

## STUDI KOMPARATIF MODEL TIME SERIES UNTUK PERAMALAN SUHU DAN ANALISIS KORELASI FAKTOR IKLIM MENGGUNAKAN METODOLOGI CRISP-DM

Angela Melia Gunawan<sup>1</sup>, Sharon Tan<sup>2</sup>, Christopher Andreas<sup>\*3</sup>

<sup>1,2,3</sup> Universitas Ciputra Surabaya, Surabaya

Email: <sup>1</sup>agunawan18@student.ciputra.ac.id, <sup>2</sup>sharontan01@student.ciputra.ac.id,

<sup>3</sup>christopher.andreas@ciputra.ac.id

\*Penulis Korespondensi

(Naskah masuk: 14 Juli 2025, diterima untuk diterbitkan: 03 Maret 2026)

### Abstrak

Perubahan iklim global telah memicu kebutuhan akan metode peramalan suhu yang lebih akurat dan adaptif. Penelitian ini bertujuan untuk mengevaluasi dan membandingkan performa tiga model peramalan dalam memprediksi suhu rata-rata bulanan di Kota Delhi, India. Penelitian ini menggunakan pendekatan *time series forecasting* berbasis statistik (SARIMA dan Holt-Winters Additive) serta *machine learning* (Random Forest). Selain itu, dilakukan analisis korelasi menggunakan metode Pearson Correlation untuk mengukur hubungan linier antara suhu dengan variabel iklim lain seperti kelembapan, kecepatan angin, dan tekanan udara. Penelitian ini mengadopsi metodologi CRISP-DM yang mencakup tahap *business understanding*, *data understanding*, *data preparation*, *modeling*, dan *evaluation*. Dataset yang digunakan bersumber dari Kaggle dan mencakup data harian dari tahun 2013 hingga 2017 yang diolah menjadi data bulanan. Hasil evaluasi menunjukkan bahwa model SARIMA memiliki performa terbaik dengan nilai MSE sebesar 1.613 dan MAPE sebesar 3.73%, yang tergolong prediksi akurasi tinggi. Analisis korelasi Pearson menunjukkan bahwa kelembapan dan tekanan udara memiliki korelasi negatif terhadap suhu rata-rata, sedangkan kecepatan angin menunjukkan korelasi positif, dan seluruh hubungan tersebut signifikan secara statistik. Hasil ini mengindikasikan bahwa variasi suhu tidak berdiri sendiri, melainkan dipengaruhi oleh interaksi kompleks antar faktor iklim. Temuan ini menegaskan pentingnya mempertimbangkan variabel pendukung dalam pengembangan model peramalan suhu agar hasil prediksi menjadi lebih komprehensif. Secara keseluruhan, hasil penelitian ini berkontribusi pada pemilihan model peramalan yang optimal untuk wilayah tropis dan mendukung pengambilan keputusan berbasis data dalam mitigasi perubahan iklim dan perencanaan kota.

**Kata kunci:** metodologi CRISP-DM, prediksi iklim, SARIMA, holt-winters additive, random forest, analisis korelasi

## COMPARATIVE STUDY OF TIME SERIES MODELS FOR TEMPERATURE FORECASTING AND CORRELATION ANALYSIS OF CLIMATE FACTORS USING CRISP-DM METHODOLOGY

### Abstract

Global climate change has triggered the need for more accurate and adaptive temperature forecasting methods. This study aims to evaluate and compare the performance of three forecasting models in predicting the monthly average temperature in Delhi City, India. This study uses a statistical time series forecasting approach (SARIMA and Holt-Winters Additive) and machine learning (Random Forest). In addition, a correlation analysis was performed using the Pearson Correlation method to measure the linear relationship between temperature and other climate variables such as humidity, wind speed, and air pressure. This research adopts the CRISP-DM methodology which includes the stages of business understanding, data understanding, data preparation, modeling, and evaluation. The dataset used is sourced from Kaggle and includes daily data from 2013 to 2017 which is processed into monthly data. The evaluation results show that the SARIMA model has the best performance with an MSE value of 1.613 and a MAPE of 3.73%, which is classified as a high accurate prediction. Pearson's correlation analysis shows that humidity and air pressure have a negative correlation with temperature, while wind speed shows a positive correlation, and all of these relationships are statistically significant. These results indicate that temperature variations do not occur independently, but are influenced by complex interactions between climate factors. These findings emphasize the importance of considering supporting variables in the development of temperature forecasting models in order to produce more comprehensive predictions. Overall, the

results of this study contribute to the selection of optimal forecasting models for tropical regions and support data-driven decision-making in climate change mitigation and urban planning.

**Keywords:** CRISP-DM methodology, climate prediction, SARIMA, holt-winters additive, random forest, correlation analysis

1. PENDAHULUAN

Perubahan iklim global telah menjadi perhatian utama dunia, ditandai oleh peningkatan kejadian suhu ekstrem, kekeringan berkepanjangan, dan pergeseran pola cuaca yang signifikan (Anggraeni dkk., 2023). India, sebagai negara beriklim tropis dengan tingkat kepadatan penduduk yang tinggi dan urbanisasi yang masif, menjadi salah satu wilayah yang paling rentan terhadap dampak perubahan iklim tersebut. Dalam konteks ini, peramalan cuaca, khususnya suhu, memegang peranan vital dalam berbagai sektor kehidupan, mulai dari perencanaan aktivitas masyarakat, pengelolaan energi, hingga perumusan strategi mitigasi risiko bencana berbasis iklim.

Seiring meningkatnya urgensi isu perubahan iklim, tren penelitian menunjukkan pergeseran dari metode statistik konvensional menuju pendekatan berbasis *Machine Learning* (ML). Model klasik seperti ARIMA dan Holt-Winters masih digunakan untuk menangkap pola musiman, namun berbagai studi komparatif terbaru menunjukkan bahwa model ML seperti Random Forest dan *neural network* sering memberikan performa lebih baik dalam menangani kompleksitas dan non-linearitas data iklim (Dhwani dkk., 2025). Selain itu, integrasi berbagai pendekatan serta analisis multivariat semakin banyak diterapkan untuk memahami hubungan antar variabel iklim dan meningkatkan akurasi prediksi (Utami dan Intan, 2024).

Data iklim memiliki karakteristik kompleks yang bersifat musiman, non-linier, dan sering kali dipengaruhi oleh berbagai faktor eksternal seperti kelembapan, tekanan atmosfer, dan kecepatan angin. Namun, pendekatan konvensional dalam peramalan suhu seringkali kurang efektif, kurang akurat, dan memerlukan waktu pemrosesan yang lama, sehingga menimbulkan tantangan tersendiri dalam menghasilkan prediksi yang andal (Shaiba, 2022). Keterbatasan tersebut mendorong perlunya eksplorasi metode yang lebih adaptif dan presisi, salah satunya melalui penerapan metode *Machine Learning* (ML). Dengan kemampuan ML dalam menangani data kompleks dan pola non-linear, penelitian ini menjadi urgensi untuk mengembangkan model prediksi suhu yang akurat sebagai kontribusi terhadap peningkatan sistem peringatan dini dan pengambilan keputusan berbasis data.

