Seminar Nasional Sains Data 2025(SENADA 2025)
UPN "Veteran" Jawa Timur
E-ISSN 2808-5841
P-ISSN 2808-7283

Implementasi Ekosistem Big Data Menggunakan Hadoop untuk Analisis Prediksi Curah Hujan dengan Metode Random Forest di Kota Bandar Lampung

Kemas Veriandra Ramadhan[1], Chalifia Wananda[2], Haikal Dwi Syaputra[3], Dwi Sulistiania[4], Ardika Satria[5], Vidia[6]

[1,2,3,4,5] *Program Studi Sains Data, Fakultas Sains, Institut Teknologi Sumatera*

[1]*kemas.122450016@student.itera.ac.id*

[2] *haikal.122450067@student.itera.ac.id*

[3]*chalifia.122450076@student.itera.ac.id*

[4]*dwi.121450079@student.itera.ac.id*

[5]*ardika.satria@sd.itera.ac.id*

[6]*vidia@sd.itera.ac.id*

** Corresponding E-mail: luluk.muthoharoh@sd.itera.ac.id*

Abstract: Accurate rainfall prediction is crucial for disaster mitigation and water resource planning, particularly in disaster-prone regions such as Bandar Lampung, Indonesia. This study aims to develop a rainfall prediction system using the Random Forest algorithm implemented within the Hadoop Big Data ecosystem. Daily weather data from BMKG (2020–2024) were processed through a Medallion Architecture (Bronze, Silver, Gold) and trained using Spark MLlib classification pipelines. The process includes data cleaning, transformation, rainfall categorization, and model evaluation using precision, recall, and f1-score metrics. Results show that the model performs well in classifying light rain but lacks accuracy in predicting minority classes such as moderate and heavy rain. These findings highlight the need for class balancing techniques and minority data augmentation. The system offers a practical implementation of Hadoop technologies for localized climate forecasting and supports the development of Big Data-based early warning systems.
Keywords: *Rainfall, Random Forest, Hadoop, Big Data, Spark, Climate Prediction*

Abstrak: Prediksi curah hujan yang akurat sangat penting dalam mendukung mitigasi bencana dan perencanaan sumber daya air, khususnya di wilayah rawan bencana seperti Kota Bandar Lampung. Penelitian ini bertujuan untuk mengembangkan sistem prediksi curah hujan berbasis algoritma Random Forest yang dijalankan pada ekosistem Big Data Hadoop. Data cuaca harian dari BMKG tahun 2020–2024 diolah menggunakan arsitektur Medallion (Bronze, Silver, Gold) dan diproses melalui pipeline Spark MLlib untuk pelatihan model klasifikasi. Proses melibatkan pembersihan data, transformasi, kategorisasi curah hujan, serta evaluasi model menggunakan metrik presisi, recall, dan f1-score. Hasil menunjukkan bahwa model sangat baik dalam mengklasifikasikan hujan ringan, namun belum optimal dalam memprediksi kelas minor seperti hujan sedang dan lebat. Temuan ini menunjukkan perlunya penanganan ketidakseimbangan kelas dan pengayaan data minor. Sistem ini memberikan kontribusi nyata dalam integrasi teknologi Hadoop untuk prediksi iklim lokal dan mendukung pengembangan sistem peringatan dini berbasis Big Data.
Kata kunci: *Curah hujan, Random Forest, Hadoop, Big Data, Spark, Prediksi Iklim*

I. PENDAHULUAN

Curah hujan merupakan salah satu parameter klimatologis yang paling krusial dalam sistem peringatan dini bencana di Indonesia. Kota Bandar Lampung, sebagai bagian dari wilayah rawan bencana di Pulau Sumatera, kerap mengalami dampak dari cuaca ekstrem seperti banjir, tanah longsor, dan kekeringan. Prediksi curah hujan yang akurat sangat dibutuhkan oleh instansi pemerintah, sektor pertanian, dan masyarakat luas untuk mitigasi risiko bencana serta perencanaan sumber daya air[1]. Namun, tantangan muncul karena data cuaca yang besar, heterogen, dan terus berkembang sulit ditangani oleh sistem konvensional.

