasi untuk keperluan manajemen sumber daya air dan perencanaan pertanian.

KAJIAN LITERATUR

Linear Regression

Teknik yang berguna untuk memodelkan hubungan linier suatu variabel independen (prediktor) terhadap variabel dependen (hasil). Tujuannya adalah untuk menemukan garis (fungsi linear) yang paling baik menggambarkan hubungan tersebut. Dalam konteks prediksi, Linear Regression digunakan untuk membuat prediksi numerik berdasarkan input variabel. Model *Linear Regression* memprediksi nilai output dengan menghitung jumlah tertimbang dari input, ditambah dengan konstanta yang disebut bias atau *intercept* (Putri et al., 2023).

Random Forest

Model yang terdiri dari banyak pohon keputusan yang dibangun secara acak dari sebagian data pelatihan. Model ini dapat digunakan untuk klasifikasi maupun regresi, dengan menggunakan metode voting atau rata-rata untuk menggabungkan hasil dari setiap pohon. Model *Random Forest* memiliki keunggulan dalam mengatasi masalah *overfitting*, variabilitas, dan ketergantungan pada variabel.

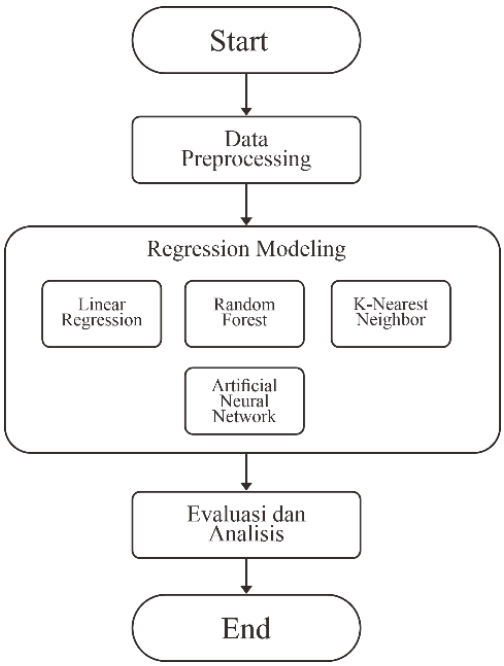
Artificial Neural Network (ANN)

Model komputasi terinspirasi dari struktur dan fungsi jaringan saraf biologis manusia. ANN terdiri dari unit komputasi yang disebut neuron buatan, yang diorganisir dalam lapisan-lapisan dan saling terhubung dengan bobot. Neuron-neuron ini bekerja bersama-sama untuk memproses informasi, mengenali pola, dan melakukan tugas tertentu. ANN digunakan dalam *machine learning* untuk pemodelan dan prediksi, serta untuk menangani masalah kompleks yang sulit dipahami oleh pendekatan algoritmik konvensional (Anandaraj et al., 2018).

K-Nearest Neighbors (KNN)

Model yang berdasarkan pada prinsip kesamaan atau kedekatan antara data. Model ini tidak membangun fungsi atau aturan secara eksplisit, tetapi menggunakan data pelatihan sebagai basis pengetahuan. Model ini dapat digunakan untuk klasifikasi maupun regresi, dengan menggunakan metode mayoritas atau rata-rata dari k tetangga terdekat untuk menentukan kelas atau nilai respons dari data baru. *Model K-Nearest Neighbors* dapat menggunakan berbagai macam ukuran jarak, seperti Euclidean, Mahalanobis, atau jarak berbasis *Random Forest* (Cosenza et al., 2020).

METODE PENELITIAN



Gambar 1. Diagram alir sistem

Data Collecting

Data harian selama 5 tahun (Desember 2018 hingga Desember 2023) dari Stasiun Klimatologi Banten digunakan dalam penelitian ini untuk memprediksi evaporasi. Data tersebut meliputi evaporasi, suhu udara, kelembaban udara, tekanan udara, radiasi matahari, dan kecepatan angin.