Beberapa penelitian sebelumnya telah menunjukkan bahwa pendekatan modern mampu meningkatkan akurasi prediksi suhu secara signifikan.

Tabel 1. Penelitian Terdahulu

| Penulis              | Metode          | Fokus dan Kontribusi Penelitian                                                                                                                                                                                                                                                     |
|----------------------|-----------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Li dan Qian (2024)   | Hybrid CNN-LSTM | Mengembangkan model deep learning hibrida untuk prediksi suhu harian di Delhi menggunakan data historis jangka panjang (1996–2017). CNN mengekstraksi pola lokal, LSTM menangkap dependensi temporal. Menunjukkan performa unggul dengan MSE 3,26217 dan RMSE 1,80615.              |
| Shah dkk. (2024)     | SARIMAX         | Menggunakan model statistik berbasis faktor eksogen untuk memprediksi suhu maksimum bulanan di Ahmedabad (1993–2022). Mempertimbangkan variabel eksternal seperti polusi dan urbanisasi, dengan RMSE 1,0265. Menunjukkan efektivitas pendekatan statistik dengan variabel tambahan. |
| Elshewey dkk. (2023) | WD-SARIMAX      | Mengintegrasikan dekomposisi wavelet dan SARIMAX untuk menangani non-stasioneritas suhu harian di Delhi (2013–2017). Menghasilkan RMSE 1,67 dan R <sup>2</sup> 0,91, serta menunjukkan keunggulan pendekatan hibrida statistik-machine learning.                                    |

Namun, mayoritas studi sebelumnya hanya berfokus pada satu jenis pendekatan dan belum melakukan evaluasi komparatif yang menyeluruh antar model-model tersebut. Selain itu, keterkaitan antar variabel iklim seperti suhu, kelembapan, kecepatan angin, dan tekanan udara belum banyak dieksplorasi secara kuantitatif.

Urgensi penelitian ini adalah menyediakan model peramalan suhu yang akurat dan realistis untuk mendukung pengambilan keputusan berbasis data dalam kebijakan adaptasi dan perencanaan berkelanjutan di India, khususnya bagi sektor pemerintah, pertanian, energi, dan lingkungan.

Berdasarkan kesenjangan tersebut, penelitian ini dirancang untuk mengevaluasi dan membandingkan kinerja tiga model peramalan *time series*, yaitu SARIMA, Holt-Winters *Additive*, dan Random Forest, dalam memprediksi suhu rata-rata bulanan di Delhi, India. Penelitian ini juga menganalisis korelasi antar faktor iklim seperti kelembapan dan kecepatan angin dalam memengaruhi suhu, menggunakan pendekatan eksplorasi data dan statistik korelasi multivariat. Penelitian ini secara sengaja memilih untuk mengevaluasi model ketiga model tersebut karena jumlah data realtif terbatas serta telah diagregasi menjadi bulanan, sehingga model

kompleks berisiko *overfitting* dan kurang optimal. Selain itu, ketiga model tersebut dipilih untuk membandingkan pendekatan statistik dan *machine learning* yang lebih sederhana namun interpretatif, efisien secara komputasi, dan realistis untuk implementasi praktis.

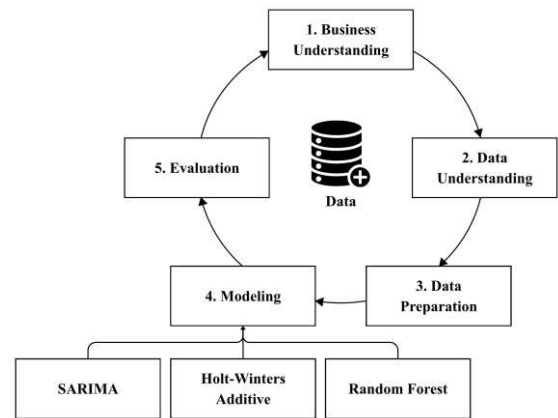
Dengan mengadopsi metodologi CRISP-DM, penelitian ini disusun secara sistematis mulai dari pemahaman data, eksplorasi, pemodelan, hingga evaluasi performa prediksi. Penelitian ini diharapkan dapat memberikan kontribusi dalam menentukan model *time series* yang paling efektif untuk peramalan suhu bulanan di wilayah perkotaan tropis, serta memperkaya literatur dengan pendekatan komparatif yang belum banyak diangkat. Hasil dari penelitian ini diharapkan mendukung pengambilan keputusan dalam kebijakan iklim, ketahanan energi, dan sistem peringatan dini berbasis data iklim lokal.

Kebaharuan penelitian ini terletak pada studi komparatif yang membandingkan model *time series* sederhana dan interpretatif dengan pendekatan *machine learning non-deep learning* dalam rangka CRISP-DM yang terstruktur dan *reproducible*, sehingga meskipun model yang digunakan bersifat umum, penelitian ini memberikan bukti empiris mengenai efektivitas model sederhana dibanding pendekatan yang lebih kompleks pada data suhu bulanan yang terbatas. Analisis korelasi dalam penelitian ini diposisikan sebagai analisis multivariat yang tidak hanya mengeksplorasi hubungan kuantitatif antar faktor iklim yang sering diabaikan dalam peramalan univariat, tetapi juga memberikan justifikasi ilmiah terhadap performa model melalui interpretasi berbasis teori meteorologi modern (Utami dan Intan, 2024.).

## 2. METODE PENELITIAN

Analisis dilakukan menggunakan metode *time series forecasting* dengan model SARIMA, Holt-Winters *additive*, dan *Random Forest* untuk menentukan model yang paling baik berdasarkan data iklim. Metode *time series forecasting* dengan akurasi tertinggi akan memberikan prediksi iklim yang lebih akurat. Selain itu, analisis korelasi juga dilakukan untuk melihat hubungan antara faktor-faktor iklim dan lebih memahami dinamika suhu serta pengaruh pendorongnya.

Penelitian ini dilakukan dengan menggunakan pendekatan CRISP-DM (*Cross Industry Standard Process for Data Mining*), yaitu metodologi standar dalam proses data mining yang terdiri dari enam tahapan utama (Schröer dkk., 2021). Namun, dalam konteks penelitian ini, proses dilakukan hingga tahap *evaluation*, tanpa mencakup tahap *deployment*. Adapun lima tahapan yang digunakan tampak dalam bagan alir berikut.



