JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 9, No. 2, Juni 2024, Pp. 628-638

PERBANDINGAN AKURASI ALGORITMA C4.5 DAN *K-NEAREST NEIGHBORS* UNTUK KLASIFIKASI CURAH HUJAN BERDASARKAN IKLIM INDONESIA

Muhammad Fauzan Nasrullah*[1)], Rd. Rohmat Saedudin[2)], Faqih Hamami[3)]
1. Telkom University, Indonesia
2. Telkom University, Indonesia
3. Telkom University, Indonesia

Article Info
Kata Kunci: Curah hujan; Iklim Indonesia; Klasifikasi; Algoritma C4.5; *K-Nearest Neighbor; Data Mining*

Keywords: *Rainfall; Indonesian Climate; Classification; C4.5 Algorithm; K-Nearest Neighbor*

Article history:
Received 18 February 2024
Revised 3 March 2024
Accepted 17 March 2024
Available online 1 June 2024

DOI :
https://doi.org/10.29100/jipi.v9i2.4655

* Corresponding author.
Muhammad Fauzan Nasrullah
E-mail address:
fauzannash@student.telkomuniversity.ac.id

ABSTRAK
Indonesia memiliki iklim tropis yang dominan, karenanya Indonesia mengalami variasi suhu yang terbatas, namun memiliki variasi curah hujan yang beragam. Variabilitas curah hujan juga tidak lepas dengan dampak yang diberikannya pada berbagai aspek kehidupan manusia dan aktivitas bisnis. Oleh karena itu informasi curah hujan merupakan aspek penting dalam pengambilan keputusan. Namun, tentunya perlu tahapan dan metode untuk melakukan proses Analisa. Maka dari itu penelitian ini bertujuan untuk mencari metode terbaik antara C4.5 dan K-Nearest Neighbors yang termasuk algoritma pada *data mining* untuk mengklasifikasi data curah hujan. Kedua algoritma digunakan untuk membangun model klasifikasi berdasarkan atribut-atribut yang relevan. Kemudian, model-model tersebut diuji dan dievaluasi menggunakan berbagai metrik seperti Akurasi, Precision, Recall dan F1-Score. Dalam penelitian ini juga menerapkan *Tuning Hyperparameter* dengan metode RandomizeSearchCV untuk mendapatkan parameter terbaik yang dapat menghasilkan nilai akurasi yang maksimal. Hasil penelitian menunjukkan bahwa kedua algoritma memiliki kinerja yang baik dalam klasifikasi curah hujan. Jika berdasarkan nilai akurasi yang didapat dengan parameter default dari kedua algoritma, C4.5 memiliki nilai akurasi yang lebih tinggi sebesar 81.42%, sementara *K-Nearest Neighbors* hanya sebesar 78.10%. Namun setelah menggunakan parameter terbaik hasil dari penerapan Tuning Hyperparameter RandomizedSearchCV, perubahan nilai akurasi yang cukup signifikan terjadi pada K-Nearest Neighbors yang didapati sebesar 83.37%, sementara C4.5 bertambah menjadi 82.56%

ABSTRACT
Indonesia has a dominant tropical climate, which is why it experiences limited temperature variations but diverse rainfall patterns. The variability of rainfall is closely intertwined with the impacts it exerts on various aspects of human life and business activities. Therefore, rainfall information constitutes a crucial aspect in decision-making. However, of course, there is a need for stages and methods to conduct the analysis process. Hence, this study aims to determine the superior method between C4.5 and K-Nearest Neighbors, both of which are algorithms in data mining, for classifying rainfall data. Both algorithms are employed to construct classification models based on relevant attributes. Subsequently, these models are tested and evaluated using various metrics such as Accuracy, Precision, Recall, and F1-Score. In this study, Hyperparameter Tuning is also applied using the RandomizedSearchCV method to obtain optimal parameters that can yield maximum accuracy. The research findings indicate that both algorithms perform well in rainfall classification. When considering the accuracy values obtained with the default parameters of both algorithms, C4.5 exhibits a higher accuracy rate of 81.42%, whereas K-Nearest Neighbors only achieves 78.10%. However, after utilizing the best parameters resulting from the implementation of Hyperparameter Tuning with RandomizedSearchCV, a significant accuracy improvement is observed in K-Nearest Neighbors, which reaches 83.37%. Meanwhile, C4.5's accuracy increases to 82.56%.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 9, No. 2, Juni 2024, Pp. 628-638

I. PENDAHULUAN

SEBUAH bentuk ketidakseimbangan ekosistem yang terjadi di bumi, atau yang lebih kita kenal dengan pemanasan global (*global warming*) terjadi akibat proses peningkatan suhu rata-rata atmosfer, laut, dan daratan di bumi. Selain berdampak langsung terhadap kenaikan suhu permukaan, pemanasan global juga menyebabkan terjadinya perubahan iklim, yang berdampak pada kehidupan manusia. Iklim dapat diartikan sebagai suatu kondisi cuaca pada suatu daerah dalam kurun waktu yang lebih lama [1].

Informasi iklim/cuaca merupakan sebuah informasi penting yang tak dapat dipisahkan dari kegiatan manusia terlebih pada sektor pembangunan seperti sektor pertanian, perkebunan, kehutanan, transportasi, pengairan, lingkungan hidup, pertambangan dan energi mitigasi bencana dan lain-lain [2]. Seperti permasalahan yang diangkat dalam penelitian ini yaitu curah hujan yang memiliki berbagai dampak signifikan terhadap variabilitas iklim. Menurut Sipayung [3], karakteristik iklim suatu daerah juga dapat dilihat dari variasi cuaca hujannya. Variabilitas curah hujan juga tidak lepas dengan dampak yang diberikan seperti menyebabkan beberapa bencana alam seperti banjir dan kekeringan. Hal ini membuat masyarakat sekitar merasakan dampak yang begitu signifikan dari variabilitas curah hujan.

