E-ISSN : 2623-1700
P-ISSN : 1693-3672

# Penerapan Algoritma LightGBM Untuk Klasifikasi Hujan Harian (Studi Kasus : DKI Jakarta)

Anwar Septian[1], Farid Akbar[2], Daniel Meissel Yehezkiel[3], Zulfan Alden Nurafdi[4], Fachri Amsury[5], Riza Fahlapi[6].
Program studi teknologi informasi
Fakultas teknik informatika, universitas bina sarana informatika
17230303@bsi.ac.id[1], 17230311@bsi.ac.id[2], 17231074@bsi.ac.id[3], 17230219@bsi.ac.id[4], fachri.fcy@bsi.ac.id[5], riza.rzf@bsi.ac.id[6].

**Abstrak**

Cuaca dipengaruhi oleh beberapa faktor yaitu Suhu, Kelembaban Udara, Curah Hujan. Prediksi cuaca sangat diperlukan oleh masyarakat, untuk menjadi bahan pertimbangan perencanaan kegiatan yang akan dilakukan oleh masyarakat, khususnya bagi masyarakat yang tinggal di daerah kota Metropolitan seperti DKI Jakarta. Penelitian ini bertujuan untuk mengembangkan sistem prediksi cuaca multi-variabel 72 jam kedepan menggunakan algoritma LightGBM dengan menggunakan strategi rekuren. Data yang digunakan di penelitian ini adalah data cuaca historis per jam selama satu tahun total data yang digunakan di penelitian ini berjumlah 8.760 baris dari 5 titik grid di daerah Jabodetabek. Dan ada empat model LightGBM yang dilatih untuk memprediksi Suhu, Kelembaban, Curah Hujan, dan Status Hujan. Validasi model dilakukan menggunakan *Time Series Cross-Validation* (TSCV) *5-split*. Hasil penelitian menunjukan kinerja model yang sangat presisi pada horizon jangka pendek. Model klasifikasi status hujan mencapai *F1-Score* rata rata 0.930 dan *AUC* 0.954 pada prediksi 12 jam kedepan, yang mengindikasikan kemampuan deteksi hujan yang akurat. Dan model mengalami penuruan performa pada prediksi 72 jam kedepan akibat sifat strategi rekuren, meskipun mengalami penurunan performa, model tetap mempertahankan F1-Score yang solid sebesar 0.776. Penelitian ini menyimpulkan bahwa LightGBM dengan strategi rekuren sangat efektif sebagai sistem peringatan dini jangka pendek dengan akurasi tinggi.
**Kata kunci:** Prediksi Cuaca, LightGBM, Klasifikasi Hujan, Prediksi Rekuren.

**Abstract**

Weather is influenced by several key factors, namely Temperature, Humidity, and Precipitation. Accurate weather forecasting is essential for the public to plan daily activities, particularly for people living in metropolitan areas like DKI Jakarta. This study aims to develop a 72-hour ahead multi- variable weather prediction system using the Light Gradient Boosting Machine (LightGBM) algorithm with a recursive strategy. The study utilizes one year of hourly historical weather data, totaling 8,760 rows, collected from five grid points in the Jabodetabek area. Four separate LightGBM models were trained to predict Temperature, Humidity, Precipitation, and Rain Status. Model validation was performed using a 5-split Time Series Cross-Validation (TSCV). The results demonstrate highly precise model performance in the short-term horizon. The rain status classification model achieved an average F1-Score of 0.930 and an AUC of 0.954 at the 12-hour ahead prediction, indicating accurate rain detection capability. While the model experienced a performance decline at

E-ISSN : 2623-1700
P-ISSN : 1693-3672

the 72-hour ahead prediction due to the nature of the recursive strategy, it maintained a solid F1-Score of 0.776. Thisstudy concludes that LightGBM with a recursive strategy is highly effective as a high-accuracy short- term early warning system.**Keywords:** Weather Forecasting, LightGBM, Rain Classification, Recursive Prediction.