Gambar 1. Bagan Alir Metodologi CRISP-DM

### 2.1. Business Understanding

Data iklim harian merupakan data cuaca dan iklim yang dikumpulkan di kota Delhi, India. Data ini berisi informasi harian mengenai suhu rata-rata, kelembapan, kecepatan angin, dan tekanan udara. Dalam konteks analitis, dataset ini dimanfaatkan untuk pengembangan model peramalan suhu rata-rata bulanan serta untuk mengidentifikasi hubungan statistik antar variabel iklim yang diamati.

Penelitian ini bertujuan untuk mengevaluasi dan membandingkan beberapa model *time series* dalam memprediksi suhu rata-rata bulanan menggunakan data historis dan menganalisis korelasi antara faktor-faktor iklim. Secara teknis, permasalahan yang diselesaikan adalah menentukan model dengan tingkat akurasi terbaik berdasarkan metrik evaluasi yang digunakan. Luaran yang diharapkan berupa model peramalan dengan performa optimal serta informasi kuantitatif mengenai kekuatan dan arah hubungan antar variabel iklim.

### 2.2. Data Understanding

Data yang digunakan adalah data sekunder, yang diperoleh dari sumber data Kaggle. Dataset ini merupakan data iklim harian di kota Delhi, India, yang mencakup periode dari 1 Januari 2013 hingga 24 April 2017 (Rao, 2019). Data tersebut berisi 1576 baris dan 4 atribut atau kolom, antara lain sebagai berikut:

Tabel 2. Deskripsi Atribut pada Dataset

| Atribut             | Keterangan                                                                              |
|---------------------|-----------------------------------------------------------------------------------------|
| <i>date</i>         | Tanggal pengambilan data, dalam format YYYY-MM-DD.                                      |
| <i>meantemp</i>     | Suhu rata-rata harian, diperoleh dari beberapa pengukuran setiap 3 jam dalam satu hari. |
| <i>humidity</i>     | Kelembapan udara harian, diukur dalam satuan gram uap air per meter kubik udara.        |
| <i>wind_speed</i>   | Kecepatan angin harian, diukur dalam kilometer per jam (km/jam).                        |
| <i>meanpressure</i> | Tekanan udara rata-rata harian, diukur dalam satuan atmosfer (atm).                     |

Dalam penelitian ini, dilakukan pemilihan variabel berdasarkan tujuan analisis yang akan dilakukan. Untuk pemodelan *time series forecasting*,

variabel target ( $y$ ) yang diprediksi adalah suhu rata-rata bulanan (*meantemp*), sedangkan variabel deret waktu ( $t$ ) adalah tanggal pengambilan data (*date*), yang menunjukkan urutan kronologis data. Pemilihan *meantemp* sebagai target didasarkan pada pentingnya informasi suhu dalam konteks perubahan iklim dan perencanaan mitigasi. Sementara itu, untuk analisis korelasi antar variabel iklim, variabel independen ( $x$ ) yang digunakan adalah kelembapan udara (*humidity*), kecepatan angin (*wind\_speed*), dan tekanan rata-rata (*meanpressure*), dengan *meantemp* tetap menjadi variabel dependen ( $y$ ). Analisis ini bertujuan untuk mengevaluasi sejauh mana variabel kelembapan dan kecepatan angin berkaitan secara statistik terhadap suhu rata-rata bulanan, yang dapat memberikan pemahaman lebih lanjut mengenai interaksi antar faktor iklim.

*Exploratory Data Analysis* (EDA) juga dilakukan dengan tujuan untuk memahami karakteristik data. Pada tahap ini, dilakukan pengecekan terhadap kualitas data seperti keberadaan nilai yang hilang (*missing values*), keunikan data, dan distribusi nilai pada setiap variabel. Selain itu, EDA juga digunakan untuk mengidentifikasi pola, tren, dan hubungan antar variabel yang dapat memberikan wawasan awal sebelum dilakukan pemodelan atau analisis lebih lanjut.

Dalam dataset ini, dilakukan pemeriksaan terhadap seluruh variabel. Analisis meliputi visualisasi statistik deskriptif, distribusi nilai, serta korelasi antar variabel untuk mendapatkan gambaran umum kondisi cuaca yang tercatat. Proses ini membantu dalam menentukan langkah *preprocessing* data selanjutnya.

Analisis korelasi dalam penelitian ini menggunakan metode *Pearson Correlation* dan diposisikan sebagai analisis eksploratif dan interpretatif untuk memahami hubungan linier antar variabel iklim sebagai pelengkap hasil peramalan, bukan sebagai metode seleksi fitur dalam pemodelan.

### 2.3. Data Preparation

Setelah melalui tahap eksplorasi awal, data selanjutnya diproses untuk keperluan pemodelan dengan tujuan memastikan kualitas, konsistensi, dan format yang sesuai. Tahap ini mencakup berbagai transformasi data, antara lain konversi tipe data, penanganan duplikasi, identifikasi dan penanganan *outlier* pada variabel tertentu, penetapan indeks waktu, serta pemeriksaan terhadap keberadaan nilai hilang (*missing values*) pada dataset yang digunakan.

Data iklim harian yang telah dibersihkan kemudian diagregasi ke dalam frekuensi bulanan dengan titik acuan akhir bulan (*month-end*, *ME*). Proses agregasi dilakukan dengan menghitung nilai rata-rata bulanan untuk setiap variabel yang diamati, yaitu *meantemp*, *humidity*, *wind\_speed*, dan *meanpressure*. Dataset hasil *resampling* ini selanjutnya digunakan sebagai dasar dalam pemodelan deret waktu (*time series*).

Variabel target utama dalam pemodelan adalah suhu rata-rata (*meantemp*), yang diekstraksi sebagai variabel  $y$ . Variabel ini kemudian dibagi menjadi data latih (*training set*) dan data uji (*testing set*) dengan proporsi 80% dan 20%. Pembagian dilakukan tanpa pengacakan guna mempertahankan urutan temporal data, yang esensial dalam analisis deret waktu.

### 2.4. Modeling

Penelitian ini menggunakan metode *time series* dan *regression*. Metode *time series forecasting* yang digunakan adalah SARIMA, Holt-Winters *additive*, dan *Random Forest*. Pada tahap ini, dilakukan proses pembentukan model dengan menggunakan masing-masing metode yang telah dipilih. Selanjutnya, setiap metode yang dipilih dibangun dan dilatih (*training*) menggunakan data historis yang telah melalui tahap *preprocessing* sebelumnya.

Proses ini melibatkan pemilihan parameter yang optimal untuk masing-masing model. Untuk metode SARIMA, pemilihan parameter order ( $p$ ,  $d$ ,  $q$ ) dan musiman ( $P$ ,  $D$ ,  $Q$ )<sup>8</sup> dilakukan secara otomatis menggunakan pendekatan *auto\_arima*, yang mengevaluasi berbagai kombinasi parameter berdasarkan kriteria seperti AIC (*Akaike Information Criterion*) untuk menemukan konfigurasi terbaik. Sedangkan, pada metode Holt-Winters *additive*, proses pelatihan menggunakan fungsi *model.fit*, dimana parameter *smoothing* diestimasi secara otomatis oleh algoritma selama proses *fitting* untuk meminimalkan kesalahan prediksi.