Pada bulan Juli 2022 curah hujan rendah di Kabupaten Lombok Barat mengakibatkan beberapa wilayah terdampak kekeringan. Selanjutnya berdasarkan hasil Rekap Data Kejadian Bencana Pusat Pengendalian Operasi Penanggulangan Bencana (Pusdalops PB) terdapat 10 kejadian banjir dikarenakan curah hujan tinggi di Kabupaten Probolinggo dihitung sejak bulan Maret 2020 lalu (BPBD, 2022). Hal seperti itu tentunya dapat menghambat aktivitas manusia di berbagai sektor, salah satu sektor yang sangat dipengaruhi variabilitas curah hujan yaitu sektor pertanian, karena hal itu menyebabkan dinamika pergeseran musim hujan dan kemarau yang membuat peningkatan terjadinya resiko gagal panen [4].

Berdasarkan peristiwa-peristiwa yang sudah terjadi, informasi iklim/cuaca kini menjadi suatu aspek penting dalam sistem informasi [5]. Hal tersebut disebabkan karena berdampak pada berbagai aspek kehidupan manusia dan aktivitas bisnis, seperti pada perusahaan pertanian yang mendapatkaan manfaat dalam memprediksi curah hujan yang akurat untuk mengatur jadwal tanam, pemeliharaan tanaman, dan kegiatan pertanian lainnya secara efisien [6]. Selain itu perusahaan di bidang energi terbarukan juga mendapatkan manfaat dari informasi curah hujan dalam mengoptimalkan produksi energi dan memprediksi potensi ketersediaan sumber energi [7]. Dan masih banyak perusahaan lainnya pada macam-macam bidang yang dapat terbantu dengan informasi iklim/cuaca untuk pengambilan keputusan yang lebih tepat dan efisien [8]. Oleh karena itu informasi iklim/cuaca mempunyai nilai yang sangat strategis dalam pengambilan keputusan [9]. Namun, tentunya perlu berbagai tahap dan metode untuk melakukan proses analisa terhadap iklim suatu wilayah, salah satunya ialah proses *data mining*.

Data mining merupakan serangkaian proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar [10]. Data mining juga dapat diartikan sebagai penggalian informasi baru dari sejumlah besar data untuk membantu dalam pengambilan keputusan. Di dalamnya pun terdapat banyak teknik seperti Algoritma C4.5, *K-Nearest Neighbors* dan sebagainya. Istilah *data mining* kini mulai populer seiring perkembangan zaman yang serba digital saat ini. Namun, terdapat beberapa tantangan dalam implementasi *data mining* di masa kini. Seperti struktur data yang cenderung heterogen yang berasal dari berbagai sumber, karenanya diperlukan pemrosesan awal dan penggabungan data terlebih dahulu [11]. Lalu perkembangan metode juga menyebabkan banyak lahirnya algoritma yang rumit dan memerlukan banyak parameter untuk mencapai hasil yang optimal. Terlepas dari hal itu, tentu data mining akan semakin dibutuhkan pada zaman ini dan zaman yang akan dating [12], maka dari itu peneliti ingin membuktikan bahwa implementasi data mining dapat memberikan solusi yang baik untuk klasifikasi curah hujan berdasarkan iklim di Indonesia. Namun, banyaknya jenis algoritma dalam data mining tentunya memiliki nilai positif dan negatifnya masing-masing bagi setiap algoritma[13]. Oleh karena itu dalam penelitian ini bertujuan untuk mencari algoritma dengan hasil akurasi terbaik dalam mengklasifikasi curah hujan dari dua algoritma klasifikasi, yaitu *K-Nearest Neighbors* dan C4.5.

Penelitian dengan topik serupa juga sudah dilakukan seperti pada penelitian [14] yang berjudul Perbandingan Metode Data Mining untuk Prediksi Curah Hujan dengan Algoritma C4.5, Naïve Bayes, dan KNN, mendappatkan hasil dari ketiga algoritma yang diuji, didapatkan bahwa algoritma C4.5 mendapatkan hasil akurasi terbaik sebesar 88.03%. Selanjutnya pada penelitian [15] yang berjudul Analisis Teknik Data Mining Algoritma C4.5 dan K-Nearest Neighbors untuk Mendiagnosa Penyakit Diabetes Melitus, juga membandingkan hasil akurasi dari dua algoritma klasifikasi tersebut sehingga mendapatkan hasil bahwa algoritma KNN memiliki hasil akurasi yang lebih tinggi sebesar 79.14% daripada algoritma C4.5 dengan hasil akurasi sebesar 76.10%. Dari beberapa penelitian

terdahulu yang sudah dilakukan dapat dikatakan kedua algoritma sudah cukup baik dalam mengklasifikasi sebuah data, namun dengan perbedaan dataset, teknik dan *tools* yang dilakukan mungkin dapat membuat salah satu dari kedua algoritma tersebut terkadang lebih unggul dan tidak unggul.

Oleh karenanya dalam penelitian ini menggunakan *K-Fold Cross Validation* dalam memvalidasi dari kedua model untuk melihat konsistensi dari kedua model tersebut. Selain itu penulis juga menggunakan *Tuning Hyperparameter* dengan pendekatan *RandomizedsearchCV* untuk meningkatkan kinerja kedua model agar menemukan parameter yang paling tepat sesuai dengan dataset yang digunakan. Sehingga pada akhirnya penulis dapat membandingkan kinerja algoritma C4.5 dan *K-Nearest Neighbors* dalam mengklasifikasi curah hujan berdasarkan iklim Indonesia.