## PENDAHULUAN

Cuaca mempengaruhi kehidupan manusia diberbagai sektor, yang mana kondisi cuaca merupakan hal yang penting dan tidak akan lepas dari kehidupan manusia. Cuaca sangat sulit diprediksi karena banyak faktor sebagai variabel yang menentukan cuaca, tidak bisa hanya melihat kondisi udara pada waktu yang relatif singkat. Cuaca di pengaruhi oleh beberapa atribut yaitu, tekanan, kecepatan angin, curah hujan, suhu, dan fenomena atmosfer sebagai komponennya [1].

Pada tahun 1922, Lewis Fry Richardson menerbitkan buku berjudul "*Weather Prediction by Numerical Process*", buku itu menggambarkan visi tentang bagaimana persamaan dinamis atmosfer dapat digunakan untuk meramalkan cuaca, dibuku itu diceritakan memerlukan penggunaan "*forecast factory*" dengan tenaga kerja 64.000 *"computer"* manusia yang menghitung algoritma secara manual bersamaan. Walaupun Richardson gagal menghilangkan gelombang frekuensi tinggi dari perhitungannya sendiri, yang menyebabkan solusi perkiraan yang tidak realistis dan tidak stabil, tapi Richardson memiliki kemampuan memprediksi cuaca numerik (NWP) dan menunjukan bahwa

prediksi cuaca dapat dilakukan[2]. Aktivitas pemerintahan dan bisnis Indonesia yang terpusat di pulau Jawa khusus nya DKI Jakarta[3], dengan itu kami memilih Kota DKI Jakarta sebagai studi kasus, sebagai pusat pemerintahan dan bisnis, prediksi cuaca ini diharapkan membantu masyarakat untuk melakukan aktivitas sehari – hari. Seiring dengan kemajuan teknologi dan ketersediaan data yang semakin besar, pendekatan berbasis *machine learning* mulai banyak digunakan untuk meningkatkan akurasi serta efisiensi dalam proses prediksi cuaca[4].

Salah satu algoritma *machine learning* yang menunjukkan performa unggul dalam prediksi data kompleks adalah *Light Gradient Boosting Machine* (LightGBM). LightGBM merupakan algoritma *gradient boosting* yang dikembangkan oleh Microsoft dan dirancang untuk memberikan kecepatan pelatihan tinggi, efisiensi memori, serta hasil prediksi yang akurat[5]. Dengan memanfaatkan data

E-ISSN : 2623-1700
P-ISSN : 1693-3672

historis seperti suhu, kelembapan, tekanan udara, curah hujan, dan kecepatan angin, LightGBM dapat mempelajari pola dari data masa lalu untuk memprediksi kondisi cuaca di masa mendatang.

Dan oleh karena algoritma Light GBM ini memiliki beberapa keunggulan dibanding algoritma lain, kami memilih algoritma LightGBM ini untuk menghasilkan sistem klasifikasi prediksi cuaca yang lebih cepat dan akurat. Berdasarkan latar belakang tersebut, rumusan masalah dalam penelitian ini adalah "Bagaimana Performa algoritma LightGBM dalam memprediksi empat variabel cuaca (Suhu, Kelembaban, Curah Hujan, dan Status Hujan) secara rekuren untuk 72 jam kedepan, ditinjau dari metrik evaluasi regresi (RMSE/MAE) dan klasifikasi *(F1 Score)*".

## METODE PENELITIAN

### 1.1 Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode Prediksi Deret Waktu Rekuren *(Recursive Time Series Forecasting)* untuk memprediksi kondisi 72 jam kedepan secara per jam algoritma utama yang digunakan adalah *Light Gradient Boosting Machine* (LightGBM). Penelitian difokuskan pada empat variabel cuaca utama sebagai target prediksi yaitu: Suhu Udara, Kelembaban Relatif, Curah Hujan , Dan Status Hujan.

Setiap variabel diprediksi menggunakan model terpisah, sehingga total terdapat empat model LightGBM.

Penelitian ini dilakukan melalui beberapa tahap, yaitu pengumpulan data historis, pembersihan data, pembuatan fitur *(feature engineering)*, pembagian dataset, pelatihan model, validasi, dan evaluasi model.

### 1.2 Sumber dan Perolehan Data

Data penelitian diperoleh dari Open-Meteo Archive API, berupa data cuaca historis data per jam untuk wilayah DKI Jakarta, Tangerang, Bekasi, Bogor, dan Karawang.