Berbeda dengan kedua metode tersebut, *Random Forest* pada dasarnya merupakan algoritma *supervised learning* yang umum digunakan untuk regresi dan klasifikasi, sehingga tidak secara langsung memodelkan dependensi waktu. Oleh karena itu, dalam penelitian ini data deret waktu ditransformasikan terlebih dahulu ke dalam bentuk *supervised learning* melalui pembentukan *lag features*. Pada metode ini, parameter seperti jumlah pohon ( $n\_estimators$ ) dan kedalaman maksimum pohon ( $max\_depth$ ) menggunakan nilai *default*, sehingga model dilatih tanpa penyesuaian eksplisit terhadap parameter tersebut.

Setelah itu, ketiga model tersebut kemudian digunakan untuk melakukan prediksi pada data *testing*. Data *testing* ini merupakan data yang dipisahkan dari data *training* dan tidak pernah dilibatkan dalam proses *training*, sehingga dapat digunakan untuk menguji kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya. Proses prediksi ini bertujuan untuk menghasilkan estimasi nilai variabel target pada periode waktu mendatang, yang kemudian akan dianalisis lebih lanjut pada tahap evaluasi.

### 2.5. Evaluation

Tujuan akhir dari penelitian ini adalah untuk melakukan peramalan serta menganalisis korelasi

antar variabel yang diteliti. Peramalan dilakukan berdasarkan model terbaik yang diperoleh dari perbandingan akurasi model SARIMA, Holt-Winters *additive*, dan *Random Forest*. Dalam hal ini, ada 2 indikator yang digunakan untuk mengevaluasi kemampuan dan keakuratan model. Indikator tersebut adalah MSE dan MAPE.

Jika  $y_t$  adalah nilai aktual pada waktu  $t$  dan  $\hat{y}_t$  adalah nilai estimasi pada waktu  $t$  yang diperoleh dari hasil pemodelan, maka *error* yang dihasilkan dapat dituliskan sebagai berikut (Andreas dkk., 2021 dalam Sediono dkk., 2023):

$$\epsilon_t = y_t - \hat{y}_t \quad (1)$$

Selanjutnya, ukuran MSE dapat didefinisikan dimana  $n$  adalah jumlah data deret waktu sebagai berikut:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_t - \hat{y}_t)^2 \quad (2)$$

Ukuran akurasi model juga dapat dilihat dari nilai MAPE. Nilai MAPE dapat dituliskan sebagai berikut:

$$MAPE = \left( \frac{1}{n} \sum_{i=1}^n \frac{e_t}{y_t} \right) \cdot 100\% \quad (3)$$

Nilai MAPE memiliki interpretasi di setiap rentang nilai yang diperoleh. Interpretasi tersebut disajikan dalam Tabel 3 (Moreno dkk., 2013 dalam Sediono dkk., 2023).

Tabel 3. Kriteria Nilai MAPE

| Nilai MAPE      | Interpretasi           |
|-----------------|------------------------|
| Dibawah 10%     | Prediksi akurat tinggi |
| Diantara 10-20% | Prediksi baik          |
| Diantara 20-50% | Masuk akal             |
| Diatas 50%      | Tidak akurat           |

### 3. HASIL DAN PEMBAHASAN

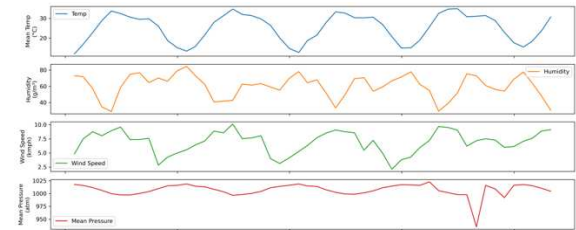
Analisis statistik deskriptif terhadap data dilakukan untuk memperoleh gambaran awal mengenai karakteristik masing-masing variabel. Tabel 4 menyajikan ringkasan statistik dari empat atribut pada data, yaitu suhu rata-rata udara (*meantemp*), kelembapan (*humidity*), kecepatan angin (*wind\_speed*), dan tekanan rata-rata (*meanpressure*).

Tabel 4. Statistik Deskriptif pada Dataset

|       | Mean Temp (°C) | Humidity (g/m <sup>3</sup> ) | Wind Speed (km/jam) | Mean Pressure (atm) |
|-------|----------------|------------------------------|---------------------|---------------------|
| Count | 52             | 52                           | 52                  | 52                  |
| Mean  | 25.22          | 60.29                        | 6.92                | 1007.22             |
| Std   | 7.01           | 14.23                        | 1.96                | 12.8                |
| Min   | 12.07          | 28.94                        | 2.12                | 934.95              |
| 25%   | 18.81          | 53.4                         | 5.54                | 1001.32             |
| 50%   | 27.43          | 62.62                        | 7.32                | 1009.53             |
| 75%   | 30.94          | 70.94                        | 8.58                | 1015.67             |
| Max   | 34.95          | 84.57                        | 10.11               | 1023.08             |

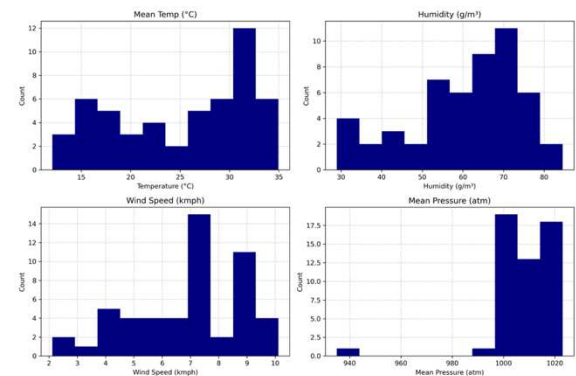
Plot *time series* untuk atribut-atribut pada dataset dapat dilihat pada Gambar 2. Hubungan antara tanggal dari tahun 2013 hingga 2017 dalam analisis deret waktu ini menunjukkan variabilitas

suhu, kelembapan, kecepatan angin, dan tekanan rata-rata. Suhu udara bervariasi antara 12.07-34.95°C, kelembapan 28.94-84.57 g/m<sup>3</sup>, kecepatan angin 2.12-10.11 km/jam, dan tekanan rata-rata antara 934.95-1023.08 atm.