II. METODE PENELITIAN

A. Curah Hujan

Curah hujan dapat diartikan sebagai banyak atau ketinggian air hujan yang terkumpul pada tempat yang datar, tidak menguap, tidak meresap, dan tidak mengalir dalam jangka waktu tertentu. Satuan yang selalu digunakan dalam mengukur curah hujan ialah satuan milimeter atau inchi, namun untuk di Indonesia sendiri satuan curah hujan yang digunakan ialah dalam satuan milimeter (mm). Pada umumnya jika curah hujan sebanyak satu milimeter, maka pada suatu tempat yang datar seluas satu meter persegi, air akan tertampung dengan ketinggian satu milimeter atau setara dengan satu liter [16]. Menurut BMKG terdapat lima kriteria curah hujan sebagai berikut :

TABEL I
KATEGORI CURAH HUJAN

Status	Rentang Curah Hujan
Berawan	0 mm/hari
Sangat Ringan	< 5 mm/hari
Ringan	5 – 20 mm/hari
Sedang	21-50 mm/hari
Lebat	51-100 mm/hari

B. *Data Mining*

Data Mining atau penggalian data dalam bahasa Indonesia dapat diartikan sebagai suatu proses dimana kecerdasan buatan, matematika, teknik statistik dan *machine learning* digunakan untuk mengekstrasi dan mengidentifikasi sebuah informasi yang memiliki sebuah *value* dan pengetahuan yang terkait dari berbagai *database* besar [17]. *Data mining* juga sering disebut sebagai *knowledge discovery in database* (*KDD*), merupakan kegiatan yang meliputi pengumpulan, pemakaian data, historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar [18]. Sederhananya *data mining* dapat didefinisikan sebagai proses untuk mendapatkan informasi yang memiliki *value* dari gudang basis data yang besar (Tan et al., 2006).

C. Algoritma C4.5

Algoritma C4.5 adalah algoritma yang digunakan untuk membentuk pohon keputusan (*Decision Tree*). Pohon keputusan sendiri merupakan sebuah metode klasifikasi dan prediksi yang kerap dikenal dan digunakan oleh banyak orang. Pohon keputusan berguna untuk mengeksplor data dan menemukan hubungan yang tersembunyi dari variabel atau atribut yang digunakan. Algoritma C4.5 juga merupakan pengembangan dari algoritma ID3 yang dikembangkan oleh J.Ross Quinlan, algoritma ini juga merupakan algoritma untuk membangun sebuah pohon keputusan [20]

D. Algoritma *K-Nearest Neighbors*

K-Nearest Neighbors atau yang disingkat KNN merupakan salah satu algoritma yang bertujuan untuk mengklasifikasikan objek baru berdasarkan data pembelajaran (*neighbors*) yang jaraknya paling dekat dengak objek tersebut. Dekat atau jauhnya *neighbors* dapat dihitung berdasarkan jarak *euclidean* [21]. KNN juga memiliki kelebihan seperti tangguh dan efektif terhadap data pembelajaran yang memiliki banyak *noise* dan ukuran yang besar.

E. *Confusion Matrix*

Confusion matrix adalah sebuah alat yang digunakan untuk mengevaluasi kinerja model klasifikasi dalam tugas *data mining* atau *machine learning*. Ini digunakan untuk menggambarkan perbandingan antara hasil klasifikasi

model dengan kelas sebenarnya dari data yang diuji. *Confusion matrix* terdiri dari empat sel atau nilai, yaitu *True Positive* (TP), *False Positive* (FP), *True Negative* (TN), dan *False Negative* (FN), matrks ini yangdigunakan untuk menghitung berbagai metrik evaluasi kenerja model, seperti akurasi, *precision, recall,* dan *F1-score* [22].

F. *K-Fold Cross Validation*

K-fold cross-validation adalah metode statistik yang digunakan untuk mengevaluasi performa model atau algoritma yang telah dibuat. Pada tahap pelatihan, dataset akan dibagi menjadi dua bagian, yaitu data latih dan data validasi. Model akan dilatih menggunakan data latih dan dievaluasi menggunakan data validasi sebanyak k-fold kali [23].

G. *SMOTE*

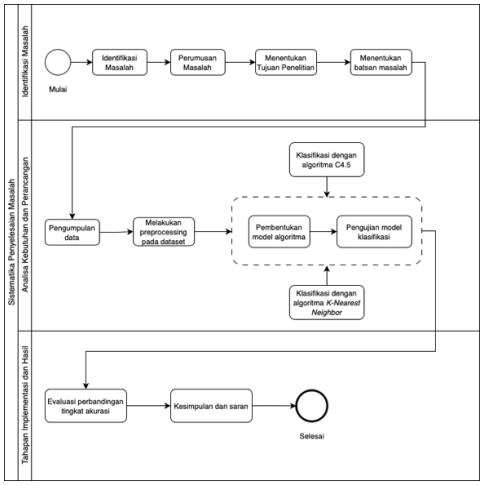
SMOTE atau singkatan dari *Synthetic Minority Oversampling Technique* adalah sebuah teknik untuk mengatasi ketidakseimbangan jumlah sampel pada kelas minoritas dengan cara menciptakan sampel sintetis. Teknik ini akan menambahkan sampel sintetis hingga jumlah sampel pada kelas minoritas menjadi seimbang dengan jumlah sampel pada kelas mayoritas [24].

H. *Hyperparameter*

Hyperparameter adalah parameter yang tidak dihitung oleh model itu sendiri selama proses pembelajaran (*training*), tetapi harus ditentukan oleh pengguna sebelum model dilatih. Dalam *machine learning* dan *deep learning*, algoritma memiliki hyperparameter yang mempengaruhi cara model belajar dan beroperasi. Pengaturan *hyperparameter* yang tepat sangat penting dalam mencapai kinerja model yang optimal. Proses penentuan *hyperparameter* yang baik dapat melibatkan eksperimen dan validasi silang (*cross-validation)* untuk menemukan kombinasi *hyperparameter* yang menghasilkan hasil terbaik pada data yang belum pernah dilihat sebelumnya [25].