Data Preprocessing

Pemrosesan data melibatkan langkah-langkah seperti menghapus data ganda, memperbaiki kesalahan, dan memeriksa konsistensi data. Untuk mendapatkan data berkualitas tinggi, penelitian ini memastikan proses preprocessing yang melibatkan identifikasi dan penghapusan data ganda, eliminasi data yang tidak lengkap, serta normalisasi data.

Regression Modeling

Penelitian ini membuat model regresi, seperti *Linear regression, Random Forest*, Jaringan Saraf Tiruan (ANN), dan *K-Nearest Neighbors* (KNN). Saat melakukan pemodelan, dataset dibagi dengan rasio 80% sebagai data latih dan 20% sebagai data uji.

Evaluasi Model

Setelah melalui fase pemodelan, hasil regresi dari model-model yang dikembangkan dievaluasi menggunakan beberapa metrik kinerja utama. Tiga metrik utama yang digunakan untuk mengukur akurasi prediksi adalah *R-squared* (R^2) dan *Root Mean Square Error* (RMSE) (Ardiansyah, 2023).

R square adalah ukuran yang menunjukkan sejauh mana variabel dependen (endogen) dipengaruhi oleh variabel independen (eksogen). Rentang nilai R square adalah 0 sampai 1, dan apabila nilai mendekati 1, mengindikasikan dampak signifikan dari variabel independen pada variabel dependen. Sebaliknya, nilai yang mendekati 0 menandakan dampak yang kecil dari variabel independen terhadap variabel dependen (Ardiansyah, 2023).

$$R^2 = 1 - \frac{SS\ Error}{SS\ Total} = 1 - \frac{\sum(yi-y)^2}{\sum(yi-y)^2}$$

SS Error = variasi dari residu
SS Total = variasi total
yi = Aktual respon ke-i
y = Rerata
yj = Prediksi respon ke-i

RMSE adalah salah satu metode untuk mengukur kesalahan suatu model dalam memprediksi data kuantitatif. RMSE adalah nilai rata-rata dari perbedaan antara nilai yang diprediksi oleh model dan nilai yang sebenarnya, yang kemudian dikuadratkan dan diakarkan. Model yang memiliki RMSE yang rendah berarti lebih akurat (Ardiansyah, 2023). Rumus matematika untuk RMSE adalah sebagai berikut:

$$RMSE = \sqrt{\frac{(y' - y)^2}{n}}$$

n = Jumlah data
y' = Nilai data prediksi
y = Nilai data aktual

HASIL DAN PEMBAHASAN

Penelitian ini memanfaatkan *Jupyter IDE*, sebuah platform berbasis web yang memudahkan pembuatan dan berbagi dokumen kode yang dapat dijalankan dengan kemampuan visualisasi. Selain itu, Jupyter juga menyediakan alat untuk membersihkan data, mentransformasi data, mensimulasikan nilai numerik, melakukan pemodelan statistik, visualisasi data, dan mengimplementasikan pembelajaran mesin.

Pemahaman Data

Penelitian ini menggunakan data evaporasi, suhu udara, kelembaban udara, radiasi matahari, tekanan udara, dan kecepatan angin yang tercatat di stasiun klimatologi Banten. Cuplikan dataset parameter cuaca dari stasiun klimatologi Banten dapat dilihat pada Gambar 2.

	Time	Evaporasi	Radiasi_Matahari	Suhu	Kecepatan_Angin	Tekanan_Udara	RH
0	2018-12-01 00:00:00	1.7	954	29.7	0.9	1006.1	75.633333
1	2018-12-02 00:00:00	4.3	1023	30.4	1.6	1005.2	73.600000
2	2018-12-03 00:00:00	4.6	1839	27.6	0.7	1006.6	89.866667
3	2018-12-04 00:00:00	1.4	1682	27.6	0.8	1006.4	84.366667
4	2018-12-05 00:00:00	1.4	641	29.2	1.2	1005.2	79.533333
...	...	...	...	...	...	...	...
1839	2023-12-14 00:00:00	6.2	2316	32.2	1.8	1006.1	71.333333
1840	2023-12-15 00:00:00	5.3	3027	32.1	1.4	1005.8	68.666667
1841	2023-12-16 00:00:00	5.2	2767	31.9	1.7	1005.8	67.666667
1842	2023-12-17 00:00:00	5.2	3131	32.0	2.1	1006.0	69.000000
1843	2023-12-18 00:00:00	7.9	3435	32.5	2.0	1006.1	67.666667