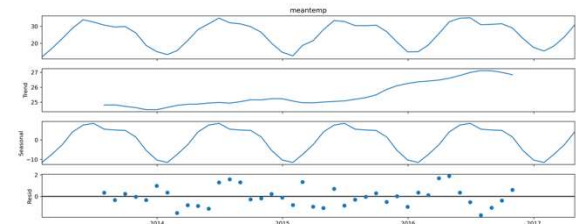
Gambar 2. Plot *Time Series* pada Dataset

Untuk eksplorasi atribut dalam dataset, histogram pada Gambar 3 menunjukkan bahwa suhu relatif merata dengan puncak pada 30-35°C. Kelembapan dominan pada 60-75 g/m<sup>3</sup> dan cenderung ke kanan. Kecepatan angin paling sering berada pada 6-9 km/jam, sedangkan tekanan udara terkonsentrasi di kisaran 1000-1020 atm dengan sedikit nilai ekstrem.



Gambar 3. Visualisasi Histogram pada Dataset

Selain itu, dekomposisi atribut *meantemp* pada Gambar 4 menunjukkan adanya tren meningkat dan pola musiman yang konsisten. Pola ini mengindikasikan bahwa data tidak stasioner dan dipengaruhi oleh fluktuasi musiman. Oleh karena itu, digunakan tiga model yang mampu menangani tren dan musiman, yaitu SARIMA, Holt-Winters *Additive*, dan *Random Forest*, untuk dibandingkan dalam memprediksi suhu.

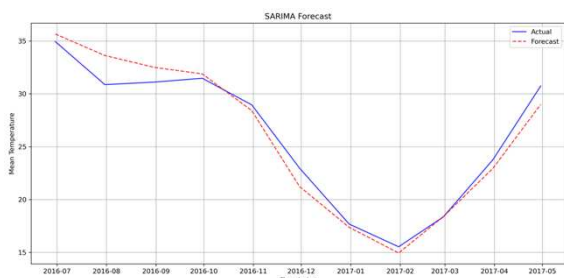


Gambar 4. Visualisasi Dekomposisi Data pada Atribut *Meantemp*

### 3.1. Model SARIMA

Pemodelan SARIMA mencakup dua kelompok parameter, yaitu komponen non-musiman ( $p, d, q$ ) dan komponen musiman ( $P, D, Q$ )<sup>s</sup>. Dalam penelitian ini, pemilihan parameter dilakukan berdasarkan evaluasi karakteristik data, seperti stasioneritas dan pola musiman yang berulang setiap 12 periode ( $s = 12$ ). Hasil identifikasi menunjukkan bahwa model terbaik adalah ARIMA(1, 0, 0)(2, 1, 1)[12] dengan *intercept*. Nilai  $d = 0$  menunjukkan bahwa data sudah stasioner secara tren, sedangkan nilai  $D = 1$  mengindikasikan bahwa diperlukan *differencing* satu kali untuk mengatasi pola musiman. Komponen *autoregressive* dan *moving average* pada bagian musiman dan non-musiman masing-masing menunjukkan kontribusi signifikan dalam membentuk model yang mampu menangkap pola historis data secara akurat.

Gambar 5 menunjukkan perbandingan antara nilai aktual dan hasil prediksi dari model SARIMA terhadap data suhu rata-rata. Secara umum, model SARIMA mampu mengikuti pola musiman dan tren data dengan cukup baik. Model berhasil menangkap penurunan suhu yang signifikan pada akhir tahun dan peningkatan Kembali pada awal hingga pertengahan tahun berikutnya. Meskipun terdapat sedikit perbedaan pada beberapa titik, secara keseluruhan prediksi menunjukkan tingkat kesesuaian yang baik dengan nilai aktual.



Gambar 5. Visualisasi Hasil Prediksi Model SARIMA

Hasil estimasi parameter model SARIMA ditampilkan pada Tabel 5, yang menunjukkan parameter signifikansi dan *p-value* masing-masing parameter. Dari tabel tersebut, terlihat bahwa hanya parameter  $ar.L1$  (*autoregressive lag* ke-1 non-musiman) yang signifikan secara statistik, dengan *p-value* sebesar 0.049 (lebih kecil dari 0.05). Hal ini menunjukkan bahwa nilai masa lalu satu periode sebelumnya berkontribusi signifikan terhadap nilai saat ini. Sebaliknya, parameter musiman seperti  $ar.S.L12$ ,  $ar.S.L24$ , dan  $ma.S.L12$  memiliki *p-value* yang lebih besar dari 0.05, yang mengindikasikan bahwa kontribusinya terhadap model tidak signifikan secara statistik. Meskipun demikian, parameter-parameter tersebut tetap digunakan dalam model untuk menangkap pola musiman yang kompleks.

Model memiliki nilai MSE sebesar 1.613, sehingga nilai RMSE adalah sebesar 1.27. Hal ini menunjukkan bahwa rata-rata deviasi prediksi dari

nilai aktual adalah sekitar 1.27°C. Nilai ini mengindikasikan bahwa model menghasilkan *error* yang cukup rendah secara umum.

Tabel 5. Ringkasan Hasil Estimasi Parameter Model SARIMA

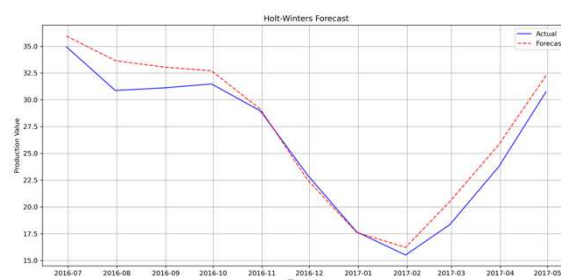
|                 | Parameter Signifikansi | <i>p-value</i> |
|-----------------|------------------------|----------------|
| <b>ar.L1</b>    | Signifikan             | 0.049          |
| <b>ar.S.L12</b> | Tidak signifikan       | 1.000          |
| <b>ar.S.L24</b> | Tidak signifikan       | 1.000          |
| <b>ma.S.L12</b> | Tidak signifikan       | 1.000          |

Sementara itu, nilai MAPE sebesar 0.0373 atau 3.73% menunjukkan bahwa kesalahan rata-rata prediksi terhadap nilai aktual berada di bawah 5%, yang dapat dikategorikan sebagai tingkat kesalahan yang rendah. Persentase nilai MAPE yakni 3.7% termasuk ke dalam kategori prediksi yang memiliki tingkat akurasi tinggi (*highly accurate forecasting*).

### 3.2. Model Holt-Winters Additive

Pemodelan Holt-Winters *additive* dilakukan untuk meramalkan data suhu rata-rata dengan mempertimbangkan adanya tren dan pola musiman yang bersifat aditif. Model ini mengandalkan tiga komponen utama, yaitu level (nilai dasar), tren (arah perubahan nilai dari waktu ke waktu), dan musiman (pola berulang dalam periode tertentu). Dalam penelitian ini, digunakan musiman dengan periode 12, sesuai dengan pola musiman tahunan pada data.