I. Sistematika Penyelesaian Masalah

Dalam sistematika penyelesaian masalah dilakukan melalui tiga tahapan yaitu identifikasi masalah, analisa kebutuhan dan perancangan, dan tahapan implementasi dan hasil. Selengkapnya dijelaskan melalui gambar dari sistematika penyelesaian masalah pada gambar 1.



Gambar. 1. Sistematika penyelesaian masalah

J. Pengumpulan Data

Pada penelitian ini, peneliti melakukan perbandingan tingkat akurasi pada dua algoritma klasifikasi, yaitu *K-Nearest Neighbors* dan C4.5. Menggunakan data iklim Indonesia yang telah didapat melalui situs penyedia data yaitu Kaggle dengan alamat https://www.kaggle.com/datasets/greegtitan/indonesia-climate/. Data yang digunakan ialah data dengan periode waktu 5 tahun terakhir yakni dari tahun 2016 hingga 2020.

TABEL II
DATASET CLIMATE DATA DAILY IDN

date	Tn	Tx	Tavg	RH_avg	RR	ss	ff_x	ddd_x	ff_avg	ddd_car	station_id
01-01-2016	24.8	30.8	27.6	81.0	0.0	7.0	9.0	110.0	7.0	E	96001
01-02-2016	24.0	30.0	26.2	88.0	0.0	6.0	8.0	110.0	6.0	E	96001
01-03-2016	24.0	31.4	27.2	87.0	6.5	4.5	6.0	110.0	4.0	E	96001
01-04-2016	24.0	31.0	27.6	86.0	1.4	7.5	8.0	100.0	5.0	E	96001
01-05-2016	24.4	31.4	28.0	84.0	NaN	10.0	5.0	70.0	3.0	E	96001

K. Identifikasi Data

Pada tahap ini peneliti melakukan pemahaman dan analisis data yang sudah berhasil diperoleh tentang jenis data, format data, dan struktur data itu tersebut. Ini juga melibatkan seleksi data dengan memilih atribut yang relevan atau diperlukan untuk proses selanjutnya. Pada tahap ini, atribut yang tidak relevan atau tidak memiliki dampak signifikan pada tujuan klasifikasi dapat dihapus dari data. Atribut yang akan digunakan untuk proses pengolahan data selanjutnya yaitu Tavg, RH_avg, RR, ss dan ddd_x, atribut tersebut dipilih berdasarkan faktor yang dapat memengaruhi curah hujan. Setelah melakukan proses seleksi data, selanjutnya peneliti menentukan variabel yang akan digunakan untuk penelitian. Terdapat dua jenis variabel dalam penelitian ini yaitu variabel dependen dan variabel independen. Pembagian kedua variabel dapat dilihat pada table 3.

TABEL III
VARIABEL PENELITIAN

Variabel	Nama Variabel	Definisi
Variabel Independen (X)	Tavg	Rata-rata temperatur (°C)
	RH_avg	Rata-rata kelembaban (%)
	ss	Durasi sinar matahari (*hour*)
	ddd_x	Arah angin dengan kecepatan maksimum (º)
	ff_avg	Kecepatan angin rata-rata (m/s)
Variabel Dependen (Y)	RR	Curah hujan (mm) Dengan kategori 1 = Hujan Ringan 2 = Hujan Sedang 3 = Hujan Lebat 4 = Hujan Sangat lebat

L. *Data Preprocessing*

Pada tahap ini peneliti melakukan proses persiapan data sebelum diproses lebih lanjut. Tujuan dari tahap ini ialah untuk mengubah data mentah (*raw data*) menjadi data yang lebih siap digunakan atau dapat diolah sistem komputer. Dalam tahapan *Data Preprocessing*, tahapan dimulai dari *data cleansing* dengan menghilangkan atau menghapus data yang tidak lengkap, tidak akurat, atau duplikat dari data yang digunakan. Apabila memang terdapat data yang tidak lengkap atau bernilai *null*, maka akan dilakukan pencarian *mean* dan atau median dari setiap kolom untuk dimasukkan kedalam data yang bernilai *null*. Dapat dilihat bahwa masih terdapat nilai *null* (NaN) di dalam dataset, jika dihitung dari setiap kolomnya menghasilkan jumlah nilai *null* seperti pada di tabel 4.

TABEL IV
JUMLAH NILAI *NULL* SETIAP VARIABEL

Nama Variabel	Jumlah Nilai *Null*
Tavg	29355
RH_avg	30286
RR	76390
ss	21390
ddd_x	2215
ff_avg	1468

Setelah melakukan perhitungan jumlah nilai *null,* selanjutnya peneliti melakukan imputasi *missing value* untuk mengatasi data yang tidak memiliki nilai (*null*). Pada penelitian ini imputasi *missing value* dilakukan dengan menggantikan atau mengisi nilai yang hilang (*null*) dengan rata-rata (*mean*) dan median dari setiap variabel yang sama. Hasil dari proses tersebut dapat dilihat pada tabel 5 dan 6.

TABEL V
DATASET SEBELUM *DATA CLEANSING*

Tavg	RH_avg	RR	ss	ddd_x	ff_avg
27.6	81.0	0.0	7.0	110.0	7.0
28.0	84.0	NaN	10.0	70.0	3.0
27.2	87.0	6.5	4.5	110.0	4.0
28.1	78.0	NaN	3.0	260.0	2.0
28.4	81.0	NaN	6.5	260.0	2.0

TABEL VI
DATASET SETELAH DATA CLEANSING

Tavg	RH_avg	RR	ss	ddd_x	ff_avg
27.6	81.0	0.0	7.0	110.0	7.0
28.0	84.0	1.6	10.0	70.0	3.0
27.2	87.0	6.5	4.5	110.0	4.0
28.1	78.0	1.6	3.0	260.0	2.0
28.4	81.0	1.6	6.5	260.0	2.0

Tahapan berikutnya yaitu *data labeling*, bertujuan untuk memberikan label pada dataset yang digunakan dengan kriteria tertentu. Label diberikan pada atribut RR atau curah hujan yang dikategorikan berdasarkan probabilitas curah hujan oleh BMKG. Data tersebut dapat dilihat pada tabel 7. Setelah dilakukan proses *data labelling* pada dataset berdasarkan kriteria tertentu, tampilan data yang sudah diberi label akan seperti pada tabel 8.