Gambar 2. *Dataset*

Data Preprocessing

Penelitian ini melibatkan langkah pembersihan data pada tahap awal pra-pemrosesan. Hal ini dilakukan untuk memperbaiki atau menghapus data yang rusak dan tidak relevan.

Eksplorasi Data

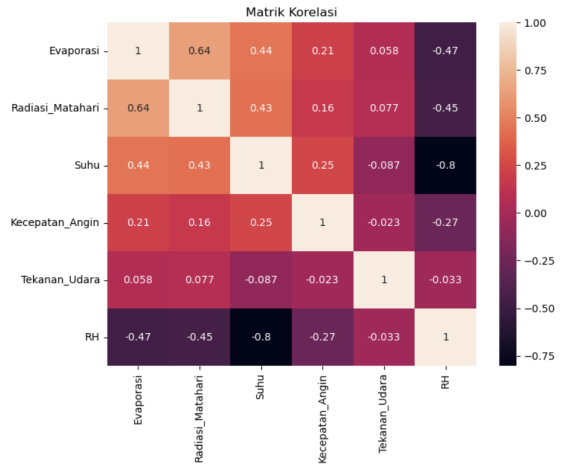
Setelah melakukan tahap data preprocessing, langkah selanjutnya adalah eksplorasi data untuk memahami karakteristik dan pola-pola yang terdapat

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

dalam dataset. Tabel 1 memberikan gambaran umum tentang distribusi nilai-nilai pada setiap parameter cuaca. Rata-rata dan deviasi standar membantu dalam menilai tingkat variasi dan kestabilan data. Selain itu, rentang antara nilai minimum dan maksimum memberikan informasi mengenai sebaran nilai secara keseluruhan. Analisis lebih lanjut terhadap data ini akan menjadi dasar untuk pembentukan model prediksi

Tabel 1. Deskripsi Statistik Data

Parameter	Mean	Stdev	Min	Max
Evaporasi	3,840	1,634	0,000	9,900
Rad. Matahari	1804,621	881,917	19,000	3816,0
Suhu	29,675	1,440	24,900	33,700
Kec. Angin	1,577	0,693	0,100	4,900
Tek. Udara	1006,587	1,331	1002,6	101,40
RH	77,508	7,556	50,000	97,333



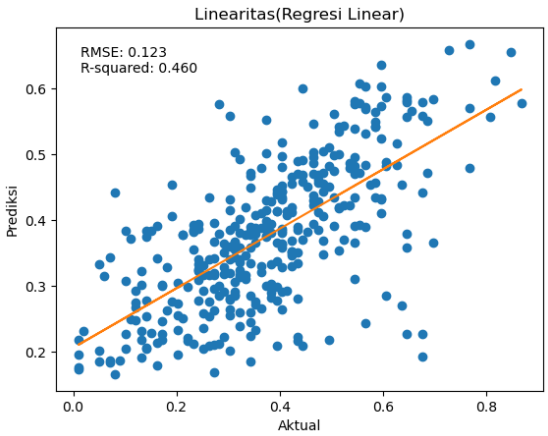
Gambar 3. Matrik Korelasi

Matriks korelasi merupakan alat visual yang efektif untuk menampilkan hubungan antar variabel dalam suatu dataset. Dari Gambar 3, terlihat bahwa evaporasi memiliki korelasi yang kuat dengan radiasi matahari, dengan nilai 0,64. Disusul oleh suhu dengan nilai 0,44, kecepatan angin dengan nilai 0,21, tekanan udara sebesar 0,058, dan korelasi terendah terjadi pada RH dengan nilai -0,47.