Hasil peramalan model Holt-Winters *Additive* ditampilkan pada Gambar 6, yang membandingkan nilai aktual dan hasil prediksi. Secara umum, model ini mampu mengikuti fluktuasi data aktual dengan baik, terutama dalam mengidentifikasi pola naik-turun yang terjadi secara periodik. Model berhasil menangkap penurunan tajam pada akhir tahun dan peningkatan kembali pada awal tahun berikutnya. Meskipun terdapat sedikit deviasi antara nilai aktual dan prediksi, bentuk kurva prediksi secara keseluruhan selaras dengan data historis.



Gambar 6. Visualisasi Hasil Prediksi Model Holt-Winters Additive

Tabel 6 menunjukkan ringkasan hasil estimasi parameter model Holt-Winters *additive*, yang menunjukkan nilai koefisien dari masing-masing parameter. Nilai *smoothing level* yaitu 0.377 menunjukkan bahwa model mempertimbangkan nilai aktual saat ini dalam memperbarui prediksi, namun masih memprioritaskan informasi dari pola historis. Parameter *smoothing trend* dan *seasonal* yang hampir



nol menunjukkan bahwa pola tren dan musiman sudah cukup stabil dan tidak terlalu fluktuatif.

Tabel 6. Ringkasan Hasil Estimasi Parameter Model Holt-Winters Additive

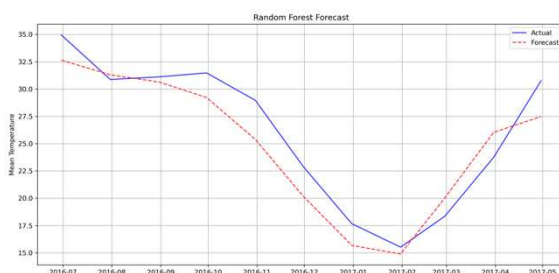
|                    | Nilai Koefisien          |
|--------------------|--------------------------|
| Smoothing level    | 0.3767575                |
| Smoothing trend    | $6.2009 \times 10^{-13}$ |
| Smoothing seasonal | $1.8375 \times 10^{-12}$ |

Nilai MSE untuk model ini adalah 2.354, yang setara dengan RMSE sebesar 1.53. Hal ini mengindikasikan bahwa deviasi rata-rata prediksi terhadap nilai aktual berada pada kisaran  $1.53^{\circ}\text{C}$ . Selain itu, nilai MAPE sebesar 0.0497 atau 4.97% menunjukkan bahwa rata-rata kesalahan relative prediksi terhadap nilai aktual berada di bawah ambang 5%. Dengan demikian, model ini termasuk dalam kategori peramalan dengan tingkat akurasi yang tinggi (*highly accurate forecasting*).

### 3.3. Model Random Forest

Model *Random Forest* digunakan untuk meramalkan suhu rata-rata berdasarkan nilai suhu sebelumnya yang dijadikan sebagai fitur input (*lag features*). Pendekatan ini memanfaatkan teknik *ensemble learning*, yang menggabungkan banyak pohon keputusan untuk meningkatkan akurasi dan stabilitas prediksi. Dalam penelitian ini, model dibangun menggunakan parameter  $n\_estimators = 100$ , yang berarti terdapat 100 pohon keputusan yang dibentuk dalam proses *ensemble*, serta  $random\_state = 42$  untuk memastikan reproduktibilitas hasil pelatihan. Sementara itu, parameter lainnya seperti *max\_depth*, *min\_samples\_split*, dan *min\_samples\_leaf* menggunakan nilai default dari library, sehingga model dilatih tanpa penyesuaian kompleks terhadap struktur pohon. Konfigurasi ini dipilih untuk menjaga keseimbangan antara performa prediksi dan efisiensi komputasi.

Gambar 7 memperlihatkan hasil perbandingan antara nilai aktual dan prediksi model. Secara umum, model mampu mengikuti pola tren dan musiman data suhu dengan cukup baik. Model dapat menangkap penurunan dan kenaikan suhu yang terjadi secara periodik, meskipun respons terhadap perubahan ekstrem cenderung lebih lambat dibandingkan model lainnya. Hal ini tampak pada perbedaan nilai prediksi yang cukup mencolok pada bulan November 2016 hingga Januari 2017, serta pada bulan April 2017.



Gambar 7. Visualisasi Hasil Prediksi Model *Random Forest*

Evaluasi performa menunjukkan bahwa model menghasilkan MSE sebesar 4.92, yang setara dengan RMSE sebesar  $2.22^{\circ}\text{C}$ . Sementara itu, MAPE sebesar 0.0778 atau 7.78% menunjukkan bahwa rata-rata kesalahan masih lebih besar dibandingkan 2 model sebelumnya, namun masih berada di dalam kategori peramalan dengan tingkat akurasi yang tinggi (*highly accurate forecasting*).

Tabel 7. Evaluasi Model

| Model                 | MSE   | MAPE   |
|-----------------------|-------|--------|
| SARIMA                | 1.613 | 0.037  |
| Holt-Winters Additive | 2.354 | 0.0497 |
| Random Forest         | 4.92  | 0.078  |

Pada Tabel 7 yang menunjukkan komparasi antara model SARIMA, Holt-Winters, dan *Random Forest*, ditemukan perbedaan kinerja yang signifikan. Model SARIMA menunjukkan akurasi prediksi tertinggi di antara ketiganya, dengan nilai MSE sebesar 1.613 dan nilai MAPE sebesar 3.73%. Nilai MAPE ini mengindikasikan kesalahan rata-rata sekitar 3.73% dari nilai aktual, menempatkan SARIMA dalam kategori peramalan dengan akurasi sangat tinggi dan menjadi pemodelan yang paling akurat dibandingkan 2 model lainnya.

Namun demikian, dari sisi waktu komputasi terdapat perbedaan yang cukup mencolok. Model SARIMA membutuhkan waktu komputasi paling lama, yaitu sekitar 17 detik untuk dataset ini. Sebaliknya, model Holt-Winters dan Random Forest memiliki waktu komputasi yang jauh lebih cepat, yakni kurang dari 1 detik. Hal ini menunjukkan adanya *trade-off* antara akurasi dan efisiensi waktu, di mana SARIMA unggul dalam ketepatan prediksi, tetapi membutuhkan waktu proses yang lebih lama dibandingkan dua model lainnya yang lebih efisien secara komputasi.