BMKG secara rutin menghasilkan data iklim harian yang kaya informasi, termasuk suhu, kelembaban, curah hujan, dan parameter angin. Di sisi lain, kemajuan dalam teknologi Big Data dan machine learning telah membuka peluang baru untuk menganalisis data tersebut secara lebih efisien dan presisi [5]. Berbagai studi sebelumnya menunjukkan bahwa algoritma klasifikasi seperti Random Forest efektif digunakan dalam prediksi curah hujan dan kejadian cuaca ekstrem [6]. Namun, masih terbatasnya infrastruktur analitik yang mendukung pemrosesan data skala besar menjadi penghambat utama dalam penerapan sistem prediksi yang handal dan berkelanjutan.

Belum banyak studi di Indonesia yang mengimplementasikan sistem prediksi curah hujan berbasis ekosistem Hadoop secara end-to-end. Infrastruktur Hadoop memiliki kemampuan penyimpanan dan pemrosesan data besar secara terdistribusi, namun penggunaannya masih minim dalam konteks lokal untuk prediksi iklim. Selain itu, integrasi antara data historis cuaca dan algoritma machine learning pada arsitektur Lakehouse berbasis Medallion masih jarang dijumpai. Oleh karena itu, diperlukan upaya perancangan sistem prediktif yang scalable dan efisien menggunakan Hadoop, Apache Spark, Hive, dan Random Forest, khususnya untuk wilayah Bandar Lampung.

Penelitian ini bertujuan untuk mengembangkan sistem prediksi curah hujan menggunakan algoritma Random Forest yang dijalankan pada ekosistem Hadoop. Data curah hujan harian dari BMKG Bandar Lampung diolah melalui pipeline Medallion Architecture (Bronze–Silver–Gold), diproses menggunakan Apache Spark untuk transformasi dan pelatihan model, serta dianalisis dengan Hive SQL untuk visualisasi hasil. Sistem dibangun dalam lingkungan Docker multi-container dan mengimplementasikan skenario pemrosesan batch yang efisien serta mendukung skalabilitas data.

Sistem ini diharapkan dapat memberikan prediksi curah hujan yang lebih akurat dan real-time, mendukung kebijakan mitigasi bencana, serta menjadi fondasi untuk pengembangan sistem peringatan dini iklim berbasis Big Data di masa depan. Selain itu, proyek ini memberikan contoh nyata penerapan teknologi Hadoop dalam konteks lokal dan menyediakan arsitektur sistem yang dapat direplikasi untuk wilayah lain di Indonesia.

II. METODE PENELITIAN

II.1 Pendekatan Sistem dan Sumber Data

Untuk memprediksi kategori curah hujan di Kota Bandar Lampung penelitian ini menggunakan pendekatan *big data classification*. Algoritma yang digunakan adalah Random Forest Classifier yang diimplementasikan dalam ekosistem Apache Hadoop, khususnya dengan pemanfaatan Apache Spark MLlib untuk pelatihan model klasifikasi. Sistem dirancang berdasarkan arsitektur data lake Medallion yang terdiri dari tiga lapisan yaitu Bronze Layer, Silver Layer, dan Gold Layer. Arsitektur sistem berjalan di atas cluster Hadoop berbasis Docker, dengan komponen utama seperti Hadoop HDFS, Apache Spark, Apache Hive, Apache Ambari, dan Apache Superset.

Data yang digunakan adalah data historis curah hujan harian di Kota Bandar Lampung yang diperoleh dari Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) Stasiun Pajang Kota Bandar Lampung pada rentang tahun 2020-2024, dimana data dengan 11 variabel yaitu tanggal pengamatan, suhu minimum harian, suhu maksimum harian, suhu rata-rata harian, kelembaban relatif rata-rata harian, curah hujan harian, lama penyinaran matahari, kecepatan angin maksimum harian, arah angin