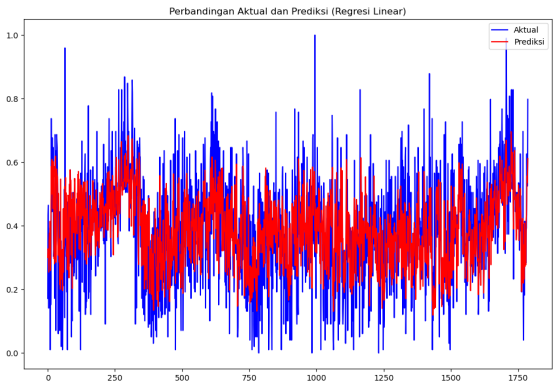
Regression Modeling

Dalam tahap ini, model prediksi dibuat dengan memanfaatkan *Python* dan beberapa *library* pendukungnya, yaitu numpy, pandas, scipy, matplotlib, dan scikit-learn (sklearn). Proses evaluasi dilakukan dengan menggunakan model *Linear regression, Random Forest*, ANN, dan KNN. Data dibagi secara acak, dengan proporsi 80% untuk data latih dan 20% untuk data uji dan validasi. Hal ini memastikan bahwa data uji tidak terlibat dalam proses pelatihan, agar dapat memberikan ukuran yang objektif terhadap performa model baik saat pelatihan maupun setelahnya. Keseluruhan proses ini diimplementasikan menggunakan Jupyter. Kinerja model dinilai dengan menggunakan metrik RMSE dan Rsquare, sementara nilai prediksi dibandingkan secara langsung dengan nilai observasi aktual.

Pada Gambar 4 dan 5, terlihat grafik dan scatter plot perbandingan antara prediksi dan nilai aktual evaporasi menggunakan model regresi linear. Nilai RMSE yang diperoleh sebesar 0,123, sementara R2 adalah 0,460.



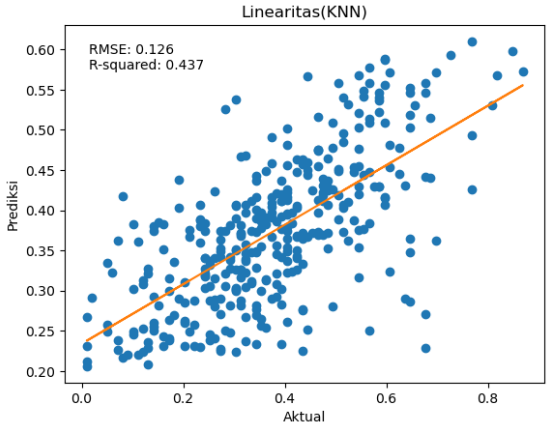
Gambar 4. Plot scatter prediksi dan aktual model *linear regression*



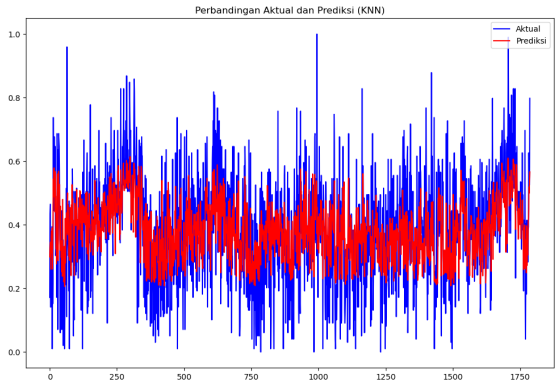
Gambar 5. Grafik prediksi dan aktual model *linear regression*

Untuk model K-Nearest Neighbors (KNN), grafik dan scatter plot perbandingan antara prediksi dan nilai aktual evaporasi dapat dilihat pada Gambar 6 dan 7. Nilai RMSE yang diperoleh sebesar 0,126, sementara R2 adalah 0,437.