Detail data sebagai berikut:

- Jumlah data : 8.760 baris (jumlah dari 365 hari × 24 jam)
- Periode data : satu tahun kebelakang dihitung dari tanggal terakhir model dijalankan
- Variabel dasar : Suhu Udara

E-ISSN : 2623-1700
P-ISSN : 1693-3672

*(temprature*)

, Kelembaban (*humidity)*, Kecepatan angin (*wind_speed_10m*), tekanan permukaan *(surface_pressure)* , curah hujan *(precipitation)*, radiasi gelombang pendek *(shortwave_ radiation).*

Data historis ini kemudian diproses dan di gunakan untuk pelatihan model. Pemisahan *training* dan *testing* dilakukan melalui skema *Time Series Cross-Validation (TSCV)* dengan 5 *splits*, yang secara inheren mempertahankan urutan waktu data untuk pengujian yang realistis

### 1.3 *Preprocessing Dat*a

#### 1.3.1 Pembersihan Data dan Pra-Pemrosesan Awal

Tahap ini bertujuan untuk memastikan kualitas dan konsistensi data sebelum masuk ke tahap *feature engineering*, dan berikut tahapan Pembersihan Data :

1. Konsolidasi dan pengukuran waktu : Data dari lima lokasi Grid di gabungkan berdasarkan *timestamp* dan diurutkan secara kronologis untuk memfasilitasi analisis deret waktu. Penyeragaman zona waktu (asia/jakarta) telah di lakukan pada tahap pengambilan data dari API Open Meteo.
2. Penangganan Nilai Hilang *(Missing Values)*: Nilai yang hilang (NaN) dalam data cuaca diisi menggunakan metode *Forward Fill (ffill)*, dimana nilai terakhir yang valid digunakan untuk mengisi celah waktu berikutnya. Metode ini diaplikasikan pada semua kolom cuaca.
3. Penghapusan Baris Target Kosong : Setelah proses *feature engineering* dan penentuan target prediksi satu jam ke depan (H+1), baris yang memiliki nilai target prediksi kosong (NaN), terutama di akhir dataset akibat pergeseran waktu, dihapus untuk memastikan hanya data yang lengkap yang digunakan untuk pelatihan model.

#### 1.3.2 Transformasi dan Normalisasi

Setelah tahap *feature engineering*, semua fitur input numerik yang akan digunakan untuk melatih model di

E-ISSN : 2623-1700
P-ISSN : 1693-3672

normalisasi. Normalisasi dilakukan menggunakan objek *StandardScaler* dari pustaka *Scikit-learn*. Tujuan dari penskalaan ini adalah untuk :

1. Menyeragamkan skala nilai antar variabel fitur.
2. Memastikan setiap fitur memberikan kontribusi yang setara dalam pelatihan model LightGBM.
3. Meningkatkan konvergensi dan performa model secara keseluruhan.

*StandardScaler* dilatih hanya pada data training di dalam *loop Time Series Cross-Validation (TSCV)*, dan kemudian digunakan untuk mentransformasi data *testing* untuk menghindari data *leakage.*

### 1.3.3 *Feature Engineering*

*Feature engineering* dilakukan untuk memperkaya representasi data sehingga model dapat belajar pola cuaca yang lebih kompleks Berikut daftar *Feature Engineering* dalam penelitian ini :

**a) Fitur Siklik Waktu (4 Kolom)**

Fitur ini mempresentasikan pola periodik waktu harian dan tahunan, penting untuk menangkap fluktuasi Suhu dan Kelembaban :

- sin(jam), cos(jam): mempresentasikan pola harian 24 jam.
- sin(hari dalam setahun), cos(hari dalam setahun): mempresentasikan pola musiman 1 tahun.

**b) Fitur Lagging (42 Kolom)**

Fitur *lagging* merepresentasikan kondisi cuaca beberapa jam dan hari sebelumnya untuk prediksi deret waktu .

- Horizon lag: 1, 6, 12, 24, 48, 72, 168 jam
- Variabel yang di cakup : Diterapkan pada semua variabel cuaca dasar (Suhu, Kelembaban, Curah hujan , Arah Angin dan Radiasi Gelombang

**c) Fitur *Rolling* Window**

Fitur ini berfungsi untuk merangkum data dalam satu periode waktu terakhir untuk melihat tren jangka pendek dan volatilitas.