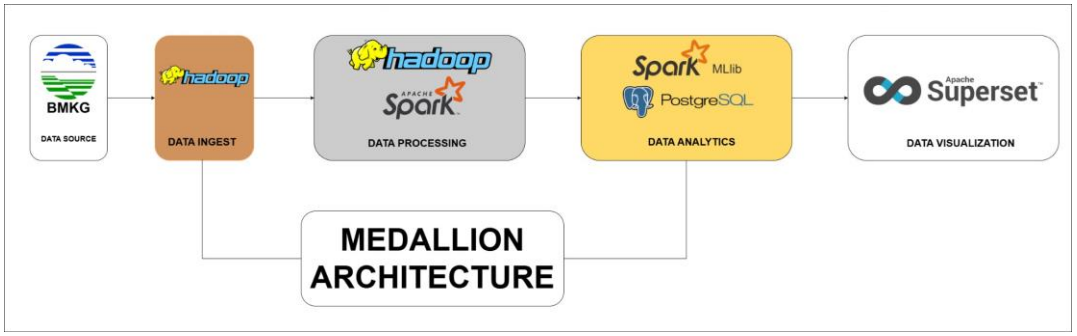
maksimum, kecepatan angin rata-rata harian, dan arah angin utama. Akan tetapi yang akan digunakan untuk prediksi cuaca hanya 5 variabel yaitu tanggal pengamatan, suhu rata-rata, kelembapan relatif rata-rata, curah hujan, dan lama penyinaran matahari.

Tabel 1. Dataset Tahun 2020-2024

No	date	temp_avg	humidity_avg	rainfall	sunshine
1	1/1/2020	27.9	83	42.4	3.4
2	1/2/2020	28.4	80	-	1.2
3	1/3/2020	27.9	84	8	5
4	1/4/2020	28.1	84	33	5
…	…	…	…	…	…
1708	12/31/2024	29.5	79	0.3	0.1

II.2 Alur dan Tahapan Pemrosesan Data

1. Data Mentah (Bronze Layer) : Data curah hujan mentah diambil secara berkala pada sumber data dan disimpan ke HDFS dalam format CSV.
2. Preprocessing (Silver Layer) : Proses pembersihan dan transformasi data, yang dilakukan menggunakan Apache Spark, mencakup penghapusan duplikasi dan nilai null, *parsing* tanggal dan waktu, *encoding* fitur kategorikal, serta konversi ke format kolumnar efisien seperti Parquet.
3. Data Klasifikasi (Gold Layer) : Data dianalisis dan dilabeli untuk klasifikasi. Fitur-fitur tambahan juga dapat ditambahkan, seperti kategori waktu (pagi/siang/malam) atau rata-rata kelembaban harian.



Gambar 1. Pipeline Arsitektur

II.3 Pelatihan Model Klasifikasi

1. Preprocessing

   Pada langkah ini dimulai dari pembersihan dan transformasi pada data yang bagiannya yakni pemilihan fitur dimana fitur tanggal tidak akan dipakai dalam pelatihan model ini karena data bertipe *datetime* dan dalam pelatihan model tidak menggunakan algoritma berbasis waktu. Selanjutnya lakukan pengecekan data yang hilang *(null)* pada tiap fitur, dimana pada fitur

'curah hujan' dilakukan pengisian data yang hilang dengan tanda (-) dan angka 8888 atau 9999 (data tidak diukur) dengan cara imputasi berdasarkan mean. Untuk fitur 'arah angin terbanyak' melakukan transformasi data pada fitur arah angin terbanyak (ddd_car), yang dimana fitur arah angin terbanyak memiliki type data huruf (string) yang tidak seragam dengan fitur lain yang memiliki type data angka (numerik). Setelah dilakukan pembersihan dan transformasi lakukan normalisasi fitur numerik menggunakan Min-Max Scaling. Min-Max normalization merupakan metode normalisasi dengan melakukan transformasi linier terhadap data asli sehingga menghasilkan keseimbangan nilai perbandingan antar data saat sebelum dan sesudah proses. Lalu lakukan encoding label pada curah hujan untuk klasifikasi.

2. Splitting Data
   Dataset dibagi menjadi 80% data latih dan 20% data uji dengan metode *randomSplit()* dari Spark. Untuk data latih menggunakan dataset dari tahun 2020-2023, sedangkan data uji menggunakan dataset tahun 2024.
3. Pelatihan Model
   Algoritma yang digunakan adalah *RandomForestClassifier* dari Spark MLlib. Model tersebut akan dilatih pada data latih dan juga data uji.
4. Evaluasi Model
   Evaluasi model yang digunakan pada penelitian ini adalah metric evaluasinya accuracy, recall, precision. Confusion matriks digunakan untuk menilai seberapa baik model mengklasifikasikan masing-masing kelas.