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

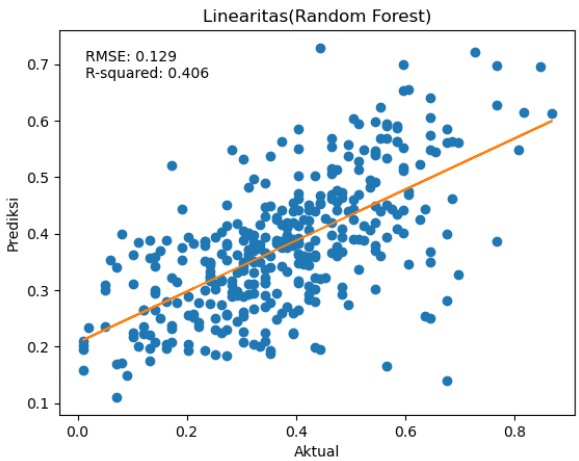


Gambar 6. Plot scatter prediksi dan aktual model *K-Nearest Neighbors* (KNN)

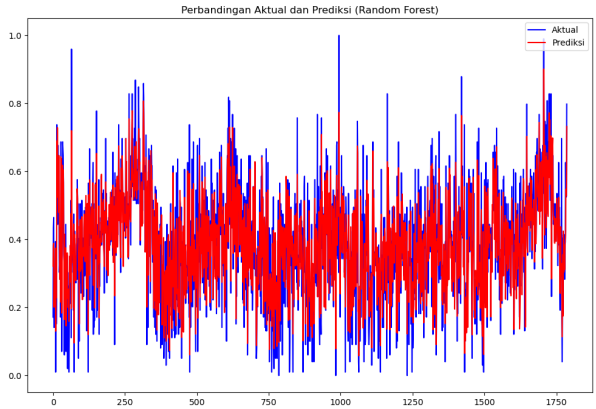


Gambar 7. Grafik prediksi dan aktual model *K-Nearest Neighbors* (KNN)

Grafik dan scatter plot perbandingan antara prediksi dan nilai aktual evaporasi model *Random Forest* dapat dilihat pada Gambar 8 dan 9. Nilai RMSE yang diperoleh sebesar 0,129, dan R2 adalah 0,406.

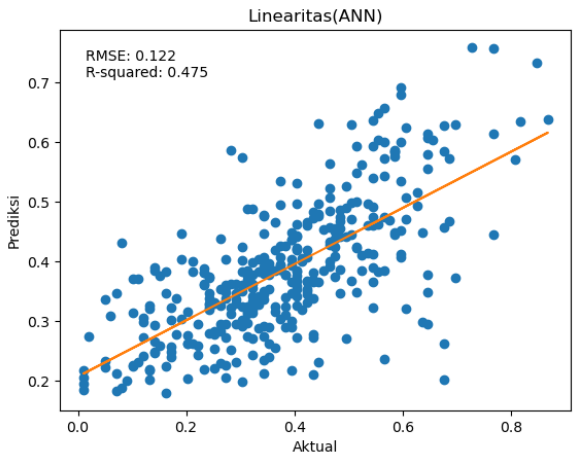


Gambar 8. Plot scatter prediksi dan aktual model *Random Forest*

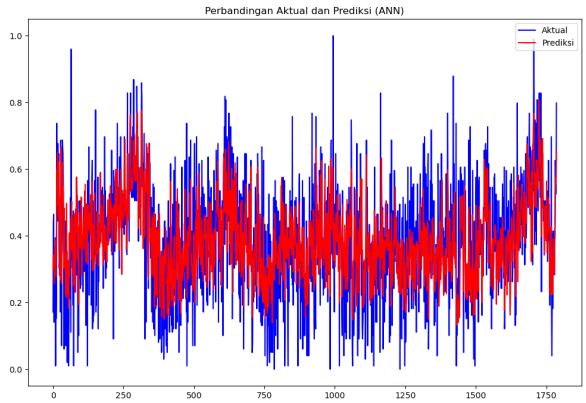


Gambar 9. Grafik prediksi dan aktual model *Random Forest*

Selanjutnya, untuk model *Artificial Neural Network* (ANN), grafik dan scatter plot perbandingan antara prediksi dan nilai aktual evaporasi dapat dilihat pada Gambar 10 dan 11. Didapatkan nilai RMSE sebesar 0,122, dan R2 adalah 0,475.