- Jendela waktu: 24 jam, 72 jam, 168 jam
- Jenis Statistik : Dihitung untuk *Rolling mean* (rata - rata bergerak)

E-ISSN : 2623-1700
P-ISSN : 1693-3672

dan *Rolling Standard Deviation (Std)* (deviasi standard bergerak)

**d) Total Fitur**

Sebagai hasil dari proses *feature engineering* yang komprehensif, jumlah fitur input yang digunakan untuk melatih setiap model LightGBM berkisar 150 hingga 200 kolom. Kompleksitas fitur ini,

yang mencakup informasi *lagging* dan tren jangka panjang, memungkinkan model untuk menangkap pola temporal dan spasial yang lebih rumit, yang esensial untuk meningkatkan akurasi dalam prediksi cuaca rekuren.

### 1.4 Pembagian Dataset

Pembagian dataset dilakukan dengan metode yang dirancang Khusus untuk deret waktu yaitu *Time Series Cross-Validation (TSCV)*. Metode ini dipilih untuk menghindari data *leakage* dan memastikan bahwa model selalu diuji pada data yang secara kronologis lebih baru dibandingkan data pelatihan. Dalam implementasi ini:

1. Tidak Ada Pemisahan *Holdout Statis* (80:22) : seluruh dataset historis digunakan dalam proses TSCV, yang berfungsi ganda sebagai *training* dan *validation set*
2. Konfigurasi TSCV : Kami menggunakan objek *TimeSeries Split* dengan total 5 *splits*.
3. Proses Validasi: Pada setiap split, model dilatih pada data dari periode waktu yang lebih awal dan diuji (divalidasi) pada data dari periode waktu berikutnya. Metrik akhir (*F1 Score*, RMSE) dihitung sebagai rata-rata dari hasil kelima split tersebut, menghasilkan evaluasi kinerja model yang andal dan stabil.

### 1.5 Model Machine Learning

Penelitian ini menggunakan algoritma *Light Gradient Boosting Machine* (LightGBM). Alogritma ini dipilih karena keunggulannya pada kecepatan pelatihan , efisiensi memori, serta performa tinggi dalam

E-ISSN : 2623-1700
P-ISSN : 1693-3672

menangani dataset berukuran besar dan bertipe *sparse*.

LightGBM di implementasikan dalam skema *Multi Output Single Model*, dimana empat model terpisah dilatih untuk memprediksi empat variabel target utama :

Empat model LightGBM dilatih secara terpisah sebagai berikut:

a) Model 1 – Prediksi Suhu Udara (Regresi)
b) Model 2 – Prediksi Kelembaban Relatif (Regresi)
c) Model 3 – Prediksi Curah Hujan (Regresi) Memprediksi nilai numerik curah hujan dalam mm/jam
d) Model 4 – Prediksi Status Hujan (Klasifikasi Biner)

Pada model klasifikasi (*LGBMClassifier*), status hujan ditentukan sebagai:

- **1** = Hujan ketika Curah Hujan **> 0.0 mm /jam.**
- **0** = Tidak Hujan ketika Curah Hujan **= 0 mm/jam.**

Untuk mengatasi ketidakseimbangan kelas (*class imblance*) dalam data (dimana 'Tidak Hujan' lebih dominan), bobot kelas disesuaikan selama pelatihan untuk meningkatkan kemampuan model memprediksi kasus 'Hujan'

### 1.6 Evaluasi Model

Evaluasi model pada test set dalam *split Time Series Cross-Validation(TSCV).* Metrik evaluasi dibagi berdasarkan jenis tugas model:

#### 1.6.1 Untuk Model Regresi (Suhu, Kelembaban, Curah Hujan)

Kinerja model regresi diukur menggunakan metrik berbasis kesalahan, yang mengukur seberapa dekat prediksi model dengan nilai aktual *ground truth*:

- **MAE** *(Mean Absolute Error):* Mengukur rata-rata besaran kesalahan tanpa memperhatikan arah kesalahan.
- **RMSE** *(Root Mean Square Error):* Memberikan bobot lebih besar pada kesalahan yang besar dan merupakan metrik utama untuk menilai akurasi prediksi nilai.