### 3.4. Analisis Korelasi

Berdasarkan Tabel 8, terdapat hubungan signifikan antara suhu rata-rata (*meantemp*) dengan ketiga variabel lainnya, yaitu kelembapan (*humidity*), kecepatan angin (*wind speed*), dan tekanan udara rata-rata (*meanpressure*). Koefisien korelasi antara *meantemp* dan *humidity* sebesar -0.611 menunjukkan korelasi negatif sedang menuju kuat, artinya ketika kelembapan meningkat, suhu cenderung menurun. Namun, hubungan ini tidak mutlak, karena kelembapan absolut merupakan jumlah uap air dalam suatu volume udara dan tidak dipengaruhi oleh suhu, kecuali jika terjadi proses kondensasi (Makruf dkk., 2025). Korelasi tersebut kemungkinan dipengaruhi faktor lain seperti tekanan udara, pembentukan awan, dan curah hujan. Tekanan udara rendah dapat memicu udara naik, terjadi pendinginan adiabatik, kondensasi, dan pembentukan awan yang berpotensi menurunkan suhu (Catursari, 2024).

Sebaliknya, *wind speed* memiliki koefisien korelasi positif sebesar 0.639, yang menandakan korelasi positif sedang menuju kuat, sehingga

semakin tinggi kecepatan angin, maka suhu rata-rata cenderung meningkat. Hal ini dapat dijelaskan melalui adveksi, yaitu perpindahan massa udara hangat ke wilayah pengamatan, serta pencampuran vertikal akibat turbulensi di lapisan batas atmosfer yang dapat mendistribusikan panas ke permukaan (Pal dkk., 2021).

Sementara itu, *meanpressure* memiliki korelasi negatif sebesar -0.577 terhadap *meantemp*, yang menunjukkan bahwa tekanan udara yang lebih tinggi cenderung berkaitan dengan suhu yang lebih rendah. Hal ini bertentangan dengan penelitian yang dilakukan oleh Fatmi (2020) yang menyebutkan bahwa semakin tinggi suatu tempat dari permukaan laut, maka tekanan udaranya akan semakin rendah. Tekanan udara yang rendah menyebabkan udara mengembang. Proses pengembangan udara ini mengakibatkan penurunan suhu melalui mekanisme pendinginan adiabatik. Oleh karena itu, semakin rendah tekanan udara, maka suhu udara juga cenderung semakin rendah.

Namun, perbedaan ini tidak serta-merta menunjukkan bahwa data statistik tersebut salah. Sebaliknya, korelasi negatif antara tekanan udara dan suhu rata-rata bisa dipengaruhi oleh faktor-faktor lain, seperti variasi musiman, sistem cuaca, tutupan awan, serta dinamika atmosfer lokal dan global.

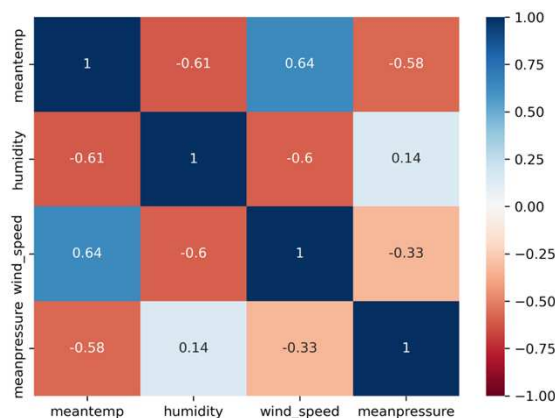
Tabel 8. Hasil Uji Korelasi

|                     |                    | <i>meantemp</i> |
|---------------------|--------------------|-----------------|
| <i>humidity</i>     | Koefisien korelasi | -0.611          |
|                     | <i>p-value</i>     | 0.000           |
| <i>wind_speed</i>   | Koefisien korelasi | 0.639           |
|                     | <i>p-value</i>     | 0.000           |
| <i>meanpressure</i> | Koefisien korelasi | -0.577          |
|                     | <i>p-value</i>     | 0.000           |

Seluruh hubungan ini signifikan secara statistik, sebagaimana ditunjukkan oleh *p-value* ketiga variabel sebesar 0.000, yang lebih kecil dari ambang batas signifikansi 0,05. Hasil ini memberikan indikasi bahwa faktor-faktor iklim seperti kelembapan, kecepatan angin, dan tekanan udara memiliki peran penting dalam memengaruhi suhu rata-rata di wilayah atau periode yang diamati. Hal ini diperkuat oleh visualisasi *heatmap* pada Gambar 8, yang menunjukkan pola warna konsisten dengan arah dan kekuatan korelasi antar variabel, dimana warna lebih gelap dan intens mencerminkan korelasi yang lebih kuat baik positif maupun negatif.

#### 4. KESIMPULAN

Berdasarkan hasil analisis, model SARIMA terbukti memberikan performa terbaik dibandingkan dua model lainnya, dengan nilai MSE sebesar 1.613 dan MAPE sebesar 3.73%, yang mengindikasikan tingkat kesalahan yang sangat rendah dan akurasi prediksi yang sangat tinggi. Hal ini menunjukkan bahwa SARIMA merupakan pilihan model yang paling optimal untuk peramalan suhu dalam konteks data yang digunakan.

Gambar 8. Visualisasi *Heatmap* pada Dataset

Selain itu, hasil uji korelasi menunjukkan bahwa suhu rata-rata memiliki hubungan yang signifikan dengan kelembapan, kecepatan angin, dan tekanan udara rata-rata. Kelembapan dan tekanan udara masing-masing menunjukkan korelasi negatif dengan suhu, sementara kecepatan angin menunjukkan korelasi positif. Seluruh korelasi ini signifikan secara statistik dengan *p-value* < 0.05. Temuan ini menunjukkan bahwa faktor-faktor iklim tersebut memainkan peran penting dalam memengaruhi variasi suhu dan dapat menjadi variabel pendukung yang relevan dalam model peramalan suhu.

Meskipun demikian, penelitian ini memiliki beberapa keterbatasan. Pertama, dataset yang digunakan terbatas pada periode tahun 2013–2017 sehingga belum mencerminkan dinamika perubahan iklim dalam jangka panjang. Kedua, data yang digunakan merupakan agregasi bulanan, sehingga variasi harian yang lebih detail tidak sepenuhnya tertangkap dalam model. Ketiga, model Random Forest tidak melalui proses tuning parameter secara mendalam, sehingga performanya mungkin belum optimal. Selain itu, penelitian ini belum memasukkan variabel eksternal lain seperti curah hujan, radiasi matahari, atau indeks polusi yang berpotensi memengaruhi suhu.

Berdasarkan keterbatasan tersebut, penelitian selanjutnya disarankan untuk menggunakan dataset dengan rentang waktu yang lebih panjang dan resolusi waktu yang lebih tinggi, seperti data harian atau mingguan, guna menangkap pola yang lebih kompleks. Selain itu, pengembangan model dapat dilakukan dengan pendekatan hybrid atau deep learning, seperti LSTM atau CNN-LSTM, serta penerapan teknik optimasi hiperparameter untuk meningkatkan performa model machine learning. Integrasi variabel eksogen tambahan juga direkomendasikan agar model peramalan suhu menjadi lebih komprehensif dan adaptif terhadap dinamika iklim yang semakin kompleks.