Gambar 10. Plot scatter prediksi dan aktual model *Artificial Neural Network (ANN)*



Gambar 11. Grafik prediksi dan aktual model *Artificial Neural Network (ANN)*

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

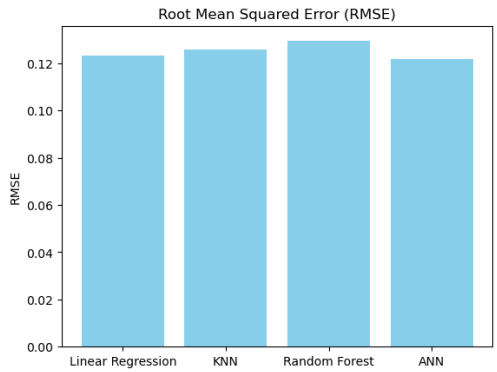
Evaluasi

R-squared dan Root Mean Square Error (RMSE) digunakan menemukan model regresi yang memiliki kinerja terbaik. Tabel 2 di bawah ini merangkum hasil pemodelan regresi.

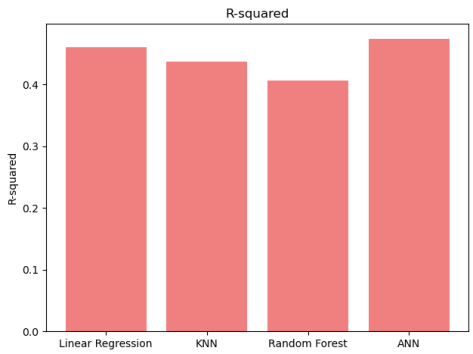
Tabel 2. Hasil Evaluasi

Model	RMSE	R2 Score
Linear Regression	0,123	0,460
K-Nearest Neighbors	0,126	0,437
Random Forest	0,129	0,406
Artificial Neural Network	0,122	0,475

Berdasarkan tabel 2, metode R2 dan RMSE menunjukkan bahwa model terbaik adalah *Artificial Neural Netwok* (ANN) dengan nilai R2 = 0,475 (47,5%) dan RMSE = 0,122. Model ANN memiliki RMSE terkecil dan R2 terbesar. Kemudian diikuti oleh *linear regression* dengan RMSE= 0,123 dan R2 = 0,460, *K-Nearest Neighbors* dengan RMSE = 0,126 dan R2 = 0,437, dan yang terakhir adalah *Random Forest* dengan RMSE=0,129 dan R2=0,406. Visualisasi hasil evaluasi menggunakan metode *R-squared* (R2) dan RMSE dapat dilihat pada Gambar 12 dan 13.



Gambar 12. Visualisasi Hasil Perhitungan RMSE



Gambar 13. Visualisasi Hasil Perhitungan *R-squared*

Berdasarkan hasil pada gambar 12 diats, model ANN memiliki nilai RMSE paling rendah, yang mengindikasikan bahwa model ini mampu menghasilkan prediksi dengan tingkat kesalahan paling kecil dibandingkan model lainnya. Linear Regression juga menunjukkan performa yang cukup baik dengan nilai RMSE yang relatif rendah, diikuti oleh KNN dengan selisih yang tidak terlalu signifikan. Sebaliknya, Random Forest memiliki nilai RMSE tertinggi, yang menunjukkan bahwa model ini menghasilkan kesalahan prediksi yang lebih besar dibandingkan ketiga model lainnya.