#### 1.6.2 Untuk Model Klasifikasi (Status Hujan)

Kinerja model klasifikasi diukur

E-ISSN : 2623-1700
P-ISSN : 1693-3672

menggunakan metrik yang sensitif terhadap hasil benar positif dan salah positif, mengingat adanya ketidaksamaan kelas dalam data cuaca:

- *Accuracy* : Persentase prediksi yang benar secara keseluruhan .
- *Precision, Recall, F1-Score*: Metrik - metrik ini sangat penting untuk menilai kinerja prediksi hujan (kelas minoritas), dengan *F1-Score* berfungsi sebagai keseimbangan harmonis antara *Precision* dan *Recall*, menjadikanya metrik evaluasi utama
- *ROC-AUC(Receiver Operating Characteristic - Area Under the Curve)*: Mengukur kemampuan model untuk membedakan antara kelas positif (hujan) dan negatif(tidak hujan).

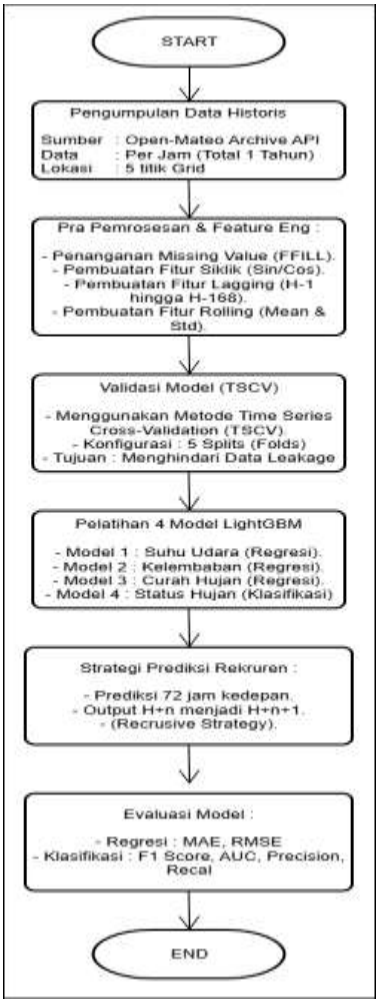
Sebagai evaluasi tambahan setelah pelatihan , Kurva Prediksi (garis prediksi 72 jam ) digunakan untuk analisis visual guna memvalidasi perilaku model pada prediksi waktu jangka panjang.

**1.7 Alur Penelitian**

Tahapan penelitian dapat diringkas sebagai beriku

E-ISSN : 2623-1700
P-ISSN : 1693-3672

Gambar 1. *Flowcart* Pembuatan Model




## HASIL PENELITIAN

Penelitian ini berhasil mengimplementasikan empat model LightGBM yang divalidasi menggunakan *Time Series Cross-Validation* (TSCV) 5-split. Evaluasi dilakukan pada prediksi waktu utama yaitu jangka pendek (12 jam), menengah (24 jam) dan panjang (72 jam)

### 1.8 KINERJA MODEL REGRESI

Kinerja model regresi seperti Suhu, Kelembaban, Curah Hujan, diukur menggunakan *RMSE* dan *MAE*. Berikut tabel evaluasi model :

Tabel 1. Tabel Evaluasi Model Regresi

E-ISSN : 2623-1700
P-ISSN : 1693-3672

| Variabel | Metrik | 12 Jam | 24 Jam | 72 Jam |
|---|---|---|---|---|
| Suhu | *RMSE* | 0.749 | 0.863 | 0.701 |
| | *MAE* | 0.518 | 0.580 | 0.452 |
| Kelembaban (%) | *RMSE* | 3.533 | 4.359 | 3.786 |
| | *MAE* | 2.883 | 2.988 | 2.587 |
| Curah Hujan (mm) | *RMSE* | 0.999 | 0.883 | 1.023 |
| | *MAE* | 0.628 | 0.472 | 0.441 |

Berdasarkan Tabel 1, model suhu menunjukkan performa yang sangat stabil. Menariknya, RMSE pada prediksi 72 jam kedepan (0.701°C) justru sedikit lebih rendah dibanding prediksi 12 jam kedepan (0.749°C), yang menunjukkan kemampuan model menangkap pola musiman harian (*daily seasonality*) dengan sangat baik meskipun menggunakan strategi rekuren.