TABEL VII
KATEGORI LABEL CURAH HUJAN

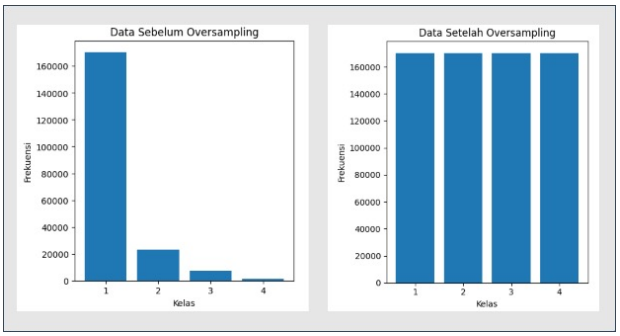
Rentang Curah Hujan	Label	Kategori Curah Hujan
0.5 – 20 mm/hari	1	Ringan
21 – 50 mm/hari	2	Sedang
51 – 100 mm/hari	3	Lebat
>100 mm/hari	4	Sangat Lebat

TABEL VIII
DATASET SETELAH *DATA LABELLING*

Tavg	RH_avg	RR	ss	ddd_x	ff_avg	label
27.6	81.0	0.0	7.0	110.0	7.0	1.0
28.0	84.0	1.6	10.0	70.0	3.0	1.0
27.2	87.0	6.5	4.5	110.0	4.0	1.0
28.1	78.0	1.6	3.0	260.0	2.0	1.0
28.4	81.0	1.6	6.5	260.0	2.0	1.0

Pada tahap selanjutnya penulis melakukan pemeriksaan apakah data jumlah data sampel yang digunakan seimbang atau tidak. Hasil pemeriksaan mendapati bahwa jumlah data setiap kelasnya tidak seimbang. Oleh karenanya penulis melakukan *data oversampling* yaitu teknik dalam pemrosesan data yang digunakan untuk mengatasi masalah ketidakseimbangan kelas dalam dataset. Dalam penelitian ini penulis menggunakan pendekatan *SMOTE* untuk melakukan data oversampling.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 9, No. 2, Juni 2024, Pp. 628-638



Gambar. 2. Grafik Jumlah Data Sebelum dan Sesudah *Oversampling*

M. Pemodelan dan Evaluasi

Setelah proses *data preprocessing* selesai dapat diartikan bahwa data sudah siap diolah untuk pemodelan data. Pada penelitian ini penulis menggunakan algoritma C4.5 dan *K-Nearest Neighbors*. *Tools* yang digunakan yaitu *Google Colab* dengan bahasa pemrograman *python*. Setelah berhasil membuat model tahapan selanjutnya yaitu melakukan tahap evaluasi untuk menguji kinerja dari masing-masing algoritma. Tahap evaluasi dilakukan dengan metode *Confusion Matrix*, *Precision*, *Recall* dan *F1-Score*.

III. HASIL DAN PEMBAHASAN

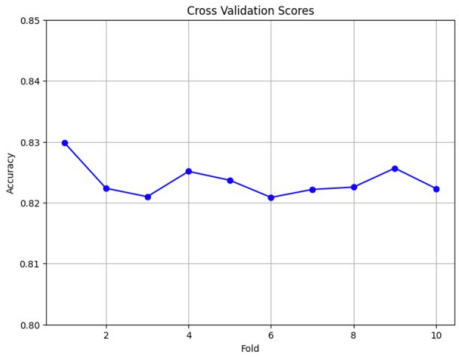
A. Implementasi Algoritma C4.5

Dalam melakukan implementasi algoritma C4.5 pada penelitian ini, penulis menggunakan fungsi *DecisionTreeClassifier* yang merupakan bagian dari modul *tree* dalam *library scikit-learn* yang digunakan untuk membangun model klasifikasi berbasis pohon keputusan. Pengujian awal algoritma C4.5 dengan menggunakan rasio perbandingan *data splitting* 80:20 mendapatkan hasil akurasi sebesar 81.42%.

Metode validasi hasil pengujian yang digunakan dalam penelitian kali ini yaitu *K-Fold Cross Validation*. Pengujian *K-Fold Cross Validation* pada algoritma C4.5 menggunakan nilai k sebanyak 10 fold dan rasio perbandingan 80:20. Hasil validasi dapat dilihat pada tabel 9 dan gambar 3 berikut.

TABEL XI
K-FOLD CROSS VALIDATION ALGORITMA C4.5

Fold ke-	1	2	3	4	5	6	7	8	9	10
Akurasi	0.829	0.822	0.820	0.825	0.823	0.820	0.822	0.822	0.825	0.822
Akurasi rata-rata	82.36%									



Gambar. 3. *K-fold Cross Validation* C4.5

Setelah melakukan pengujian, didapatkan hasil akurasi *K-Fold Cross Validation* pada algoritma C4.5 dengan akurasi rata-rata sebesar 82.36 %. Dapat dilihat juga bahwa tren tergolong stabil, dalam artian performa model yang diukur dengan metode *K-Fold Cross Validation* menunjukkan sedikit variasi antara setiap iterasi validasi. Dengan stabilitas yang baik, kemungkinan besar model tersebut juga tidak mengalami *overfitting* (terlalu sesuai dengan data pelatihan) atau *underfitting* (tidak cukup memahami pola dalam data).