#### DAFTAR PUSTAKA

ALFAQIH, L., FADILAH, M. B. & DORES, A., 2025. Perbandingan Metode Simple Linear



- Regression, Polynomial Regression, dan K-Nearest Neighbors (KNN) untuk Prediksi Transaksi Perbankan. *Just IT : Jurnal Sistem Informasi, Teknologi Informasi, dan Komputer*, 15(2), pp.373-382.
- ANGGRAENI, N. M., SUDARTI & YUSHARDI, 2023. Analisis Dampak Perubahan Iklim dan Pola Angin Pada Lingkungan Global. *Jurnal Pendidikan, Sains dan Teknologi*, 2(4), pp.1041-1047. <https://doi.org/10.47233/jpst.v2i4.1366>
- ARYATI, A., PURNAMASARI, I. & NASUTION, Y. N., 2020. Peramalan dengan Menggunakan Metode Holt-Winters Exponential Smoothing (Studi Kasus: Jumlah Wisatawan Mancanegara yang Berkunjung Ke Indonesia). *Jurnal EKSPONENSIAL*, 11(1), pp.99-106. <https://doi.org/10.30872/eksponensial.v1i1l.650>
- CATURSARI, S. A., 2024. Analisis pengaruh tekanan udara, kelembaban udara, dan suhu udara, terhadap curah hujan di Kota Surabaya. Skripsi. Universitas Islam Negeri Maulana Malik Ibrahim.
- DHWANI, N., PRANAY, P. & SHARANYA, N., 2025. Comparative Analysis of Traditional Time Series and Machine Learning Models for Multi-Year Temperature Forecasting. *Indian Journal of Science and Technology*, 18(39), pp.3145-3158. <https://doi.org/10.17485/IJST/v18i39.1263>
- ELSHEWEY, A. M., SHAMS, M. Y., ELHADY, A. M., SHOHIEB, S. M., ABDELHAMID, A. A., IBRAHIM, A. & TAREK, Z., 2023. A Novel WD-SARIMAX Model for Temperature Forecasting Using Daily Delhi Climate Dataset. *Sustainability*, 15, p.757. <https://doi.org/10.3390/su15010757>
- FATMI, N., 2020. Tekanan Udara dalam Perspektif Sains dan Al-Qur'an. *Al-Madaris: Jurnal Pendidikan dan Studi Keislaman*, 1(1), pp.31-38. <https://doi.org/10.47887/amd.v1i1.6>
- LI, B. & QIAN, Y., 2024. Weather Prediction Using CNN-LSTM for Time Series Analysis: A Case Study on Delhi Temperature Data. <https://doi.org/10.48550/arXiv.2409.09414>
- LIND, D. A., MARCHAL, W. G. & WATHEN, S. A., 2018. *Statistical Techniques in Business and Economics*. 17th ed. New York: McGraw-Hill Education.
- MAKRUF, S. M., RAMADHANI, L. R., SANDHA, F., RINI, P. A. P., NAJWA, S. A., AMRULLAH, A. H., ANATASA, S. & PRASTIA, A. S., 2025. Analisis Kelembaban Udara terhadap Tingginya Suhu di Sekaran Semarang. *Jurnal Analis*, 4(1), pp.093-100.
- NAWAWI, A., HERAWATI, S. & PRASTITI, N., 2021. Implementasi Metode Holt Winter Additive untuk Prediksi Kunjungan wisatawan Nusantara Kabupaten Sumenep. *Jurnal SimanteC*, 10(1), pp.25-30. <https://doi.org/10.21107/simantec.v10i1.12466>
- PAL, S., CLARK, N. E., LEE, T. R., CONDER, M. & BUBAN, M., 2021. When and where horizontal advection is critical to alter atmospheric boundary layer dynamics over land: The need for a conceptual framework. *Atmospheric Research*, 264. <https://doi.org/10.1016/j.atmosres.2021.105825>
- RADHI, M., AMALIA, A., SITOMPUL, D. R. H., SINURAT, S. H. & INDRA, E., 2022. Analisis Big Data dengan Metode Exploratory Data Analysis (EDA) dan Metode Visualisasi menggunakan Jupyter Notebook. *Jurnal Sistem Informasi dan Ilmu Komputer*, 4(2), pp.23-27. <https://doi.org/10.34012/jurnalsisteminformasidanilmukomputer.v4i2.2475>
- RAO, S. V., 2019. *Daily Climate Time Series Data*. [online] Kaggle. Tersedia di: <https://www.kaggle.com/datasets/sumanthvrao/daily-climate-time-series-data> [Diakses 28 Mei 2025].
- SCHRÖER, C., KRUSE, F. & GÓMEZ, J. M., 2021. A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, 181, pp.526-534. <https://doi.org/10.1016/j.procs.2021.01.199>
- SESIONO, ANDREAS, C., MARDIANTO, M. F. F. & SULIYANTO, E. A., 2023. Forecasting the volume of electronic money transactions using ARIMAX-GARCH model and support vector regression. *AIP Conference Proceedings*, 2975(1), p.080013. <https://doi.org/10.1063/5.0187234>
- SHAIBA, H., MARZOUK, R., NOUR, M. K., NEGM, N. HILAL, A. M., MOHAMED, A., MOTWAKEL, A., YASEEN, I., ZAMANI, A. S. & RIZWANULLAH, M., 2022. Weather forecasting prediction using ensemble machine learning for big data applications. *Computers, Materials & Continua*, 73(2), pp.3367-3382. <http://dx.doi.org/10.32604/cmc.2022.030067>
- SHAH, V., PATEL, N., SHAH, D., SWAIN, D., MOHANTY, M., ACHARYA, B., GEROGIANNIS, V. C. & KANAVOS, A., 2024. Forecasting Maximum Temperature Trends with SARIMAX: A Case Study from Ahmedabad, India. *Sustainability*, 16, p.7183. <https://doi.org/10.3390/su16167183>

- THEOFANI, G. & SEDIYONO, E., 2022. Multiple Linear Regression Analysis on Factors that Influence Employees Work Motivation. *Sinkron : Jurnal Dan Penelitian Teknik Informatika*, 6(3), pp.791-798. <https://doi.org/10.33395/sinkron.v7i3.11453>
- UTAMI, W. D. & INTAN, P., 2024. Prediksi Parameter Klimatologi Menggunakan Multivariate Singular Spectrum Analysis (MSSA). *Jurnal Fourier*, 13(2), pp.65-75. <https://doi.org/10.14421/fourier.2024.132.65-75>
- WIBOWO, R. A. & KURNIAWAN, A. A., 2020. Analisis Korelasi dalam Penentuan Arah antar Faktor pada Pelayanan Angkutan Umum di Kota Magelang. *Theta Omega: Journal of Electrical Engineering, Computer and Information Technology*, 1(2).