II.4 Visualisasi dan Output

Hasil klasifikasi dan evaluasi model ditampilkan melalui beberapa komponen utama dalam ekosistem big data yaitu ;

1. Apache Hive digunakan untuk melakukan query dan inspeksi terhadap data hasil model yang telah disimpan di dalam tabel pada Gold Layer. Hal ini memungkinkan pengguna teknis seperti data analyst untuk mengeksplorasi hasil klasifikasi secara detail menggunakan bahasa SQL.
2. Apache Superset dimanfaatkan untuk menyajikan visualisasi model dalam bentuk grafik yang interaktif, termasuk distribusi kelas curah hujan (rendah, sedang, tinggi) dalam rentang waktu tertentu.

Salah satu visualisasi utama yang dihasilkan adalah grafik time-series yang menggambarkan tren perubahan kategori curah hujan berdasarkan waktu, sehingga memudahkan dalam memantau pola cuaca dan mendukung pengambilan keputusan berbasis data.

II.5 Pengujian Sistem

Beberapa jenis pengujian yang dilakukan dalam proyek ini meliputi:

- Uni Testing
  Pengujian ini dilakukan untuk memvalidasi setiap komponen atau fungsi secara terpisah, khususnya script yang digunakan untuk proses *ingestion* dan *preprocessing* data. Tujuannya adalah memastikan bahwa setiap bagian dari sistem dapat berjalan secara mandiri tanpa error sebelum diintegrasikan ke dalam pipeline utama.
- Integration Testing
  Pengujian ini bertujuan untuk memastikan bahwa alur proses dari tahap *ingestion* (pengambilan data), *transformasi* (pembersihan dan pengolahan data), *klasifikasi* (pemodelan), hingga *visualisasi* (tampilan hasil prediksi) dapat terintegrasi dengan baik dan berjalan lancar sebagai satu kesatuan sistem.
- Model Evaluation Testing
  Pengujian ini difokuskan pada evaluasi kinerja model klasifikasi yang digunakan. Penilaian dilakukan dengan menggunakan metrik evaluasi seperti akurasi, presisi, dan recall, untuk

mengetahui sejauh mana model mampu memprediksi kelas curah hujan secara tepat dan efektif.

- End-to-End Testing
  Pengujian menyeluruh dari awal hingga akhir sistem, dimulai dari pengambilan data mentah hingga tahap akhir berupa visualisasi hasil prediksi di dashboard. Pengujian ini memastikan bahwa seluruh pipeline berjalan sesuai yang diharapkan tanpa adanya kendala atau kesalahan pada setiap tahap.

III. HASIL PEMBAHASAN

III.1 Preprocessing Data

- Informasi Struktur, Perubahan Tipe dan Nama Fitur Data

Berdasarkan output informasi data, DataFrame terdiri dari 1788 baris dan 10 kolom dengan struktur yang mencakup berbagai variabel seperti suhu, kelembapan, curah hujan, jam matahari, kecepatan dan arah angin. Semua kolom memiliki lengkap 1788 entri non-null data selama 5 tahun, menunjukkan tidak adanya data yang hilang. Secara tipe data, sebagian besar kolom masih berbentuk object, termasuk pada kolom numerik seperti temp_min, temp_max, dan temp_avg, sementara hanya wind_speed_max dan wind_speed_avg yang sudah bertipe `float64`. Untuk memudahkan analisis kuantitatif, perlu dilakukan konversi tipe data dari `object` ke float64 atau int64.

Sebagai bagian dari proses preprocessing dan memudahkan pemahaman serta pemilihan fitur dalam analisis data, dilakukan perubahan nama kolom menggunakan `column_mapping`. Setiap nama kolom asli yang kurang intuitif diganti dengan nama yang lebih deskriptif dan mudah dimengerti sesuai dengan makna variabelnya, seperti mengubah TN menjadi temp_min, FF_X menjadi wind_speed_max, dan sebagainya. Dengan pemberian nama yang lebih jelas, pengguna dapat lebih mudah mengidentifikasi jenis data yang tersedia dan memilih fitur yang relevan untuk analisis selanjutnya.