Pada tahap selanjutnya penulis melakukan proses *Tuning Hyperparamater* untuk meningkatkan kinerja dan generalisasi model dengan menemukan kombinasi hyperparameter yang optimal, dengan harapan dapat

menghasilkan nilai akurasi yang lebih tinggi dari pengujian awal. Pada penelitian kali ini metode tuning hyperparameter yang digunakan adalah *Randomizedsearch Cross Validation*. Hasil *tuning hyperparameter* dan akurasi algoritma C4.5 setelah menggunakan hasil *tuning hyperparameter* dapat dilihat pada tabel 10.

TABEL X
HASIL TUNING HYPERPARAMETER ALGORITMA C4.5

No.	*Hyperparameter*	Nilai
1.	*min_samples_split*	3
2.	*criterion*	*entropy*
3.	*max_depth*	256
4.	*min_samples_leaf*	1
Akurasi		82.68%

Dapat dilihat pada tabel 10 bahwa dari hasil pengujian dengan menerapkan *tuning hyperparameter* dengan metode *Randomizedsearch Cross Validation* pada algoritma C4.5 menghasilkan peningkatan nilai akurasi sebesar 82.68%. Selanjutnya penulis melakukan proses evaluasi performa algoritma C4.5 dengan menggunakan metode *Confusion Matrix, Precission, Recall* dan *F1-Score.* Tahapan didahului dengan melakukan evaluasi dengan *Confusion Matrix.* Hasil *Confusion Matrix* dari algoritma C4.5 dapat dilihat pada tabel 11.

TABEL XI
CONFUSION MATRIX ALGORITMA C4.5

	Pred 1	Pred 2	Pred 3	Pred 4
Actual 1	28230	4870	1350	196
Actual 2	4758	24290	3849	983
Actual 3	1557	3802	27959	936
Actual 4	246	736	723	32206

Setelah mendapatkan hasil dari *Confusion Matrix*, selanjutnya penbulis akan melakukan perhitungan terhadap nilai *Recall, Precision,* dan *F1-Score* dari algoritma C4.5. Karena tipe kelas dalam penelitian ini bertipe *multiclass*, oleh karenanya perhitungan dilakukan untuk setiap kelas secara individu. Hal ini membantu dalam mengevaluasi sejauh mana model dapat mengklasifikasi data pada setiap kelas secara akurat, mengidentifikasi kelas yang mungkin memiliki masalah klasifikasi, dan membandingkan performa model antara kelas-kelas yang berbeda. Hasil perhitungan dengan rata-rata setiap kelas dapat dilihat pada tabel 12.

TABEL XII
HASIL EVALUASI ALGORITMA C4.5

Metrik	Nilai
Precision	82%
Recall	82.25%
F1-Score	82.25%

B. Implementasi Algoritma *K-Nearest Neighbors*

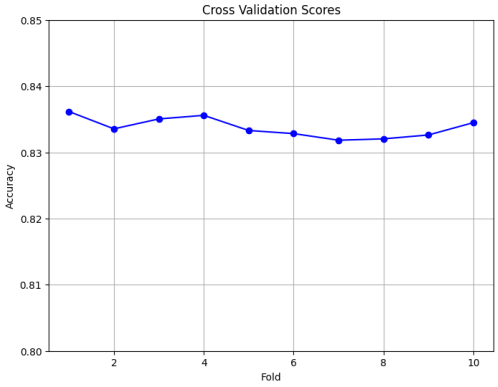
Pada tahap sebelumnya penulis sudah berhasil melakukan implementasi terhadap algoritma C4.5, tahapan selanjutnya penulis akan melakukan tahapan yang serupa terhadap algoritma *K-Nearest Neighbors*. Pengujian awal algoritma *K-Nearest Neighbors* dengan menggunakan rasio perbandingan *data splitting* 80:20 mendapatkan hasil akurasi sebesar 78.10%.

Tahap validasi juga dilakukan dengan *K-Fold Cross Validation*. Pengujian *K-Fold Cross Validation* pada algoritma *K-Nearest Neighbors* menggunakan nilai k sebanyak 10 fold dan rasio perbandingan 80:20. Hasil K-fold Cross Validation pada algoritma C4.5 dapat dilihat pada tabel 13 dan gambar 4 berikut.

TABEL XIII
K-FOLD CROSS VALIDATION ALGORITMA KNN

Fold ke-	1	2	3	4	5	6	7	8	9	10
Akurasi	0.836	0.833	0.835	0.835	0.833	0.832	0.831	0.832	0.832	0.834
Akurasi rata-rata	83.37%									

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 9, No. 2, Juni 2024, Pp. 628-638



Gambar. 4. *K-fold Cross Validation* Algoritma KNN

Setelah melakukan pengujian, didapatkan hasil akurasi *K-Fold Cross Validation* pada algoritma *K-Nearest Neighbors* dengan akurasi rata-rata sebesar 83.37 %. Hasil visualisasi dari *K-Fold Cross Validation* pada algoritma *K-Nearest Neighbors* juga menunjukkan tren yang cukup stabil. Ini berarti performa model yang diukur dengan metode *K-Fold Cross Validation* menunjukkan sedikit variasi antara setiap iterasi validasi.

TABEL XVI
HASIL *TUNING HYPERPARAMETER* ALGORITMA KNN

No.	*Hyperparameter*	Nilai
1.	*weights*	*uniform*
2.	*n_neighbors*	*1*
3.	*metric*	*minkowski*
4.	*leaf_size*	41
Akurasi		83.94%

Selanjutnya penulis juga melakukan hal yang sama dengan melakukan *tuning hyperparameter* pada algoritma *K-Nearest Neighbors*, dan mendapatkan hasil parameter serta hasil akurasi seperti pada tabel 14 diatas. Setelah melakukan pengujian dengan menerapkan *tuning hyperparameter* metode *Randomizedsearch Cross Validation* pada algoritma *K-Nearest Neighbors* terhadap rasio perbandingan 80:20, terbukti bahwa nilai akurasi mengalami peningkatan yang cukup signifikasi dari nilai akurasi pada pengujian awal sebelumnya. Peningkatan akurasi menghasilkan nilai sebesar 83.94%.