### 1.9 KINERJA MODEL KLASIFIKASI

Kinerja model klasifikasi biner (Hujan/Tidak Hujan) dievaluasi menggunakan Akurasi, *F1- Score*, *Precision, Recall*, dan *AUC*.

Tabel 2. Tabel Evaluasi Model Klasifikasi

| Waktu | Akurasi | *F1 Score* | *Precision* | *Recall* | *AUC Score* |
|---|---|---|---|---|---|
| 12 Jam | 91.6% | 0.930 | 0.966 | 0.904 | 0.954 |
| 24 Jam | 85.8% | 0.813 | 0.753 | 0.903 | 0.910 |
| 72 Jam | 87.5% | 0.776 | 0.696 | 0.889 | 0.924 |

E-ISSN : 2623-1700
P-ISSN : 1693-3672

Berdasarkan Tabel 2, model menunjukkan performa terbaik pada prediksi jangka pendek 12 jam dengan *F1-Score* mencapai 0.930 dan *Precision* 0.966. Ini berarti model sangat andal dalam memberikan peringatan hujan 12 jam kedepan dengan tingkat *false* alarm yang sangat rendah. Seiring bertambahnya prediksi waktu ke 72 jam kedepan, terlihat adanya penurunan performa yang wajar dimana *F1-Score* turun menjadi 0.776. Hal ini konsisten dengan karakteristik strategi prediksi rekuren, di mana ketidakpastian sedikit terakumulasi seiring waktu. Namun, nilai *Recall* pada prediksi 72 jam kedepan tetap tinggi (0.889), menunjukkan model masih sangat sensitif mendeteksi potensi hujan meski presisinya menurun.

## KESIMPULAN

Berdasarkan analisis dan hasil penelitian yang telah dilakukan, dapat diambil beberapa kesimpulan sebagai berikut :

1. Algoritma LightGBM terbukti sangat efektif untuk prediksi cuaca jangka pendek. Pada prediksi 12 jam, model klasifikasi hujan mencapai akurasi dan *F1- Score* superior (>0.90), menjadikannya kandidat yang layak untuk implementasi nyata. Klasifikasi hujan mencapai akurasi dan *F1-Score* superior (>0.90), menjadikannya model yang layak untuk implementasi nyata.
2. Model regresi suhu menunjukkan stabilitas tinggi di seluruh waktu dengan *error* rata-rata di bawah 0.8°C, membuktikan bahwa fitur siklik (*cyclical features*) berhasil menangkap pola harian dengan baik.
3. Strategi prediksi rekuren menunjukkan karakteristik yang diharapkan, yaitu akurasi tertinggi di prediksi 12 jam kedepan dan penurunan bertahap di prediksi 72 jam kedepan. Meskipun demikian, performa prediksi 72 jam masih dalam kategori baik (*F1 Score* > 0.75), menunjukkan robustisitas model.

## DAFTAR PUSTAKA

[1] A. M. Siregar, S. Faisal, dan A. Fauzi, “Klasifikasi untuk Prediksi Cuaca Menggunakan Esemble Learning,” vol. 13, no. 2, hal. 138–147, 2020.

[2] J. A. Brotzge *et al.*, “Challenges and Opportunities in Numerical Weather

E-ISSN : 2623-1700
P-ISSN : 1693-3672

Prediction," hal. 698–705, 2023.

[3] S. A. Reni Ria, "Dampak Dan Resiko Perpindahan Ibu Kota Terhadap Ekonomi Di Indonesia," hal. 183–203, 2020.

[4] S. Singh, Nitin, Chaturvedi, Saurabh, Akhter, "Weather Forecasting Using Machine Learning Algorithm," *2019 Int. Conf. Signal Process. Commun.*, hal. 171– 174, 2019.

[5] T.-Y. Ke, Guolin, Qi Meng, Finley Thomas, Wang Taifeng, Chen Wei, Ma Weidong, Ye Qiwei, Liu, "LightGBM : A Highly Efficient Gradient Boosting Decision Tree," no. Nips, hal. 1–9, 2017.