Dalam proses dilakukannya perubahan tipe data pada sejumlah kolom yang awalnya bersifat string menjadi tipe float agar dapat digunakan dalam analisis kuantitatif. Kolom-kolom tersebut meliputi temp_min, temp_max, temp_avg, humidity_avg, rainfall, sun_hours, wind_speed_max, dan wind_speed_avg. Perubahan dilakukan menggunakan fungsi pd.to_numeric() dengan parameter errors=coerce, yang bertujuan untuk mengubah nilai yang tidak dapat dikonversi menjadi NaN sehingga menghindari error. Dengan demikian, seluruh kolom numerik siap digunakan untuk komputasi matematis dan analisis statistik secara lebih akurat.

- Transformasi Data

Dalam proses transformasi data, nilai (-) dan 8888 yang menandakan pengukuran tidak dilakukan atau data tidak tersedia, diubah menjadi NaN untuk menunjukkan adanya nilai yang hilang. Setelah itu, dilakukan imputasi pada kolom numerik menggunakan rata-rata (mean) sebagai strategi penggantian nilai kosong. Imputasi ini dilakukan hanya pada kolom numerik seperti temp_min, temp_max, humidity_avg, dan sebagainya agar data tetap dapat digunakan dalam analisis kuantitatif tanpa kehilangan informasi penting akibat baris dengan data hilang. Dengan pendekatan ini, dataset menjadi lebih bersih dan siap untuk digunakan dalam pemodelan atau analisis selanjutnya.

Sehingga dilakukan transformasi dengan bantuan library scikit-learn (sklearn) yaitu fungsi LabelEncoder, yang mengubah data huruf ke type data angka. LabelEncoder mengubah setiap nilai dalam kolom fitur menjadi angka yang berurutan. Dimana data pada fitur arah angin terbanyak (ddd_car) yang berisi data antara lain NW (northwest), W (west), N (north), NE

(northwest), SE (southeast), E (east), S (south), SW (southwest) dan C (calm / tidak ada angin / lemah) lalu data yang berupa huruf (string) tersebut diubah ke data angka (numerik) secara berurutan menjadi NW : 0, W : 1, N : 2, NE : 3, SE : 4, E : 5, S 6, SW 7 dan C : 8.

- Pembentukan dan Pemilihan Variabel Target

Karena memiliki tujuan klasterisasi hujan, maka yang harus dilakukan yakni membentuk variabel baru dari berbagai jenis kategori pada data fitur variable target minimal minimal dua bagian (Boolean : 0 dan 1). Disini terpilihlah 4 jenis kategori pada data berdasarkan pada variabel hujan (rainfall) yang diambil dari parameter yang ada pada situs BMKG itu sendiri berdasarkan range curah hujan nya pada tabel berikut ini:

Tabel 2. Kategori Hujan

Rentang	Label	Kategori
0.5-20	0	Ringan
20-50	1	Sedang
50-100	2	Lebat
>100	3	Sangat Lebat

Didapat untuk kategori hujan ringan berada pada kategori 0 sebanyak 1627 baris data, hujan sedang dengan kategori 1 sebanyak 119 baris data, hujan lebat dengan kategori 2 sebanyak 35 baris data dan hujan sangat lebat dengan kategori 3 sebanyak 7 baris data.

III.2 Analisis Data dengan Algoritma Random Forest

- Membagi Data, Pembentukan Model, dan Evaluasi

Langkah awal dalam pembuatan model klasifikasi adalah membagi dataset menjadi dua bagian utama, yaitu set pelatihan dan set pengujian. Dalam kasus ini, kolom rainfall dan wind_dir_most dihapus serta rainfall_category dipisahkan sebagai target (y), sementara semua kolom lainnya dijadikan sebagai fitur (X) untuk memprediksi kategori hujan. Pembagian data dilakukan dengan fungsi train_test_split, di mana 80% data digunakan untuk pelatihan dan 20% sisanya untuk pengujian. Pemisahan ini memastikan bahwa model dapat dievaluasi secara objektif menggunakan data yang belum pernah dilihat sebelumnya.

Setelah data dibagi, model klasifikasi dibangun menggunakan algoritma Random Forest Classifier, yang merupakan metode ensemble berbasis pendekatan pohon keputusan. Model ini memiliki kemampuan tinggi dalam menangani data dengan dimensi yang kompleks serta cenderung lebih stabil terhadap overfitting dibandingkan model lain seperti Decision Tree. Proses pelatihan dilakukan dengan memasukkan training set ke dalam model agar dapat belajar pola dari data dan menghasilkan prediksi kategori curah hujan. Setelah model dilatih, prediksi dilakukan pada testing set untuk mengevaluasi performa akhir model.