Selanjutnya penulis juga melakukan proses evaluasi performa algoritma *K-Nearest Neighbors* dengan menggunakan metode *Confusion Matrix, Precission, Recall* dan *F1-Score.* Tahapan didahului dengan melakukan evaluasi dengan *Confusion Matrix.* Hasil *Confusion Matrix* dari algoritma *K-Nearest Neighbors* dapat dilihat pada tabel 15.

TABEL XV
CONFUSION MATRIX ALGORITMA KNN

	Pred 1	Pred 2	Pred 3	Pred 4
Actual 1	23560	6741	3046	896
Actual 2	3662	27051	2367	800
Actual 3	1369	1783	30511	591
Actual 4	176	281	175	33279

Setelah mendapatkan hasil dari *Confusion Matrix*, selanjutnya penulis akan melakukan perhitungan terhadap nilai *Recall, Precision*, dan *F1-Score* dari algoritma *K-Nearest Neighbors*. Karena tipe kelas dalam penelitian ini bertipe multiclass, oleh karenanya perhitungan dilakukan untuk setiap kelas secara individu. Hal ini membantu dalam mengevaluasi sejauh mana model dapat mengklasifikasi data pada setiap kelas secara akurat, mengidentifikasi kelas yang mungkin memiliki masalah klasifikasi, dan membandingkan performa model antara kelas-kelas yang berbeda. Hasil perhitungan dengan rata-rata setiap kelas dapat dilihat pada tabel 16.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 9, No. 2, Juni 2024, Pp. 628-638

TABEL XIV
HASIL EVALUASI ALGORITMA KNN

Metrik	Nilai
Precision	83.50%
Recall	83.40%
F1-Score	83%

C. Evaluasi Perbandingan Algoritma

Tabel 17 menampilkan hasil perbandingan dari kedua algoritma yaitu algoritma C4.5 dan *K-Nearest Neighbors.* Pertama dilakukan perbandingan pada nilai akurasi awal, pada perbandingan ini algoritma C4.5 lebih unggul dengan nilai akurasi 81.42% dibanding dengan nilai akurasi yang diperoleh algoritma *K-Nearest Neighbors* sebesar 78.10%. Setelah itu perbandingan kedua dilakukan pada nilai *K-Fold Cross Validation* dari kedua algoritma, pada perbandingan ini algoritma *K-Nearest Neighbors* unggul dengan nilai 83.37% dan 82.56% untuk algoritma C4.5 Selanjutnya perbandingan ketiga dilakukan pada nilai akurasi setelah menerapkan *Tuning Hyperparameter* dengan metode *Randomizedsearch CV*, dengan menggunakan parameter terbaik, nilai akurasi pada algoritma *K-Nearest Neighbors* mengalami kenaikan yang cukup signifikan sehingga melebihi nilai akurasi awal dan lebih besar dari algoritma C4.5. Nilai akurasi yang didapat sebesar 83.94% untuk algoritma *K-Nearest Neighbors,* sedangkan untuk algoritma C4.5 sebesar 82.68%.Selanjutnya penulis juga melakukan perbandingan pada nilai evaluasi model dari kedua algoritma, dimana pada nilai *Precision* untuk algoritma *K-Nearest Neighbors* memperoleh angka 83.25%, sedangkan untuk algoritma C4.5 memperoleh angka 82%. Selanjutnya pada nilai *Recall* untuk algoritma *K-Nearest Neighbors* memperoleh angka 83.50%, sedangkan untuk algoritma C4.5 memperoleh angka 82.25%. Sedangkan pada nilai *F1-Score* untuk algoritma *K-NearestNeighbors* memperoleh angka 83% dan 82.25% untuk algoritma C4.5. Dari hasil perbandingan ketiga evaluasi model pada kedua algoritma, dapat dilihat bahwa algoritma *K-Nearest Neighbors* menghasilkan nilai akurasi yang lebih baik dari algoritma C4.5.

Dengan hasil akurasi yang diperoleh, terbukti seperti pada penelitian-penelitian sebelumnya bahwa kedua algoritma yaitu algoritma C4.5 dan *K-Nearest Neighbors* memang mampu mengklasifikasi data curah hujan dengan baik, hal tersebut dikarenakan nilai akurasi yang diperoleh diatas angka 80%. Selanjutnya penambahan penggunaan *K-Fold Cross Validation* juga dapat memastikan bahwa kedua model cukup konsisten dan tidak mengalami *underfitting* dan *overfitting*. Lalu penambahan penggunaan *Tuning Hyperparameter* juga dapat memberikan kontribusi guna upaya meningkatkan kinerja dari kedua model sehingga memperoleh nilai akurasi yang maksimal. Dengan begitu, hasil yang sudah diperoleh dengan menambahkan metode *K-Folde Cross Validation* dan *Tuning Hyperparameter* pada penelitian ini dapat meningkatkan tingkat ketepatan hasil akurasi dan meminimalisir nilai *error*.

TABEL XVII
HASIL PERBANDINGAN KEDUA ALGORITMA

	C4.5	K-Nearest Neighbors
Accuracy	81.42%	78.10%
K-Fold Cross Validarion	82.56%	83.37%
Tuning Hyperparameter (accuracy)	82.68%	83.94%
Precision	82%	83.50%
Recall	82.25%	83.50%
F1-Score	82.25%	83%

IV. KESIMPULAN

Berdasarkan hasil keseluruhan pengujian yang dilakukan, kedua algoritma dapat dikatakan sebagai algoritma yang baik dalam mengklasifikasi curah hujan berdasarkan iklim Indonesia. Hal tersebut dikarenakan semua hasil akhir dari kedua algoritma menghasilkan angka diatas 80%. Namun satu hal yang dapat menjadi perhatian yaitu pada pengujian awal, dimana sebelum menggunakan *Tuning Hyperparameter* pada algoritma C4.5 menghasilkan nilai akurasi yang lebih tinggi dibanding nilai akurasi dari algoritma *K-Nearest Neighbors.* Hal tersebut dapat diartikan bahwa dengan menggunakan parameter default dari kedua algoritma, algoritma C4.5 lebih unggul dengan menghasil nilai akurasi yang lebih tinggi. Namun setelah menggunakan Tuning Hyperparameter, algoritma *K-Nearest Neighbors* memiliki angka yang lebih tinggi daripada algoritma C4.5 dalam hal akurasi, *K-Fold Cross Validation*, *Precision, Recall* dan *F1-Score*.