Sementara itu, gambar 13 menunjukkan nilai R-squared memberikan gambaran kemampuan masing-masing model dalam menjelaskan variasi data. Model ANN kembali menunjukkan performa terbaik dengan nilai R-squared tertinggi, yang berarti model ini paling mampu menangkap pola dan hubungan dalam data. Linear Regression berada pada posisi kedua dengan nilai yang cukup kompetitif, diikuti oleh KNN dengan performa yang sedikit lebih rendah. Random Forest kembali menunjukkan hasil terendah pada metrik ini, yang mengindikasikan bahwa kemampuannya dalam menjelaskan variabilitas data masih lebih lemah dibandingkan model lainnya. Secara keseluruhan, kedua metrik evaluasi ini secara konsisten menunjukkan bahwa ANN merupakan model dengan performa terbaik dalam penelitian ini.

KESIMPULAN

Penelitian ini membandingkan empat model regresi, yaitu *Linear Regression, Artificial Neural Network* , *K-Nearest Neighbor, dan Random Forest,* untuk memprediksi evaporasi berdasarkan parameter cuaca seperti suhu, radiasi matahari, tekanan udara, kelembaban relatif (RH), dan kecepatan angin yang diobservasi di Stasiun Klimatologi Banten. Model-model ini dievaluasi menggunakan *R-squared* dan *Root Mean Squared Error* (RMSE). Hasil dair penelitian ini menunjukkan bahwa *Artificial Neural* Network (ANN) adalah model terbaik, dengan nilai RMSE sebesar 0,122 dan *R-squared* sebesar 0,475 (47,5%). Kemudian diikuti oleh *linear regression, K-Nearest Neighbors,* dan *Random Forest* secara berturut – turut.

DAFTAR PUSTAKA

Abed, M., Imteaz, M., Ali Najah Ahmed, A.-M., & Huang, Y. (2022). Modelling monthly pan evaporation utilising Random Forest and deep learning algorithms. *Scientific Reports*, *12*, 13132. https://doi.org/10.1038/s41598-022-17263-3

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

Anandaraj, S., Rooby, J., Beerala, A., Mulukalla, V., & Koduri, S. (2018). *Strength prediction using ANN for concrete with Marble and Quarry dust*. https://doi.org/10.1109/I2C2SW45816.2018.8997326

Ardiansyah, D. (2023). Perbandingan Model Prediksi Radiasi Matahari Berbasis Mesin Pembelajaran Pada Stasiun Meteorologi Fatmawati Soekarno Bengkulu. *Megasains*, *14*(1). https://doi.org/10.46824/megasains.v14i1.129

Cosenza, D., Korhonen, L., Maltamo, M., Packalen, P., Strunk, J., Næsset, E., Gobakken, T., Soares, P., & Tomé, M. (2020). Comparison of linear regression, k-nearest neighbour and random forest methods in airborne laser-scanning-based prediction of growing stock. *Forestry*, 1–13. https://doi.org/10.1093/forestry/cpaa034

Fan, J., Chen, B., Wu, L., Zhang, F., Lu, X., & Xiang, Y. (2018). Evaluation and development of temperature-based empirical models for estimating daily global solar radiation in humid regions. *Energy*, *144*, 903–914. https://doi.org/https://doi.org/10.1016/j.energy.2017.12.091

Singh, V. P. (1988). *Hydrologic systems* (Number v. 2). Prentice Hall. https://books.google.co.id/books?id=ccsZyAEACAAJ

Vicente-Serrano, S. M., Bidegain, M., Tomas-Burguera, M., Dominguez-Castro, F., El Kenawy, A., McVicar, T. R., Azorin-Molina, C., López-Moreno, J. I., Nieto, R., Gimeno, L., & Giménez, A. (2018). A comparison of temporal variability of observed and model-based pan evaporation over Uruguay (1973–2014). *International Journal of Climatology*, *38*(1), 337–350. https://doi.org/10.1002/joc.5179