Untuk menilai kualitas model, evaluasi dilakukan menggunakan classification_report, yang menampilkan metrik utama seperti presisi, recall, F1-score, dan akurasi untuk setiap kelas dalam variabel target. Laporan ini memberikan gambaran menyeluruh tentang seberapa baik model dapat mengklasifikasikan setiap kategori hujan, seperti "Hujan Ringan", "Hujan Sedang", atau "Hujan Lebat". Selain itu, evaluasi ini juga membantu mengidentifikasi apakah model

mengalami kesulitan dalam memprediksi kelas tertentu, sehingga bisa menjadi dasar untuk peningkatan model di masa mendatang.

	precision	recall	f1-score	support
Ringan	0.91	0.99	0.95	326
Sedang	0.00	0.00	0.00	23
Lebat	0.00	0.00	0.00	8
Sangat Lebat	0.00	0.00	0.00	1
accuracy			0.90	358
macro avg	0.23	0.25	0.24	358
weighted avg	0.83	0.90	0.86	358

Gambar 2. Hasil Evaluasi

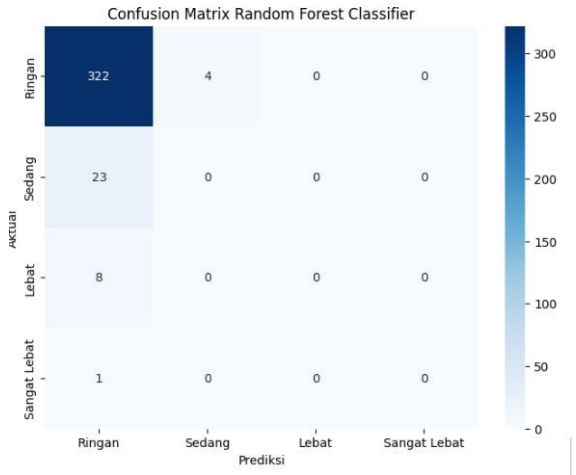
- Penjelasan Hasil dan Evaluasi

Hasil evaluasi menggunakan *classification_report* menunjukkan performa model dalam memprediksi kategori curah hujan, yaitu "Ringan", "Sedang", "Lebat", dan "Sangat Lebat". Untuk kelas "Ringan", model mencapai hasil yang sangat baik dengan nilai precision (presisi) sebesar 0.91, recall (sensitivitas) sebesar 0.99, dan f1-score sebesar 0.95. Hal ini menunjukkan bahwa model mampu mengidentifikasi kategori "Ringan" dengan sangat akurat, baik dari segi kemampuan meminimalkan false positive maupun menemukan sebagian besar kasus positif yang ada. Namun, untuk kelas lainnya seperti "Sedang", "Lebat", dan "Sangat Lebat", model sepenuhnya gagal melakukan prediksi, dengan semua metrik (precision, recall, dan f1-score) bernilai 0.00. Ini menunjukkan bahwa model tidak dapat membedakan antara kategori-kategori tersebut atau bahkan tidak menghasilkan prediksi sama sekali untuk kelas-kelas minor ini.

Secara keseluruhan, model memiliki akurasi global sebesar 0.90, yang terlihat cukup tinggi pada pandangan pertama. Namun, ketika dilihat lebih mendalam melalui macro average, yang memberikan bobot yang sama kepada setiap kelas, nilai presisi, recall, dan f1-score justru rendah, yaitu 0.23, 0.25, dan 0.24, berturut-turut. Hal ini menunjukkan bahwa meskipun model bekerja dengan baik untuk kelas mayor ("Ringan"), performanya sangat buruk untuk kelas minor ("Sedang", "Lebat", dan "Sangat Lebat"). Selain itu, weighted average, yang mempertimbangkan distribusi kelas dalam data, masih menunjukkan hasil yang relatif baik karena dominasi kelas "Ringan". Untuk meningkatkan performa model secara keseluruhan, beberapa langkah perlu diambil, seperti menambah jumlah sampel untuk kelas minor melalui teknik oversampling atau generasi data sintetis, menggunakan algoritma yang lebih sensitif terhadap keseimbangan kelas seperti Balanced Random Forest atau SMOTE memodifikasi parameter model, seperti meningkatkan kedalaman pohon atau jumlah estimator serta melakukan penyesuaian threshold untuk memperbaiki keseimbangan antara precision dan recall. Dengan demikian, model dapat menjadi lebih efektif dalam memprediksi semua kategori curah hujan secara merata.