DAFTAR PUSTAKA

[1] S. Prawirowardoyo, *Meteorology*. Bandung: ITB, 1996.
[2] N. Sunarmi *et al.*, "Analisis Faktor Unsur Cuaca terhadap Perubahan Iklim Di Kabupaten Pasuruan pada Tahun 2021 dengan Metode Principal Component Analysis," *Newton-Maxwell Journal of Physics*, vol. 3, no. 2, Oct. 2022, [Online]. Available: https://www.ejournal.unib.ac.id/index.php/nmj
[3] S. B. Sipayung, "Dampak Variabilitas Iklim Terhadap Produksi Pangan di Sumatera," vol. 2, Jun. 2005.
[4] E. Aldrian, "Sistem Peringatan Dini Menghadapi Iklim Ekstrem," vol. 10, no. 2, Dec. 2016.
[5] H. A. Tambunan and D. Saputra, "Rancang Bangun Aplikasi Prediksi Cuaca Berbasis Android," *Jurnal Bisantara Informatika (JBI)*, vol. 6, no. 2, 2022.
[6] S. Chodijah, "Strategi Komunikasi Penyampaikan Informasi Iklim Stasiun Klimatologi Sampali Medan Dalam Upaya Meminimalkan Kegagalan Panen Padi Sawah Akibat Iklim Ekstrim," *Persepsi: Communication Journal*, vol. 1, no. 1, pp. 55–69, Nov. 2018, doi: 10.30596/persepsi.v1i1.2506.
[7] J. H. Yousif, H. A. Al-Balushi, H. A. Kazem, and M. T. Chaichan, "Analysis and forecasting of weather conditions in Oman for renewable energy applications," *Case Studies in Thermal Engineering*, vol. 13, p. 100355, Mar. 2019, doi: 10.1016/J.CSITE.2018.11.006.
[8] B. Poernomo, R. Dewi, and I. Sari, "Penerapan Data Mining untuk Prakiraan Cuaca di Kota Malang Menggunakan Algoritma Iterative Dichotomiser Tree (ID3)," *JOUTICLA*, vol. 3, no. 2, 2017.
[9] Irmayani, "Penerapan Algoritma CART Klasisifikasi Sosial Ekonomi Masyarakat Kelurahan Amessangeng," *Jurnal Ilmiah Information Technology d'Computare*, vol. 10, Jul. 2020.
[10] J. Han and M. Kamber, "Designing Data-Intensive Web Applications," 2006.
[11] P. Meilina, "Penerapan Data Mining dengan Metode Klasifikasi Menggunakan Decision Tree dan Regresi," *Jurnal Teknologi Universitas Muhammadiyah Jakarta*, vol. 7, no. 1, 2015.
[12] R. Purba, "Data Mining : Masa Lalu, Sekarang dan Masa Mendatang," vol. 13, no. 1, 2012.
[13] S. Anastassia Amellia Kharis and A. Haqqi Anna Zili, "Learning Analytics dan Educational Data Mining pada Data Pendidikan," *Jurnal Riset Pembelajaran Matematika Sekolah*, vol. 6, 2022.
[14] A. Al Arif, M. Firdaus, Y. Maruhawa, S. AMIK Riau, and J. Purwodadi Panam, "Perbandingan Metode Data Mining untuk Prediksi Curah Hujan dengan Algoritma C4.5, Naïve Bayes, dan KNN," *Institut Riset dan Publikasi Indonesia (IRPI)*, pp. 187–197, Jul. 2022, [Online]. Available: https://journal.irpi.or.id/index.php/sentimas
[15] G. Karyono, "Analisis Teknik Data Mining 'Algoritma C4.5 dan K-Nearest Neighbor' untuk Mendiagnosa Penyakit Diabetes Mellitus," *Seminar Nasional Teknologi Informasi*, May 2016.
[16] Y. Suwarno, *Inovasi di Sektor Publik*. Jakarta : STIA-LAN Press, 2008.
[17] E. Turban, *Mechine Learning untuk Mengesktraksi dan Mengidentifikasi Informasi yang Bermanfaat*. 2005.
[18] B. Santosa, *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Yogyakarta: Andi, 2007.
[19] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Boston: Pearson Addison-Wesley., 2006.
[20] M. Wisnu Prihatmono and A. Felicia Watratan, "Implementasi Algoritma C4.5 Menggunakan Python Untuk Klasifikasi Kepuasan Konsumen," 2019.
[21] Rafiq Amaliyah, "Aplikasi Klasifikasi Citra Kerusakan Aspal Menggunakan Matlab 2013A," Universitas Gunadarma, 2014.
[22] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer Science+Business Media, LLC, 2006. [Online]. Available: http://research.microsoft.com
[23] Y. N. Fuada, I. D. Ubaidullah, N. Ibrahim, F. F. Talingsing, N. K. Sy, and M. A. Pramudhito, "Optimasi Convolutional Neural Network dan K-Fold Cross Validation pada Sistem Klasifikasi Glaukoma," *ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika*, vol. 10, no. 3, p. 728, Jul. 2022, doi: 10.26760/elkomika.v10i3.728.
[24] R. Siringoringo, "Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE dan K-Nearest Neighbor," vol. 3, no. 1, 2018.
[25] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.