Berdasarkan Gambar 3 yang menampilkan confusion matrix dari model Random Forest, terlihat bahwa model sangat baik dalam mengklasifikasikan curah hujan dengan kategori "Ringan", namun gagal dalam mengenali kategori lainnya seperti "Sedang", "Lebat", dan "Sangat Lebat". Dari total 322 data aktual dengan label "Ringan", model berhasil mengklasifikasikan semuanya dengan benar, sementara beberapa data dari kategori lain seperti "Sedang" (23 data), "Lebat" (8 data), dan "Sangat Lebat" (1 data) juga salah diklasifikasikan sebagai "Ringan". Hal ini menunjukkan adanya ketidakseimbangan data (*class imbalance*) di mana kelas "Ringan" mendominasi dataset, sehingga model cenderung memprediksi semua data ke kelas tersebut. Metrik evaluasi seperti recall dan precision pada kelas selain "Ringan"

memiliki nilai nol, yang berarti model tidak mampu mengenali adanya hujan selain kategori "Ringan". Akurasi model secara umum tampak tinggi, namun metrik ini menyesatkan karena tidak mencerminkan kegagalan model dalam mendeteksi kelas minoritas. Visualisasi confusion matrix juga terlihat kurang efektif karena perbedaan warna yang terlalu halus membuat informasi penting sulit terbaca, terutama untuk kelas dengan jumlah data yang sedikit. Oleh karena itu, disarankan agar visualisasi diperbaiki dengan menambahkan angka presentasi dan kontras warna yang lebih baik, serta dilakukan perbaikan pada model seperti penyeimbangan data (misalnya melalui teknik oversampling) atau penggunaan class weighting agar performa model menjadi lebih seimbang dan adil dalam mengklasifikasikan seluruh kategori curah hujan.



Gambar 3. Hasil Evaluasi

IV. KESIMPULAN

Berdasarkan hasil penelitian ekosistem Big Data Hadoop untuk analisis cuaca ekstrem di sumatera, penggunaan hadoop membantu mengelola penyimpanan data iklim yang berasal dari BMKG menjadi terorganisir dan mudah disimpan. Penggunaaan *medallion architecture* turut membantu kemudahan pengelolaan tahapan proses data. *Stack* seperti Spark dan Hive membantu kemudahan proses data yang lebih besar. Selain itu klasifikasi yang didapatkan oleh machine learning menggunakan mode RandomForest bisa dikatakan cukup baik mengenali pola data. Sehingga dapat mengklasifikasikan jenis cuaca dengan baik dan benar.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada Bapak Ardika Satria, M.Si., Ibu Luluk Muthoharoh, M.Si., dan Ibu Vidia, M.Si. yang telah membimbing serta memberikan pengetahuan dan arahan selama perkuliahan mata kuliah Analisis Big Data, yang menjadi dasar penting dalam penyusunan paper ini.

Ucapan terima kasih juga penulis sampaikan kepada teman-teman yang telah memberikan semangat dan dukungan selama proses penulisan. Secara khusus, penulis mengapresiasi tim SENADA yang telah meluangkan waktu untuk menyelenggarakan kegiatan ini dan memberikan kesempatan serta masukan berharga dalam proses review agar paper ini menjadi lebih baik.

REFERENSI

1. Badan Meteorologi, Klimatologi, dan Geofisika. (2023). *Data Cuaca Harian dan Klimatologi BMKG 2020–2024*. BMKG.

2. Zikopoulos, P., & Eaton, C. (2011). *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. McGraw-Hill.

3. Hashem, I. A. T., et al. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, 47, 98–115.

4. Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137–144.

5. Zhou, Z.-H., Feng, J., & He, Z. (2019). Machine learning for weather and climate modelling. *Nature Reviews Physics*, 1(6), 341–356.

LAMPIRAN

Link github : https://github.com/sains-data/ANALISIS-CUACA-EKSTREM-DI-SUMATERA